TIGHT BINDING BOOK

UNIVERSAL
LIBRARY

By the same author

ASTRONOMY FOR NIGHT WATCHERS
(*Faber & Faber*)

DIRECTION FINDING BY THE STARS
(*Faber & Faber*)

THE SHORTER POEMS OF WALTER SAVAGE LANDOR
(*Cambridge University Press*)

THE HEAVENS ABOVE

A Rationale of Astronomy

BY

J. B. SIDGWICK, F.R.A.S.

Member of the British Astronomical Association and of the Société Astronomique de France

Geoffrey Cumberlege
OXFORD UNIVERSITY PRESS
LONDON NEW YORK TORONTO
1948

Oxford University Press, Amen House, London E.C.4

EDINBURGH GLASGOW NEW YORK TORONTO MELBOURNE
WELLINGTON BOMBAY CALCUTTA MADRAS CAPE TOWN

Geoffrey Cumberlege, Publisher to the University

PRINTED IN GREAT BRITAIN

DEDICATED TO

K. M. S.

Singulariter sum ego donec transeam

CONTENTS

PART II: QUALITY – THE NATURE OF THINGS

LIST OF PHOTOGRAPHS

PREFACE

ONE of the less dramatic developments of the period between the wars was the appearance of Astronomy in the best-seller market—that most restricted if not most select of worlds. Whereas at one time the astronomer was not encouraged by the reading public to issue from his observatory in an official capacity, one now finds (or, at any rate, found at the height of the boom) hardly a station bookstall without its volume of celestial revelations.

During the war itself, two factors conspired to continue this trend. The blackout made visible the night sky to the townsman for the first time since the introduction of street lighting; and a variety of national duties gave both town and country dwellers plenty of opportunity for seeing it.

Thus it has come about that a wide public is in somewhat precarious possession of the more spectacular results of modern astronomical research without quite knowing how these have been established. For fine superstructures must of necessity stand upon firm foundations, and many of these astronomical dilettanti must have wondered just how such sensational results were obtained; they may, too, have speculated vaguely upon the basic principles, the trains of reasoning and the observations without which this elegant and mysterious edifice could never have been built.

Even the simplest of everyday assumptions appears, on close inspection, to be riddled with pitfalls. Indeed, the 'simpler' it is, the more difficult its substantiation frequently is: it is easier to play a first-class game of chess than to make a rational and convincing defence of the popular notions concerning time, causation, or the existence of a material world—a defence which a professional philosopher could not tear to shreds in half a minute. Ask anyone without warning why he believes that the earth rotates on its axis, or that it revolves about the sun rather than the sun about it, and he will in all likelihood be unable to provide a sufficient reason without several false starts—if then. The answer that that is what he has always been taught, or, ultimately, that that is the considered view of individuals who have studied the facts, is no real answer. It may well be his reason for holding the belief, but it is certainly not a sufficient reason.

In the pages that follow, an attempt has been made to give a systematic demonstration of the more complex facts of Astronomy, starting from simple assumptions at which not even the most sceptical

reader could cavil. So far as is convenient, the various artificial aids of the natural senses, such as the telescope and the spectroscope, are drawn into the discussion in their historical sequence. Thus, in Part I the evidence of the naked eyesight is examined, and the fullest possible use made of it by the reasoning faculty; only then is the evidence of the telescope invoked. Data provided by the spectroscope are confined to Part II, the initial chapter of which introduces the subject of spectroscopy and sketches in the background of atomic physics.

Much play has in the past been made of astronomical distances by 'popular' writers on the subject. The size of the known universe is certainly one of its more awe-inspiring features, and sizes and distances on the astronomical scale are clearly a gift to any journalist endeavouring to make a write-up exciting. The determination of all astronomical distances from the smallest to the greatest constitutes a single logical process, a fact which is often not made clear in introductions to the subject intended for the general reader. Not all astronomical amateurs would go so far as George Bernard Shaw, who once roundly accused all astronomers of being liars; but owing to the basic importance of the distance determinations, combined with the somewhat sketchy manner in which they are often treated, the average man may be forgiven a certain degree of puzzlement, if not scepticism.

For these reasons, the means whereby man has put a scale to the visible universe are in Part I presented as a continuously unfolding story, and the interrelated character of all stages in the extrapolation emphasized. The aim of the first five chapters, in other words, is to describe the size, shape and structure of the astronomical region and its contents, together with the methods by which the results have been obtained. In Part II a different approach is adopted, and the various bodies whose distances have been their primary interest in the preceding section are considered as things in themselves, rather than as anonymous points in a spatial pattern.

From time to time during the development of the argument, reference is made to evidence provided by an instrument not yet described. When this happens, however, it is never to forge a link in the main chain of the argument, but to provide subsidiary interest or contributory evidence of an already established conclusion. Brief descriptions of certain unproved hypotheses—such as the solar origin of the planets, the origin of the lunar craters, the subatomic sources of stellar energy—are inserted for the same reason.

Finally, a word of reassurance to the mathematically timid. Figures have purposely been kept to a minimum, but certain facts and trains

of reasoning are of their nature more cogently and conclusively expressed in mathematical than in verbal terms. A few figures do accordingly appear from time to time, but in no case would they present any difficulty to a schoolboy at matriculation standard. Moreover, should the reader on encountering them still prefer to skip to the next piece of firm ground, he may do so without losing the thread of the argument. For in every case the point is summarized verbally.

I should like to acknowledge here my extreme indebtedness to Professor H. H. Plaskett, whose assistance, encouragement and constructive criticism are in large measure responsible for whatever value this book may possess; also to Dr. W. H. Steavenson for reading the book in proof and making many helpful suggestions, as well as correcting a number of factual errors; and to Mrs. Doreen Marston for giving me the initial idea.

LONDON
January 1947

Part I

QUANTITY—A GEOMETRIC PICTURE

I

THE PROBLEM STATED: APPARENT MOTIONS

THE history of man's growing knowledge of the heavens, and of the movements, sizes, distances and chemical and physical constitution of the various heavenly bodies, is largely the story of his invention and progressive refinement of numerous instruments whose function is to increase the scope of his natural senses and powers of reasoning. Of these instruments, the more important are:

i. Telescopes of a variety of types.

ii. Micrometers (instruments for the accurate measurement of minute angular distances and differences of position).

iii. Cameras.

iv. The spectroscope.

v. The spectroheliograph.

vi. Mathematics, particularly trigonometry and the calculus.

With the aid of this battery of instruments, the astronomer is ab to project his inquiring mind into the furthermost corners of tl visible universe.

Naked-eye astronomy

But the first of these instruments to be discovered (excluding from consideration mathematics, a mental rather than a physical instrument) did not appear until the early seventeenth century,[1] and by that time a great deal had already been learnt of the relative positions and movements of the sun, moon, earth and the other planets. This knowledge was erected upon a foundation of naked-eye observations, refined to a certain extent by the use of primitive measuring instruments of the yard-stick and sextant type. Thus the mind of man is capable of building a considerable edifice of astronomical knowledge with no materials other than the evidences of his unaided senses. The history of the science has taught us, however, that this may only be achieved on the fulfilment of two conditions:

i. The mind must be free at the outset from preconceived ideas of what *ought* to be, and prepared to reason from the evidence of the senses *alone*.

[1] A passage in one of Leonardo da Vinci's *Notebooks* suggests that he may have constructed a primitive telescope and used it for astronomical purposes.

ii. The eye must be trained to make accurate observations of the motions and positions of the celestial objects that it perceives. This is the more easily fulfilled condition of the two.

The form of the earth

There is a preliminary point to be settled before we go out of doors to study the day and night skies and to note carefully the behaviour of the celestial bodies, in an endeavour to recreate the history of astronomy up to the time of Galileo. What is the true shape of the earth, that station from which all our observations must necessarily be made? It may be objected that the earth is spherical and that any fool is perfectly well aware of the fact. This objection, however, overlooks the professed aim of this book—to demonstrate the more complex conclusions of astronomical research by working upwards from simple assumptions that cannot rationally be denied, demonstrating each step where it can be demonstrated and at other times indicating the possible alternatives and the reason for believing one in particular to be the more probable. In short, to take the bare minimum for granted. Now, it is far from self-evident that the earth is not flat. In fact, every reader of these words has at some period during his early life received a profound shock on being told quite seriously that (i) the earth is the same shape as an orange, and that (ii) the inhabitants of Australia not only hang head downward like flies, but solemnly affirm that not they, but the inhabitants of England, live in this inconvenient attitude. It is undeniably true that the 'flat earthists' have the fact of immediate obviousness in their favour. Yet they are considered cranks. The following are some of the more important reasons why the evidence of their senses is denied by the vast majority of civilised adults.

i. However far, and in whatever direction, one travels over the surface of the earth, one never comes to an edge.

ii. Moreover, as the sixteenth century explorers discovered, it is possible to arrive at the point from which one starts merely by following one's nose for a sufficiently great distance.

iii. The telescope reveals the true shape of a number of bodies in space whose nature cannot be discovered by the naked eye. Though many of these, including those which (for reasons not yet stated) are believed most closely to resemble the earth, are spherical, not one has yet been discovered which is flat—i.e. a plane body of negligible thickness.

iv. When ships approach the horizon, their hulls disappear first, then their funnels, and lastly their masts. This can only occur in a world whose seas, and therefore that world itself, are convex.

v. The curvature of the earth's surface can actually be seen.

Since this curvature is very slight, an unruffled stretch of horizontal surface of considerable extent is required for the experiment. Such a surface is afforded by, for instance, the Corinth Canal or the Bedford Level. Suppose the reader were to drive three stakes into the centre of the canal bed, one at each end of the canal and the third midway between them. He would have to be careful to see that each stake projected above the surface of the water by the same amount, say 1 foot. If

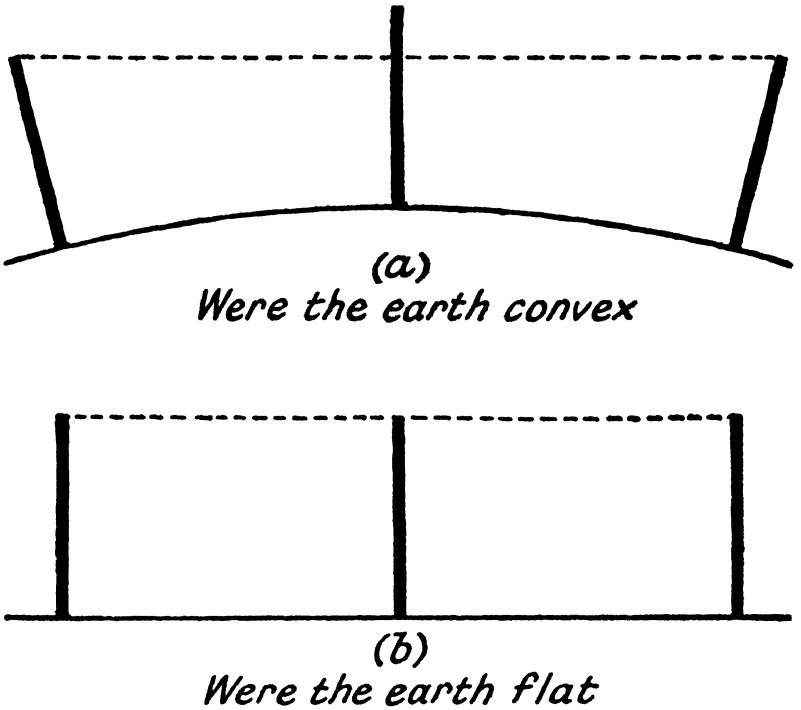


Figure 1. The stakes are the same length in each case.

then he placed a telescope upon the first stake and endeavoured to observe the top of the stake at the other end of the canal, he would find that he could not do so. His view would be obscured by the middle stake. The significance of this fact is made clear by Fig. 1.

It may also be noted that the curvature is clearly, and very strikingly, visible to the naked eye on the straight stretch of the Suez Canal immediately south of Suez.

vi. Points (i), (iv) and (v) prove that the earth's surface is convex while point (iii) suggests that it is spherical. This conclusion is confirmed by eclipses of the moon. Under certain circumstances, to be

described later, the earth passes directly between the sun and the moon. When this happens its shadow falls upon the surface of the moon. It is found that the edge of this shadow is always an arc of a circle; this is the shape of the shadow cast by a spherical body. We may, then, assume as our starting-point that the earth is not a flat, but a spherical body, and that it is suspended in space by some means still to be determined.

The size of the earth

It is of importance that we should know at the outset of our investigations, not only the shape of the earth, but also its size. It might be supposed, since we can never get far enough away from the earth to be able to see it as a whole, and since it is not practicable to walk round it with a tape measure, that its measurement must present a very difficult, if not an insoluble, problem. This is not the case, however, the theory of the process being extremely simple. Such difficulties as are encountered are practical ones.

Let us go out of doors on a clear and preferably moonless night. At once it will be seen that the sky is dotted with innumerable points of light, the stars. These stars have several interesting properties which are perceptible to the unaided eyesight, but for the moment we will confine our attention to one only. Quite a short period of observation of the southern sky will reveal the fact that the stars are moving *en bloc*—that is, without altering their relative positions, one to another—from east to west. This absence of relative motion makes it possible for us to imagine that they are fixed to the inner side of a gigantic vault with the earth at its centre, and that this vault is in rotation, carrying the stars with it.

Imagine, for a moment, a pudding basin, on the inner surface of which are marked a number of dots; the basin is pivoted at the centre of its base and may be rotated about this pivot (Fig. 2). If the basin is made to rotate through one quarter of a turn, the dots which were at *a*, *b*, *c*, *d*, *e*, and *f* will now be in positions *a'*, *b'* . . . *f'*, and the paths that they have traced out will be the dotted lines in the diagram. A study of this diagram will reveal two obvious but important facts.

i. The nearer a dot is to the centre of rotation, the shorter will be its path for a given movement of the basin.

ii. The dots on opposite sides of the pivot move in opposite directions. In the figure the rotation is clockwise, with the result that those dots above the pivot move from left to right and those below it from right to left.

Returning now to the night sky, careful observation shows that stars about half way up the southern sky move further—i.e. describe

a larger arc—in a given period of time than those at a greater altitude above the horizon; that these, in turn, move further than those directly overhead. And so on until we reach a point in the sky about midway between the northern horizon and the zenith (the point directly above the observer's head). At this point in the northern sky there happens to lie a star which, during the course of an hour's observation, apparently does not move at all. Even longer periods of observation fail to detect any motion in it, and we may therefore conclude that it must lie very near the point on the star sphere about

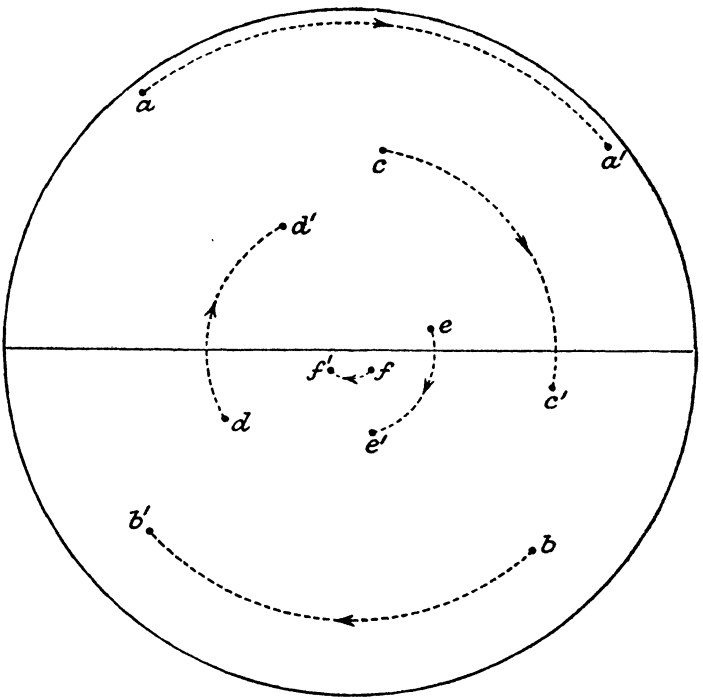


Figure 2. Motion at different distances from, and on opposite sides of, the centre of rotation.

which the whole vault revolves (see *ff'* in Fig. 2). This star is called Polaris, or the Pole Star, and the point on the star sphere close to it (about which the whole sphere revolves) the north celestial pole. The point diametrically opposed to it is the other pivot, the south celestial pole; this obviously cannot be seen from places in the northern hemisphere of the earth, any more than inhabitants of the antipodes can see Polaris. We may therefore define the celestial poles as those points upon the star sphere which have no diurnal (or twenty-four hourly) motion.

In accordance with the second result derived from our experiment

with the pudding basin, those stars above the north celestial pole—i.e. those between it and the zenith, *plus* those between the zenith and the southern horizon—revolve in an east-west direction; while those below or to the north of it revolve in the opposite direction, from west to east. These elementary facts of the diurnal rotation of the star vault will be obvious to anyone who has spent a couple of hours in intelligent observation of the night sky.

All this is preliminary to the determination of the earth's size. It has been discovered that the angular elevation of the Pole Star (which, it will be remembered, is very near the north celestial pole) above the northern horizon, depends upon the position of the observer on the earth's surface—more accurately, upon his latitude. In our latitudes, approximately midway between the equator and the north pole, Polaris is seen to lie approximately midway between the zenith and the northern horizon. As we travel north, the Pole Star rises higher and higher in the sky until, when we arrive at the north pole itself, it is directly overhead. Similarly, as we move south towards the equator it sinks lower and lower towards the northern horizon. When we reach the equator itself we find that it lies on the northern horizon; further southward movement will render it invisible, since it lies below this horizon. Put in another way, we may say that when the north celestial pole has an altitude of 90° (is overhead) our latitude is 90° (the north pole); when its altitude is 45° our latitude is 45° (we are midway between the equator and the north pole—at Turin, perhaps); and when its altitude is 0° (actually on the horizon) our latitude is 0° (the latitude of the equator). Incidentally it is to be noted that this fact provides us with another reason for believing that the earth is spherical—or at any rate not flat. For if it were, the elevation of the celestial pole above the horizon would be the same from all points of observation.

In fact, the latitude of a place is nothing more than the angular elevation of the north celestial pole (or the south celestial pole in the southern terrestrial hemisphere) above the northern (or southern) point on the horizon as observed from that place. The reader who can lay his hands on a sextant can quite easily determine his latitude. If he lives in London he will find that the angle between the Pole Star and the horizontal averages about 51½°; if he then looks in his atlas he will find that London is situated approximately midway between the 51st and 52nd parallels of latitude.

The ground is now cleared for the determination of the earth's size. First, two points on the earth's surface, due north and south of each other, are chosen, and their distance apart determined with an accuracy of at least 1 yard per hundred miles. It is in this stage of

the process that the practical difficulties are encountered. The fixing of two widely separated points whose distance apart is known to within a few inches is accomplished by triangulation survey with the most accurate instruments; even so, it is both lengthy and laborious. Then the elevation of the Pole Star[1] is taken from each station. In this way their difference in latitude is determined. Suppose, to make the explanation concrete, that the two stations are 100 miles apart, and it is found from the polar observations that their difference in latitude amounts to $1\frac{1}{2}°$. Hence, if $1\frac{1}{2}°$ latitude = 100 miles,

$$360° \text{ (the earth's circumference)}$$
$$= \frac{360}{1\frac{1}{2}} \times 100 \text{ miles}$$
$$= 25{,}000 \text{ miles, approximately.}$$

Hence, by an operation with which fourth form schoolboys are familiar, the diameter and radius (and, if necessary, the volume) of the earth may be deduced. These figures have been kept approximate and simple intentionally; in practice the determination can be, and has been, carried out with a high degree of accuracy.

Having prepared the way for our inquiry by discovering what sort of a body it is that we inhabit, and from which we must perforce make all our astronomical observations, we can proceed to the investigation of the motions and, so far as they can be deduced without the aid of instruments, the positions and distances of those other celestial bodies that are visible in the day and night skies. It is assumed that the reader undertakes to spend a short time in the open each evening, there to study the appearance of the night sky.

Diurnal rotation of the star sphere

The stars are sizeless points of light, varying in brightness and also in colour. The differences in colour shown to the naked eye by individual stars are not striking. Nevertheless, they can be noticed quite clearly in the case of a number of the brighter stars.

We have already discovered that the star sphere rotates steadily from east to west (if we are looking at that part of it that lies south of the north celestial pole—as we usually are, since it is by far the larger part) about a pivot in the northern sky which is nearly marked by the star Polaris. Since we are now engaged in practical observation, it will be as well to find this star at once. One obviously important function of Polaris is to give the direction of the true north to any sailor or other observer situated north of the equator; another,

[1] Actually, of the north celestial pole, since Polaris does not exactly mark this point.

as we have seen, to give the latitude of all places from which it is visible. Most people who know no other astronomy are familiar with that group of northern stars known variously as the Plough, the Great Bear, and the Dipper; it is represented in Fig. 3. If the observer carries his eye along the line joining the stars marked β and α for a distance equal to about five times that separating them, he will find a star which, though of only medium brightness, is nevertheless rendered unmistakable by its isolation. This is Polaris.

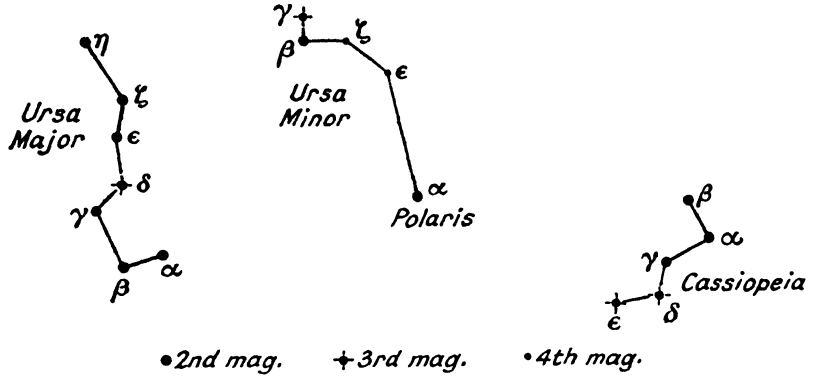


Figure 3. Three circumpolar constellations.

One feature following from the rotation of the heavens about the north celestial pole will soon be noted. The elevation of Polaris above the horizon, is, as we have seen, equal to the latitude of the observer. But this angular elevation is the shortest distance from Polaris to the horizon. Hence a star which is distant from Polaris by a number of degrees equal to the observer's latitude (51½ in the case of London) will never dip below the horizon in the course of its rotation about the pole. It will indeed just touch the horizon once in every diurnal period, and may therefore set if the northern horizon is not clear of obstructions. But all stars nearer to Polaris than this will never set at all—the whole of their circuit about the pole must be described above the horizon. Such stars are called circumpolar, and in our latitudes the stars of the Plough are an example of this class. All stars further from Polaris than the latitude of the observer will, of course, carry out a part of their circuit below the horizon, rising above it in the east and setting below it in the west. This distinction will be made clear by a glance at Fig. 4.

The reader will perhaps have been impressed by the fact that the star sphere has many points of resemblance to a model globe of the earth[1] such as is to be found in most schoolrooms. In the first place

[1] This statement does not beg the question of the earth's rotation; the demonstration of this fact will come later. A comparison is simply being made between the star sphere and the type of model mechanism of which a child's globe is an example.

the globe rotates between two pivots which correspond with the north and south celestial poles. Furthermore, if we imagine a plane to lie horizontally through the centre of the globe it will be appreciated that the pivot which we call the north celestial pole lies not immediately above it (i.e. at the zenith) but at a point intermediate between the zenith and the horizon. In the model we must call this section of the horizontal plane the north horizon. By rotating the globe and watching the parallels of latitude (which may be regarded

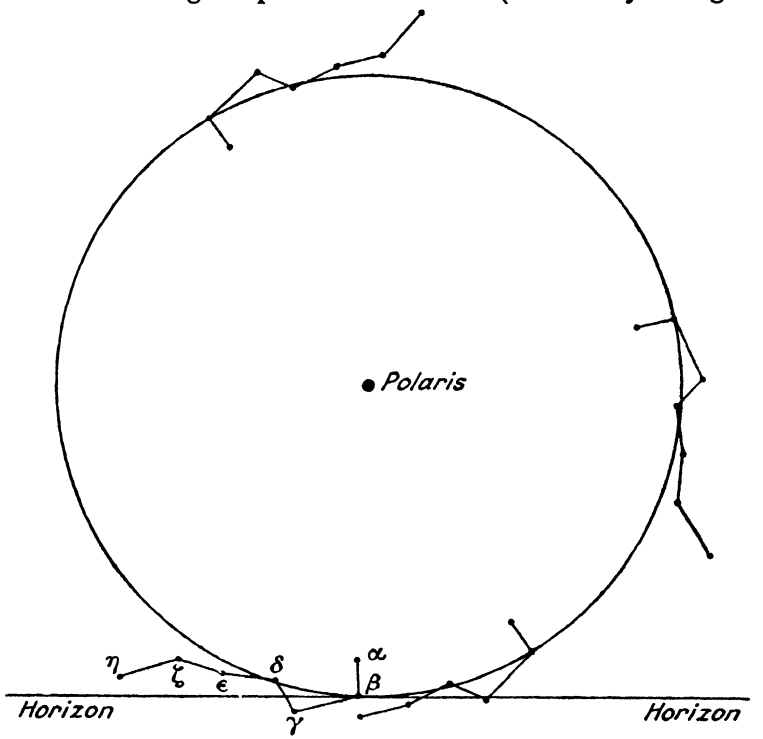


Figure 4. Circumpolar stars. The altitude of Polaris in the diagram (i.e. the radius of the circle) is 33°. The appearance represented is therefore that of the northern sky as seen from any place whose latitude is 33°, e.g. the Sea of Galilee. It will be noted that stars farther from the pole than this distance (η, ζ, ε, and γ) rise and set, and stars nearer (δ and α) are circumpolar, while β, whose angular distance from the pole is 33°, just grazes the horizon.

as the paths of certain stars in their diurnal rotation about the earth) the difference between circumpolar stars and stars that rise and set will immediately be made plain.

Midway between the pivots of the globe is drawn a great circle—i.e. a circle whose plane passes through the centre of the globe—known as the equator. Similarly, in the celestial sphere there is an imaginary circle drawn about the whole heavens, the visible section of which lies between the English observer's zenith and his southern

horizon. A moment's thought will reveal the fact that this circle, the celestial equator, will cut his horizon at points due east and due west of him, and will attain its maximum height above the horizon where it is due south (Fig. 5). By definition, the shortest distance of any point on the celestial equator from the north celestial pole (or the south

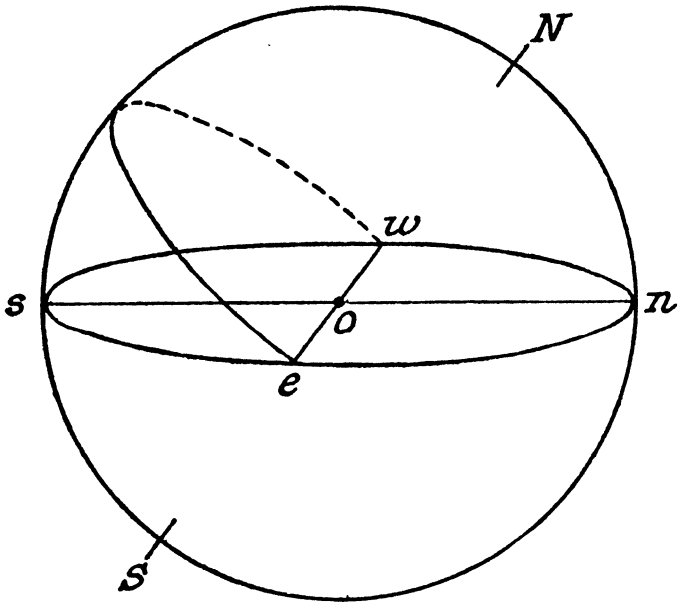


Figure 5. The circle *sSnN* represents the star sphere, bisecting which is the observer's horizon *senw*. The observer is situated at *O*, and *N* and *S* are the north and south celestial poles. It will be seen that the celestial equator, of which only the visible half is shown, cuts the horizon at points due east and due west of the observer, and reaches its maximum altitude due south of him.

celestial pole, which is below our horizon, for that matter) is 90°. In the same way all points on the terrestrial equator have a latitude of 0°, that of each pole being 90°. This is, of course, only another way of saying that the celestial equator bisects the star sphere into two hemispheres, in each of which one of the poles is centrally placed.

The sidereal day

Having noted the fact of the rotation of the star sphere about the earth, and having discovered the positions of the pole of this rotation and of the celestial equator, we must now discover the time required for one complete rotation of the sphere. The obvious way of doing this is to note the time at which some conspicuous star is due south—or, to express it in technical language, the time that it transits the meridian, the meridian being the great circle that connects the pole, the zenith and the north and south points of the horizon—then to

note the time of the next culmination, and to measure the time interval separating the two observations.

We will suppose that the observer produces a compass during the day preceding his first observation, and discovers that the point on his horizon that is due south of him is marked by a church spire. On the succeeding night he goes out at 8 p.m. and notes that there is a bright star slightly to the east of the vertical line passing through the steeple—that is to say, east of the meridian. He therefore waits, and at 8.30 precisely he decides that the star is directly over the steeple. He makes a note of the time and perhaps constructs a rough star map of the southern sky so that he will be sure of recognizing his star on the next occasion.

The following night he again begins observation at 8 p.m. and notes that the appearance of the southern sky is, so far as he can tell with the naked eye, exactly the same as on the preceding night. Not only are the stars in their same relative positions—that, as we have seen, is to be expected—but their positions in the sky, their compass bearings, appear to be the same. He therefore jumps to the conclusion that the star sphere rotates in a period of twenty-four hours. However, being of a scientific disposition, he decides to wait and see if his star really will culminate at 8.30, thus substantiating his guess. He waits less patiently this time, and when at last the star appears to be directly over the steeple he glances at his watch. He finds that the time is not, as he had expected, exactly 8.30, but about 8.26. Unfortunately, his disposition is not truly and sufficiently scientific. He comes therefore to one of two conclusions, perhaps to both. Since this observation is not in accordance with his preconceived notion that the star sphere rotates in exactly twenty-four hours (a guess based on comparatively inaccurate data, be it noted), then one or other of his observations was inaccurate and that the star was not *exactly* over the spire; or else his watch needs regulating.

But when he comes out on the third evening he is disturbed to find that the discrepancy between the expected and the observed is even greater. The star culminates at, according to his watch, 8.22 and at 8.30 it is quite obviously to the west of the spire. Clearly something has gone wrong, and as the result of a little thought he decides that his guess at twenty-four hours as being the exact period was slightly in error; further, that this error has now been doubled by the second rotation of the star sphere, thus becoming more easily detectable. He therefore begins all over again, and a week's observations carefully made and recorded show him where his mistake lay. The star sphere does not rotate in precisely twenty-four hours, but in a period about 4 minutes shorter, known as the sidereal day. Thus, if our

observer began his observations on Sunday night, and the star was then due south at 8.30, on Monday it would culminate at 8.26, and at 8.30 would be a short distance past (i.e. to the west of) the meridian; on Tuesday it would culminate at 8.22; on Wednesday at 8.18; on Thursday at 8.14. Since the star makes a complete circuit of the heavens in about twenty-four hours, it will travel an appreciable distance in a quarter of an hour and will therefore be very noticeably to the west of the meridian at 8.30 on Thursday night. It will, in fact, be to the west of the meridian by a distance equal to one ninety-sixth of its entire circuit of the star sphere.

Annual rotation of the star sphere

This discrepancy of four minutes *per diem* between the sidereal and the solar day has an important effect upon the appearance of the night-sky. At 8.30 p.m. on Monday all the stars in the southern sky will be slightly further west than they were at the same time on Sunday: stars that were on the meridian will now be slightly to the west of it; stars that were on the east horizon, now slightly above it; stars that were exactly on the west horizon will now have sunk below it and will be invisible. The difference in the aspect of the heavens occasioned by the passing of one day is of course small, since the diurnal shift accomplished in four minutes is only slight. But consider what happens as the result of a lapse of three months. The stars which were originally on the meridian at 8.30 p.m. are now on the west horizon at that hour; stars that were on the east horizon are now on the meridian; and a number of new stars which were invisible below the eastern horizon are now rising above it.

In fact, the star sphere makes one complete revolution in one year in addition to its diurnal motion. One year after his first observation, our observer will again notice that his star is on the meridian at 8.30 p.m. In the meanwhile every star that crossed his meridian between 8.30 on that original Sunday and 8.26 on the following Monday (only about half of which he could have seen, the others being above the horizon during the daytime) will have been on the meridian, or at any rate very near it, at 8.30 p.m. on one of the 364 intervening nights. In other words, the summer stars, those that occupy the southern sky on summer midnights, let us say, are not the same as the winter stars. (Note in passing that the circumpolar stars will have been visible on every one of the intervening nights, and for the whole of each of them.) Hence anyone familiar with the constellations, had he fallen asleep like Rip Van Winkle and awoken during the night, could know at once what the season was. If he heard a clock strike, thus telling him if it were early, midnight,

or later still, he would be able to make a pretty accurate guess at the month. And if in addition he were armed with a set of astronomical tables, he would be able to deduce the exact date—though not, of course, the year.

The motion of the moon

The next celestial body to be studied is the moon, and a single month's observation, for a few minutes each night, reveals the main characteristics of its motion and varying appearance. Further observations, made during subsequent months, will show that these features are recurrent or cyclic: the moon does the same things in the same order month after month.

The first thing that the observer will notice, and a few hours' or even minutes' observation are all that are required, is that the moon shares the stars' diurnal east-west motion. It rises in the east, though this may be cloaked by the fact of its rising in the daytime, and sets in the west, rising again in the east *about* twenty-four hours later. It will be remembered that the star sphere is, so to speak, rotating more rapidly than the passing of our terrestrial days (which are measured by solar, not sidereal, time); that every star arrives at the meridian, or any other selected point on the star sphere, a little ahead of solar time. Now the moon does somewhat the same thing, though in the opposite direction. In short, it is, as has been said, carried round the star sphere from east to west with the stars, but its motion appears to be slower. Thus the stars in its neighbourhood are always passing it, with the result that it has a residual west-east motion in respect of them.

Suppose, for example, that on Sunday at midnight the moon is due south of the observer. He memorises its position with reference to a bright star which is also on the meridian, but further south, i.e. directly below the moon. As he watches, he notices that the moon and the star (together with all the other stars) are moving across the sky towards the western horizon. But since, in this motion, the moon is proceeding more slowly than the star, he will on the next night observe that when the star is again on the meridian the moon has been left a certain distance behind. It is now considerably to the east of the star and he would say that besides its diurnal east-west motion the moon has a motion of its own from west to east, relative to the starry background. As a result of this motion the moon is displaced to the east at a given time on succeeding nights. Whether the observer regards the moon as travelling eastwards of its own accord while at the same time being carried westwards by the diurnal motion of the star sphere, or regards the moon as having no true eastward

motion but only as travelling westwards with the star sphere though more slowly than it, is immaterial. What is important to realize is that the moon takes more than twenty-four hours to make a diurnal circuit of the heavens, while the stars take slightly less.

The moon and the zodiac

The next thing to be noticed is that the moon, in its motion across the sky, is restricted to a comparatively narrow path: even the most astronomically ignorant person would be surprised to see the moon in the northern sky, for instance. This narrow path, which is situated between the English observer's zenith and his southern horizon, is called the zodiac; it is a parallel-sided belt, 18° wide, and out of it the moon (with, as we shall see, certain other bodies) never moves. Sometimes it is nearer one side of this highway, sometimes nearer the other, but it never jumps the curb.

The lunar phases

The third, and most interesting, point to be noticed with the naked eye is the cycle of the moon's phases and the time it requires to complete one eastward circuit of the star sphere. It will be remembered that owing to the fact of the stars revolving *en bloc* in slightly less than 24 hours, it requires a lapse of a year for a given star once more to be in the same position in the sky at the same time of night. Owing to the moon's considerable lag among the stars, it returns to the same position—say, the meridian—at the same hour after an interval of about twenty-nine and a half days.

This period is known as the lunar month, or lunation, and the moon's varying positions and appearances throughout it must now be described. For the layman, the lunar month starts when for the first time he sees the moon as a fine crescent low down in the west, soon after sunset. Two things are to be noticed: the moon is close to the sun in the sky, and the illuminated sickle lies on that side of the moon which faces the sun. Disregard of the latter fact has led artists to perpetrate absurdities times without number. As the month proceeds, the moon moves eastward along the zodiac. With its daily recession from the sun the proportion of its whole face which is illuminated increases. In a week it is on the meridian at about the hour that the sun sets; it is now half illuminated, or dichotomized, the illuminated hemisphere being that on the right-hand side, the side facing the sun. During the second week the moon continues its course away from the sun, and the terminator (i.e. the boundary between the illuminated and the unilluminated portions of its disc) moves further across its face from west to east (right to left). At the

end of the second week, midway through the lunar month, it is fully illuminated, and rises in the east at about sunset. Half way through the night, when the sun, below the horizon, has traversed half its course from the west horizon to the east, the moon will be on the meridian. As it sets in the west, the sun rises.

The moon is now as far from the sun, measuring angularly round the star sphere, as it ever can be: the two bodies are 180° apart, i.e. at diametrically opposite points on the sphere. In continuing its eastward motion, therefore, the moon begins to approach the sun again, but from the opposite side. During the first half of the month (new–full) it has been moving away from the *east* side of the solar disc. Henceforth (full–old) it is approaching its *west* side. As it does so, the terminator, which has vanished at full, reappears as before at its west limb and moves eastward across its disc; but now the opposite hemisphere is illuminated—darkness on the west, light on the east—since the moon's other limb is now nearer the sun. After three weeks the moon is about 90° to the west of the sun: it is on the meridian at sunrise, setting towards midday. It is again dichotomized, the illuminated hemisphere, being that facing the sun, is on the left-hand or east side of the disc. During the final week of the month it approaches nearer and nearer to the sun, becoming an increasingly narrow sickle as it does so, rising above the eastern horizon later each night, and setting in the west later and later in the afternoon. Towards the end of the fourth week it is a very fine sickle, rising shortly before the sun. Finally, as its eastward motion continues, it becomes invisible in the sun's glare. Thus invisible, it passes the sun and a few days later is again visible as the new moon low down in the west shortly after sunset. Another lunar month has begun.

The motion of the sun

The investigation of the sun's motions and positions is complicated by the fact that when we can observe the sun we cannot observe the stars. Since the sky is itself a featureless waste it is a great convenience to have the stars, whose motions we have studied and can allow for, as reference points. It is easy, for instance, to determine the exact position of the moon throughout the lunation, for we only have to make a direct observation of it, and then plot the observed position upon a star map. In dealing with the sun, indirect methods of study must be used.

But no special observations are required to establish our first point; common experience and what we have already discovered of the motions of the stars provide the necessary data. We know that,

whatever the season, the sun is always due south at midday.[1] We also know that the same star is not always south at midnight. Each twenty-four hours it has moved a little more than once round the star sphere and the sum of these daily increments amounts to one complete circuit in a year. Combining this knowledge of the movement of the stars with the observed fact that the sun is always due south at midday, we see that the sun must have a motion relative to the stars, and, moreover, that it must complete one west-to-east revolution of the star sphere in a year. It also makes a diurnal apparent circuit of the earth, rising in the east and setting in the west. During this interval it will have moved across the (invisible) starry background approximately one three hundred and sixty-fifth of its complete circuit. Thus, like the moon, the sun has a proper motion among the stars, though a slower one, since it accomplishes in a year what the moon does in a month.

The ecliptic

The general fact of the sun's motion among the stars is thus not very difficult to establish. But a question will have been suggested to the observer which the practical difficulty already referred to makes less easy to answer immediately. He has discovered that the moon's possibilities of position upon the star sphere are strictly limited; as we have seen, it never moves off a track 18° wide—an angular distance about equal to that covered by a stick 9 inches long when held at arm's length against the sky. Since the star sphere appears to be curved—it looks like the inner side of a hollow sphere, not like a flat surface—the lines traced out by all celestial objects in the course of their motions are likewise curved. In watching the moon pursue its curved and circumscribed path across the heavens, the observer will have been struck by the fact that this is just about, as far as he can judge with the naked eye, the same path as that pursued by the sun during the daytime. It is in endeavouring to find out more about this interesting and as yet inexplicable correspondence that he will encounter the difficulty already referred to. He is therefore forced to deduce the position of the sun in relation to the stars, from its observed position in the day sky, by an indirect method. There are several such methods to choose from. That employed by the

[1] Strictly speaking, the sun we see is not due south at 12 o'clock on every day of the year. Owing to two factors of which we shall learn more later—namely, the obliquity of the ecliptic, and the eccentricity of the earth's orbit—the sun runs sometimes fast and sometimes slow; this 'error', which is known as the Equation of Time, may be as great as a quarter of an hour. An imaginary sun whose motion is uniform, the so-called mean sun, is therefore made the basis of civil time reckoning. And it is this mean sun which is due south at noon on every day of the year.

ancient Egyptian astronomers, though not particularly accurate, is easily understood. Since, the Egyptians reasoned, we cannot observe the sun and the stars simultaneously, and yet want to know the position of the former relative to the latter, we must do the next best thing, which is to observe them in the quickest possible succession. Consequently they systematically noted the last bright star visible near the east horizon before sunrise, and by its aid deduced the approximate position of the sun in relation to the star's invisible neighbours.

Thus the path of the sun across the star sphere can be mapped and our observer's suspicions are confirmed: the sun's path does in fact lie very close to that of the moon. This path is known as the ecliptic (which is, strictly, the path traced out by the centre of the sun's disc in the course of a year) and so nearly does it coincide with that of the moon that it is possible to trace out a zone 18° wide in which the ecliptic is centrally placed and out of which the moon never moves. This is the zodiac, which is, in fact, that strip of star sphere that extends 9° on each side of the ecliptic. The zodiac is divided into twelve sections of equal size, known as the signs of the zodiac. Since the sun completes one circuit of the zodiac in a year, it passes one month in each sign.

The sun and the seasons

There is still one further characteristic of the sun's yearly motion that must be mentioned. It is a fact of common experience that the sun is higher in summer than in winter; also that the days are longer in the former season than in the latter. The first fact can be used to determine the length of the year, and in primitive times was so used. Let the reader mark on the ground a north-south line and at its southern end fix a vertical stick. Now when the stick's sun-cast shadow lies along the line the time must be midday and the sun at its maximum height above the horizon. Between sunrise and this time it will have been rising progressively higher in the sky from the eastern horizon; from then until sunset it will sink lower and lower towards the western.

The passage of the seasons provides us with a rough and ready method of determining the length of the year, but the result obtained can never be more than very approximate, except by a fluke. A study of the stick's shadow, on the other hand, allows the determination to be made with some accuracy. Suppose that every day the reader makes a scratch on the ground to mark the apex of the noon shadow. If he begins his experiment at about Christmas-time he will find that the shadow is slightly shorter each day until it reaches a minimum

value on 21 June. Thenceforward until about 21 December it will grow progressively longer again. Since the seasons are regulated by the noonday height of the sun, and the length of the shadow gauges this with fair accuracy, we are now in a position to say that the interval from 21 December to 21 June is half a year and that the total number of days in a complete year is equivalent to double the number in this interval.

The gnomon

It is interesting to note in passing that this most primitive of astronomical instruments, the gnomon, is possessed of considerable

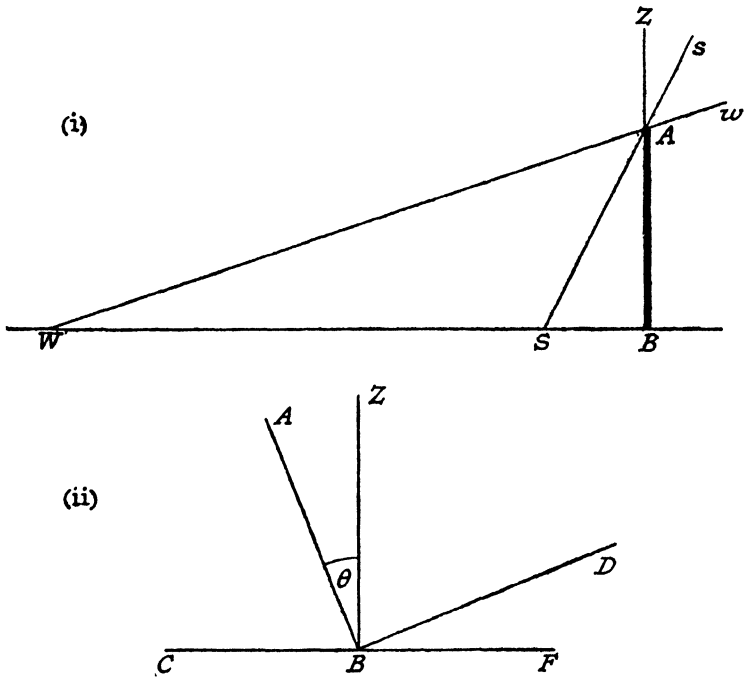


Figure 6. Determining latitude by means of the gnomon.

versatility in the hands of ingenious users. We have already seen that it can be used as a calendar, to measure the length of the year. The varying directions of the shadow on either side of the marked line record the passing of the hours, i.e. the time of day. It can be used to measure the obliquity of the ecliptic—a point to which we shall return. And, finally, it presents an alternative and approximate way of determining the observer's latitude. Suppose that W (Fig. 6 (i)) is the apex of the sun's longest shadow (on the shortest day), and S the apex at midsummer, when the shadow is shortest and the day longest. Both these distances can be measured, as can the height of

the style, *AB*. Since *WAB* and *SAB* are right-angled triangles, a simple trigonometrical operation will give the angles *WAB* and *SAB*.[1] But these angles = angles *ZAw* and *ZAs* respectively. These are clearly the sun's maximum and minimum zenith distances, and their mean equals the angular distance from the zenith to the celestial equator, since the sun is on the equator when midway between its maximum and minimum altitudes. (This fact will be made clear in the next paragraph.) We have thus succeeded in discovering the angle subtended between the celestial equator and the zenith. Referring for a moment to Fig. 6 (ii), *B* is the position of the observer, *CF* his horizon, *BA* the direction of the pole, and the angle *ABC* his latitude. *BD* is the direction of the celestial equator, the angle *ABD* being a right angle by definition. The angle *ZBD* is the zenith distance of the equator, which we have already discovered, and it is clearly equal to the angle *ABC* since both equal $(90° - \theta)$. Hence angle *ZBD* is the observer's latitude.

Obliquity of the ecliptic

It will immediately be asked how this annual bobbing up and down of the sun in the sky is connected with its west-east path among the stars. Having plotted the daily position of the sun on a star map throughout a complete year, we find that the ecliptic (its path) is neither coincident with the celestial equator, nor is it parallel to it. They are, in fact, related to one another in the same way as two hoops jammed together, one within the other (Fig. 7). Now since the latitude of any given point on earth is invariable, the altitude of the north celestial pole is also invariable. Thus the altitude of the point of intersection of the meridian and the celestial equator is the same on every day of the year. Since the sun, in its course about the ecliptic, is sometimes north of the celestial equator and sometimes south of it, its midday elevation above the horizon will vary from season to season. When it is at one or other of the points of intersection of the equator and the ecliptic—when, in fact, it lies on the equator—the days and nights will be of equal duration, the sun will rise due east and set due west (since we have seen that the celestial equator always cuts the observer's horizon at points due east and west of him), and its altitude at noon will be midway between its maximum (summer solstice) and its minimum (winter solstice). These points are called the equinoxes, and the sun arrives at

[1] Suppose, for instance, that $WB = 14$ ft., $SB = 2$ ft., and $AB = 4$ ft. Then

$$\tan \angle WAB = \frac{14}{4}, \text{ and } \tan \angle SAB = \frac{2}{4}$$

$$\text{Hence } \angle WAB \simeq 74°, \text{ and } \angle SAB \simeq 27°.$$

them on or about 21 March (vernal equinox) and 23 September (autumnal equinox).

The angle between the ecliptic and the celestial equator is about $23\frac{1}{2}°$ and is known as the obliquity of the ecliptic. It follows that when the sun is at its furthest from the celestial equator (at the

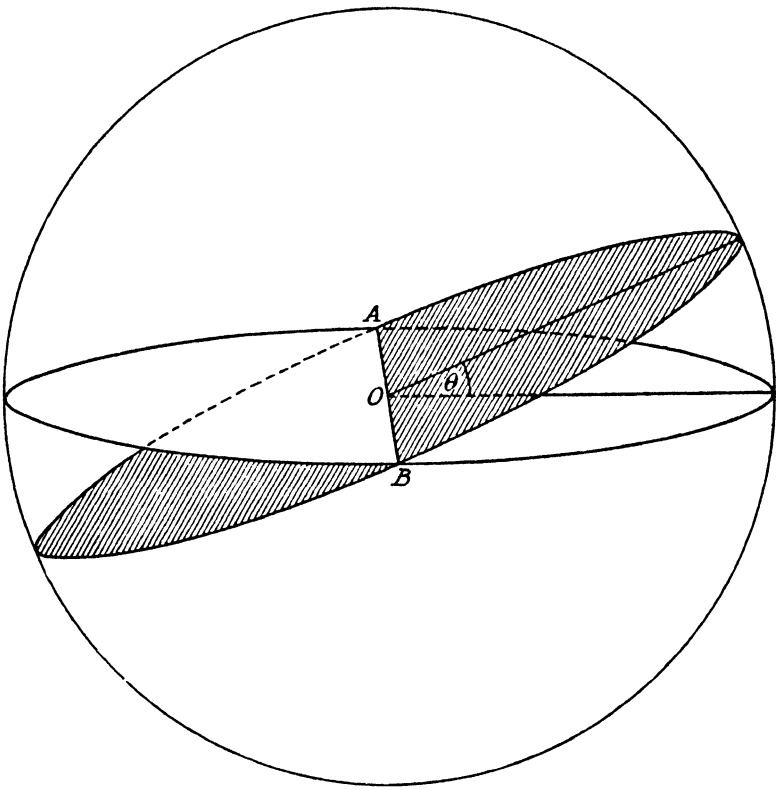


Figure 7. The circle represents the star sphere, seen from without. The intersecting great circles may represent either the celestial equator and the ecliptic, or the moon's orbit and the ecliptic. If the former, the points A and B are the equinoxes and the angle $\theta = 23\frac{1}{2}°$. If the latter, A and B are the nodes. In either case the earth is situated at O, the centre of the sphere.

solstices) its distance from it is $23\frac{1}{2}°$. Hence the midday sun at midsummer is 47° higher above the horizon than the midday sun at mid-winter.

It is quite simple for the reader to determine the value of the obliquity of the ecliptic for himself, by means of the gnomon. As before, we will assume for the sake of argument that AB (the height of the style) is 4 ft., WB (the length of the noon shadow at the winter solstice) 14 ft., and SB (the same quantity at the summer solstice) 2 ft. Now since the difference in the angular height of the sun above the horizon at the two solstices is double the angle subtended between

the ecliptic and the celestial equator, the angle *sAw* is twice as great as the obliquity of the ecliptic. But angle *sAw*=angle *WAS*, and this angle=*WAB*−*SAB*, both of which we have already determined (p. 21, n). They are, respectively, 74° and 27°. Hence angle *WAS* (=*sAw*)=47°. Hence the obliquity of the ecliptic=23½°. If the reader knows no trigonometry, he can simply make a scale drawing of, say, 1 in.: 1 ft., which shows the points *W*, *S*, *B* and *A*, and then measure the angle *WAS* with a protractor.

The planets

We now come to the last class of celestial object that repays study with the naked eye. When we made our observations of the stars we found that one of their most obvious characteristics was their immobility relative to one another. Thus, if we were to make a map of some conspicuous star group (the Great Bear, say) we should find that year after year it would be as accurate as on the day it was made. Yet if the reader observes regularly he will sooner or later discover that there are exceptions to this rule: there are stars (or so they seem) which not only change their positions relative to the vast mass of their neighbours but also relative to one another. These objects are called planets, and five are visible to the naked eye.[1] The word 'planet' is connected philologically with the idea of a wanderer. In appearance the planets are indistinguishable from stars; only their motion, which requires serial observation for its detection, can infallibly betray their different nature to the naked eye. It is widely supposed that whereas on some nights the stars twinkle, the planets never do. But this is an unsure guide, and it is extremely doubtful if anyone unfamiliar with the appearance of the night sky could distinguish stars from planets by this means alone.

Motions of the outer planets

Careful observation of the successive positions of the planets reveals the fact that they may be divided into two groups, the members of which behave in quite different ways. We will consider the planets belonging to the larger group first. They are three in number and to them the names Mars, Jupiter and Saturn have been given. Like the sun and moon they travel eastward among the stars, but their motions differ from those of the brighter bodies in two important respects:

i. they are very much slower;
ii. they are not smooth, but erratic.

[1] A sixth, Uranus, is just visible to the naked eye when its position is known beforehand. It was not known to the ancients, however, and was discovered telescopically by Herschel in the eighteenth century.

Taking the second point first, the motion of a planet belonging to this group may be described as follows (Fig. 8). When first observed we will suppose that it is moving eastward with increasing velocity. After an interval, the length of which varies from planet to planet, it begins to slow down, finally becoming stationary among the stars. Then it begins to move in the reverse direction (westward), accelerating at first, then slowing again, until once more it is stationary. Lastly it resumes its eastward motion. Thus the planets Mars, Jupiter and Saturn move eastwards in a series of loops; since more ground is always covered by the direct motion than by the retrograde

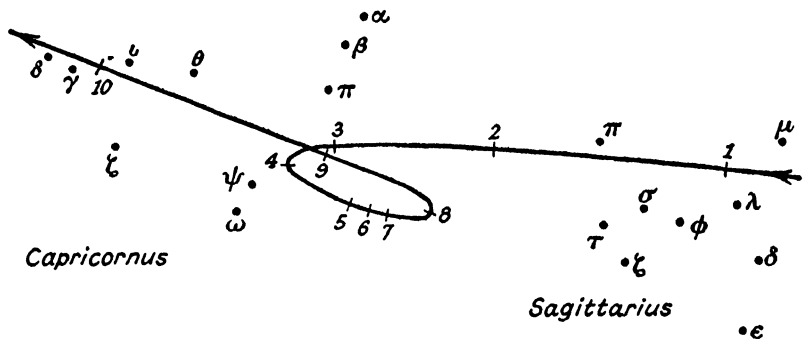


Figure 8. Successive positions of Mars in the constellations Sagittarius and Capricornus during 1939:

1. 1 April	6. 27 July
2. 1 May	7. 1 August
3. 1 June	8. 1 September
4. 1 July	9. 1 October
5. 23 July	10. 1 November

During July and August the planet was retrograding.

there is a residual eastward motion. Observations over a number of years permit the determination of the period required by each planet for one circumambulation of the star sphere. It is found that for Mars this is about one and three-quarter years, for Jupiter about twelve years, and for Saturn about twenty-nine years.

Motions of the inner planets

The second planetary group has two members, Venus and Mercury. As with all the other bodies we have so far considered, the motions of these planets are recurrent. To illustrate them, we shall describe a single cycle in the motion of Venus. Let us suppose that when first observed, Venus is situated in the south-west sky in the evening; it is visible for several hours after sunset before following the sun below the horizon. Nightly observations will show that at a given time each evening Venus is a little nearer the horizon. Thus the period of its visibility is steadily diminishing, since the sun's

glare does not permit it to be seen till some time after sunset. Eventually the planet has moved so near the sun in the sky that it is not visible until on the point of setting. Last of all, it becomes entirely invisible, for by the time that the sky is dark enough for its visibility, it has already set.

The observer need not feel stultified by this disappearance, for the situation is not without precedent in his experience. In his observations of the moon he encountered an analogous case: once in each lunar month the moon approaches so near that part of the sky occupied by the sun that it becomes invisible (new moon), appearing on the other side of the sun after a suitable lapse of time. It will obviously be worth scrutinizing the eastern horizon before dawn in the hope that Venus too will pass through the sun's glare and reappear on its western side. The observer's foresight and patience will eventually be rewarded. Low down in the east he will notice, shortly before dawn, a shining point of light which in a few minutes is rendered invisible by the rising sun. As the days pass he will notice that Venus is slightly further from the sun when first seen and is consequently visible for a longer period of time. Day by day its angular distance from the sun increases.

Unlike the moon, however, Venus does not recede continuously from the sun, eventually to approach it again from the opposite (eastern) side. When it has receded about 48° from the sun it stops in its westward progress and begins to retrace its steps towards the sun's west side. Eventually it is lost in the morning glare just as formerly, when an 'evening star', it was lost in the glare of the setting sun.

Thus Venus swings back and forth from the east of the sun (when it is visible as an 'evening star') to the west ('morning star') and back to the east once more. This cycle is endlessly repeated. Mercury's motion is of exactly the same type. It differs from Venus in two respects: it never recedes as far from the sun as its brighter neighbour (the maximum distance between Mercury and the sun is only 28°), and it completes its cycle of changes in a shorter period of time. On account of the first point of difference it is in practice very much more difficult to observe than Venus. Most people are familiar with the appearance of Venus as the 'evening star', but few who have not deliberately set out to do so have ever caught a glimpse of the fugitive Mercury.

Thus we see that neither Venus nor Mercury can ever be in opposition to the sun, as can the planets Mars, Jupiter and Saturn.

The planets and the zodiac

Lastly, let an important point be noted. All the five planets are

confined to the zodiac. Like that of the moon, their paths are neither coincident with nor parallel to the ecliptic, but are inclined to it at small angles (all less than 9° since they never move out of the 18°-wide zodiac) in the same way that the ecliptic is inclined to the equator.

Recapitulation

In the present chapter we have reviewed the main characteristics of the motions of the sun, moon, stars and planets as they may be discerned by elementary naked-eye observations. A number of facts have been omitted. The sun and moon do not, for instance, move absolutely uniformly in their courses, though the irregularities in these motions are neither so noticeable nor so important as those affecting the planets. Again, Polaris has not always marked, and will not always mark, even the approximate position of the north celestial pole. But to explain the growth of our astronomical knowledge step by step and yet to keep the account within manageable proportions one must inevitably omit less important points. All those facts that are stepping stones to the demonstration of further astronomical essentials have been included.

These leading facts may be summarized as follows:

1. The earth is approximately spherical and its size can be determined with great accuracy.

2. Every object in the sky makes one complete east-west circuit of the star sphere in *about* twenty-four hours. This circuit is known as the diurnal revolution of the body in question.

3. The star sphere rotates as a coherent body between two poles—the north and the south celestial poles.

4. The stars are mere points of light and do not change their positions relative to one another—at any rate not noticeably within the period that can be covered by the observations of a single man.

5. The period of the diurnal rotation of the star sphere being slightly shorter than twenty-four hours of solar time, the stars appear to be moving steadily westward, one complete circuit being accomplished in a year.

6. Besides its diurnal east-west motion, the moon has a monthly west-east motion. During the month it passes through a cycle of phases that has been described.

7. All the planets and the moon are confined, in their motion across the star sphere, to a zone 18° wide. It is known as the zodiac, and placed centrally within it lies the ecliptic, i.e. the yearly path traced out by the centre of the sun's disc.

8. The ecliptic is inclined to the celestial equator at an angle of $23\frac{1}{2}°$. The fact that the sun is therefore carried to either side of the celestial equator originates the terrestrial seasons, the sun being higher in summer than in winter.

9. The sun makes its west-east circuit of the star sphere in one year. (This is the same as 5, looked at from the reverse viewpoint.)

10. Besides partaking of the east-west diurnal motion of the star sphere, the planets have proper motions of their own among the stars. In this respect they may be clearly divided into two groups:

i. Mars, Jupiter and Saturn. Their paths are looped, i.e. at times they are travelling, not eastward, but westward. Nevertheless, their eastward motions always more than counterbalance their periodic retrograde motions.

ii. Venus and Mercury. They oscillate from one side of the sun to the other, but do not exhibit the looping that characterizes group i.

11. i. Jupiter requires a longer period for its circumambulation of the star sphere than Mars, and Saturn than either of them.

ii. Mercury never recedes from the sun to so great a distance as Venus, and completes one oscillation in a shorter period of time.

The problem is to find a single hypothesis which explains this variety of observed appearances and/or facts.

II

THE PROBLEM SOLVED: REAL MOTIONS

THE most obvious explanation of these phenomena is that the heavens as a whole revolve about a stationary earth in slightly under twenty-four hours, and that the sun, moon and planets, while partaking of this motion, have in addition motions of their own, for the most part in the reverse direction. This possible explanation being the first to hand, it will be convenient to develop it in ways suggested by a more detailed consideration of the visible behaviour of the sun, moon, planets and stars, and to see whether it can be made to render a coherent and reasonable account of these.

The Ptolemaic universe

Such an investigation was made by Ptolemy (*fl. c.* A.D. 130), working along lines suggested by Hipparchus some three hundred years earlier. Ptolemy made three basic assumptions, the first two being dependent upon the obvious evidence of the senses and the third upon fallacious *a priori* reasoning:

i. Since the celestial bodies appear to revolve about the earth, they do so revolve.

ii. Since the earth appears to be motionless, it is motionless.

iii. The circle is the most perfect figure, and therefore the celestial bodies must move in orbits that are either circles or else are compounded of circles. Furthermore, not to detract from this mystical perfection, their motions in their orbits must be uniform —they must neither accelerate nor decelerate.

Centrally placed in the universe lies the motionless earth. On the confines of the universe are the stars. These are regarded as being equidistant from the earth and embedded in a giant sphere centred upon the earth, the star sphere. It is the rotation of this sphere which causes the diurnal motion of the stars. Within the space between the earth and the star sphere are situated seven other spheres,[1] each of which bears an orbit[2]—either of the sun, the moon, or a planet. That the more distant bodies may not be obscured by the spheres carrying the nearer, these spheres are regarded as composed of

[1] The crystal spheres were, strictly speaking, a feature of the Aristotelian system (which Ptolemy was seeking to reconcile with the observational facts), though he never explicitly disclaims them.

[2] Actually, in most cases, a deferent: *vide infra.*

perfectly transparent crystal. We have already seen that the sun and moon pursue direct, unfaltering courses across the heavens. Hence these bodies may be regarded as moving with a uniform velocity about circular, circumterrestrial orbits.

To explain the periodic retrograding of Mars, Jupiter and Saturn in terms of uniform circular motion, Ptolemy was forced to introduce epicycles. The planet does not itself move in a circular circumsolar orbit but about a second orbit, smaller but also circular; the centre of

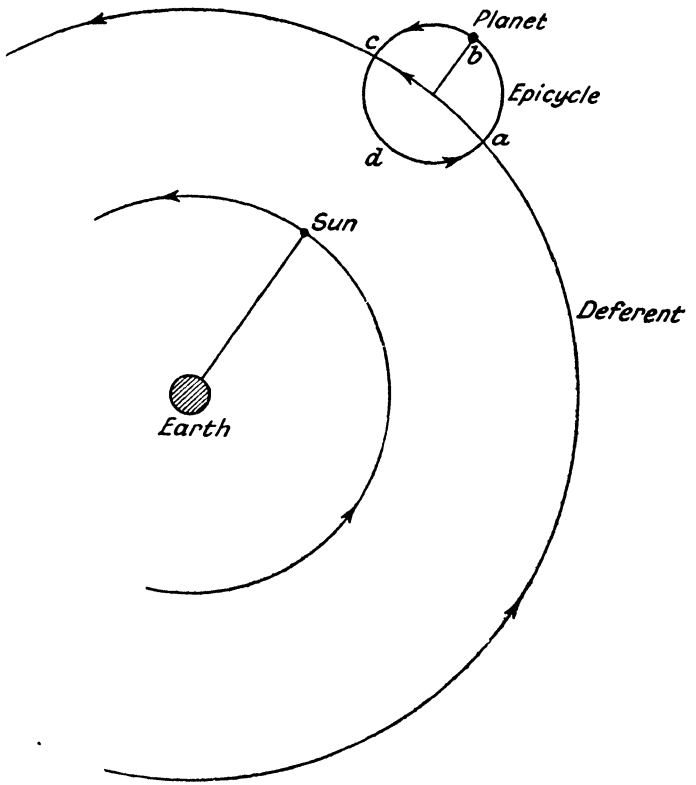


Figure 9.

this epicycle lies on the main orbit, or deferent, and travels uniformly round it (Fig. 9). Furthermore, the radius of the epicycle which connects each of the planets with its fictitious planet (i.e. the centre of the epicycle—the point that travels round the deferent) must be parallel to the line joining the earth and the sun.

A moment's thought, and the study of Fig. 9, will show that such an arrangement will in fact impart to the planet an apparently looped motion, despite the fact that it is moving with uniform velocity in an orbit composed entirely of circles—the 'perfect' figure. In copying

in this way the *particular* motion of each individual planet, Ptolemy and subsequent astronomers experienced a good deal of difficulty, but it is easy to see that in general the plan will work. The planet's motion in that part of the epicycle marked by the letters a—b will be observed as acceleratingly direct; from b—c deceleratingly direct; at c it will appear to pause momentarily, and then from c—d to move with increasing velocity in a retrograde direction; from d—a its velocity will be on the decrease again, and at a itself the planet will be momentarily stationary. Then once more it begins its direct eastward motion.

To see how Ptolemy accounted for the observed motions of Venus and Mercury, we must first comment upon his estimation of the relative distances of the various celestial bodies from the earth. Of two bodies travelling with the same velocity, that which is nearer the observer will appear to be the faster moving. Thus, to an observer sitting on an esplanade seat, the people strolling past at perhaps 2 m.p.h. move further across the field of vision during a given interval than a ship on the horizon travelling at 2 m.p.h.—or even 20 m.p.h.

We have seen that, of all the celestial bodies, that which makes a complete circuit of the heavens to its original position in the shortest time, is the moon: it circuits the star sphere in one month. Of those bodies that appear to swing back and forth on either side of the sun, Mercury moves more quickly than Venus and has the apparently smaller orbit; both of them complete one cycle from elongation to elongation in less than a year. Next comes the sun, which requires a full year to circuit the heavens. Then Mars, Jupiter and Saturn in that order. Ptolemy therefore concluded that the moon was the nearest body to the earth; then Mercury, then Venus, then the sun. Now since Mercury and Venus are limited in the distance that they may retreat from the sun, Ptolemy decided that the motion of their epicycles must be such that it is always possible to join their centres, the earth and the sun by a single straight line. Their oscillations on either side of the sun would then result simply from their epicyclic revolution, one complete revolution being accomplished in the observed period. Fig. 10 will show that if this condition is fulfilled, their epicyclic motions will in fact cause their observed motions. It will also demonstrate the interesting fact that Venus and Mercury, if they shine only by reflected light, can, under the Ptolemaic dispensation, never appear more than half illuminated to an observer on earth.

A further device that Ptolemy was forced to introduce—tied as he was by the twin considerations of explaining the observational facts whilst invoking none but 'perfect' circular motions—

was that known as movable eccentrics. This highly artificial arrangement consists of a planetary orbit whose centre does not coincide with that of the earth, but lies always on the line joining the earth and the planet in question. Thus the earth was not, after all, at the

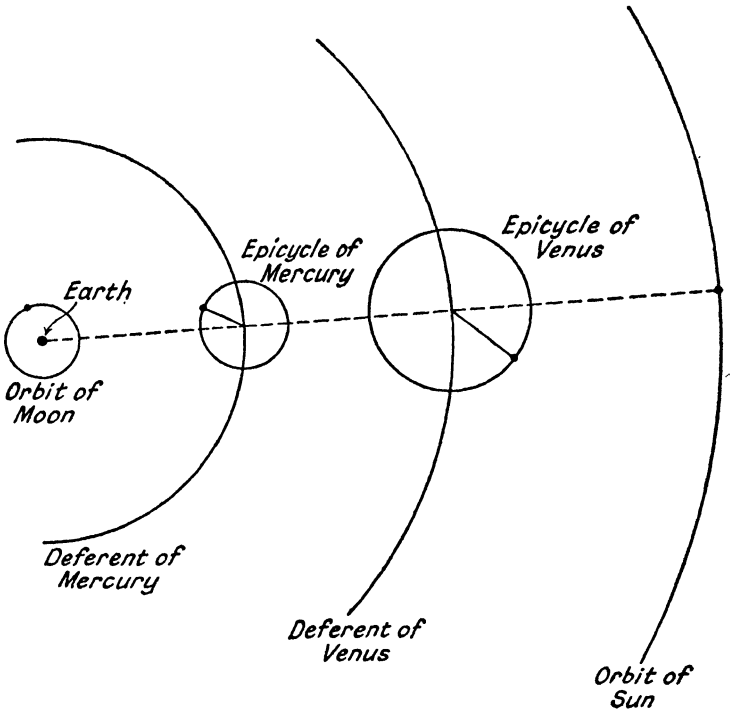


Figure 10.

centre of the universe, though it was admittedly nearer to it than any other body.

Ptolemaic difficulties

This, in general outline, is Ptolemy's geocentric theory of the structure of the solar system. Up to a point it works admirably. But as it became possible to make more and more accurate measurements of the positions of the sun, moon and planets it was found that the model, as it stood, did not exactly represent the observed phenomena. That, in fact, the sun, moon and planets do not move exactly as they should were Ptolemy's account of their relative positions and orbits correct. Consequently, the theory had to be patched up, and the more accurate astronomical observation became, the more drastic had the patching to be.

First, it was found sufficient to elaborate and proliferate the

movable eccentrics, applying them also to the epicycles. But the tale of woe did not end there. Still the observed positions of the planets did not agree with the theoretical positions (derived from the application of Ptolemy's hypothesis), and it became necessary to add further epicycles. Instead of a planet moving round an epicycle moving round a deferent—which many people will think a sufficiently complicated and unnatural arrangement—the planet was supposed to move round an epicycle which itself moved round an epicycle which in turn moved round an epicycle . . . which moved round a deferent. Soon the number of epicycles in the system, if it were to work, had mounted to eighty, and the geocentric hypothesis had become one of literally inconceivable complexity.

More fatal to its chances of being true than its superficial complexity was the fact that its complexity was not of the type that rests upon a fundamental simplicity or unity—it was impossible to formulate any general laws regulating the behaviour of the bodies concerned. Each individual body had to be given individual attention, and it was not possible to relate the corpus of particular variations and idiosyncrasies to a single elementary law. To account for the observed facts of each planetary motion it became necessary to employ a separate and highly complicated piece of juggling with the relative lengths of the radii of the epicycles (with additional new epicycles if required), with the velocity of the planet about its epicycle and of the epicycles round one another and round the deferent, and also with the degree of eccentricity of the epicycles and deferent. And then the publication of a newer and more accurate set of observations would necessitate still more juggling, still more 'saving of the phenomena'. That the mechanics of the heliocentric cosmology—to be described presently—are based upon a single, simple law (as Newton showed) and that the apparent complexity is due to the manifestation under varying conditions of this single law, is the most impressive earnest of the truth of the heliocentric hypothesis.

The geocentric hypothesis refuted

Since this is not a strictly historical survey, it is unnecessary to follow step by step the overthrow of the geocentric hypothesis. It will be enough to consider certain of its essential implications, and then to show that these are not substantiated by observed fact. For this purpose, the following points will be sufficient:

i. The solar system consists of a number of nearly concentric crystal spheres which contain the sun, moon and planets or their deferents.

ii. Neither Venus nor Mercury, supposing they were large enough to have sensible discs (the telescope was, of course, not invented till long after Ptolemy's day) could ever show more than a single hemisphere illuminated by the sun.

iii. The earth is motionless upon its axis.

iv. The earth has no orbital motion; on the contrary, it is the sun, together with the rest of the solar system, which revolves about it.

The crystal spheres

Points (ii), (iii), and (iv) are essential to the hypothesis. If it can be demonstrated that they are false, the hypothesis falls to the ground. Point (i) is not essential since there is no evidence to show that the deferents are not unsupported in space. Nevertheless, it was this point that was first called into question, and its failure to stand up to examination marked the beginning of the overthrow of the entire hypothesis. In the sixteenth century the Danish astronomer Tycho Brahe was able, from his observations of comets, to demonstrate conclusively that these objects pass freely to and fro through the space supposed to be occupied by the crystal spheres. The spheres as material objects, clearly did not exist.

The phases of Venus

More accurate observations—this time at the hands of Galileo armed with the recently invented telescope—then showed that Venus passes through a complete series of phases exactly similar to those of the moon; the same was later shown of Mercury. It is clear again, that Ptolemy was in error, at least as regards the relative positions of the earth, the sun, Venus and Mercury.

Rotation of the earth

But the foundation of the geocentric system (disregarding the irrational assumption that the circle is the 'perfect' figure and that therefore the planetary orbits must be circles or composed of circles) is that the earth is motionless. It is this conception that makes the whole complicated system of epicycles and circumterrestrial motions necessary. Once it is shown that the earth moves—that it rotates on its axis, or revolves about the sun, for instance—the basic assumption of the hypothesis is demonstrated to be false. Thence forward we may legitimately look for a hypothesis that gives a more natural and universal explanation of the facts.

In 1851 Foucault devised an experiment which establishes once and for all the fact that the earth rotates on its axis. If a heavy

weight is suspended by a considerable length of wire, no rotation of the support to which the wire is attached will alter the plane in which the weight is set swinging. Suppose that the wire is fixed to a movable beam in the roof of a lofty hall and that the weight is set swinging accurately north and south. Suppose, too, that at the outset of the experiment the beam also lies along a north-south line. Then if it is rotated slowly through 90°, so that it is now lying east-west, it will be found that the pendulum is still swinging north-south. Further rotation of the beam has no more effect: however the support is rotated, the pendulum always swings in the plane in which it is started. The reason for this is that it is easier for the wire to twist[1] than for the heavy weight to alter its direction of swing.

Here, then, in the immovable plane of the swing of a freely suspended pendulum, we have something not provided by the stars—a reference system which we know to have no motion of its own. When we see the stars apparently moving round the earth there is no *a priori* means of knowing whether the stars are really moving from east to west or whether the earth is rotating from west to east: the two motions would result in the same appearance. But the pendulum always swings in the same plane in space, and therefore if we find that the earth appears to move relative to it (or vice versa, it is the same thing) we know that it really is the earth, and not the plane of swing, that is changing its position.

Foucault used a heavy iron bob and suspended it from the 200 feet high roof of the Panthéon. To the lower side of the bob he attached a pointer of just the right length to scratch a groove in a tray of sand on the floor as it swung; in this way the direction of each swing could be recorded. After the pendulum had been swinging a few minutes, the startling discovery (it would at any rate have startled Ptolemy) was made that the direction of the swing was slowly changing relative to the walls of the room: its plane was rotating. But we know that this cannot be so. Hence it must be the earth that is rotating and carrying the Panthéon and the tray of sand with it. If the pendulum were suspended at the north pole it would be found that its plane of swing would require twenty-four hours to make one complete rotation. The earth rotates on its axis once in twenty-four hours.

This is conclusive proof that the apparent diurnal rotation of the star sphere about the earth is due to the earth's rotation in the reverse direction, from west to east. Two other experiments which confirm this result, one of them based on an entirely different principle, may be noted in passing. The first was suggested by Newton in 1679.

[1] If the bearing is frictionless, or practically so, even the torsion of the wire is eliminated.

If the earth is at rest and a weight is dropped from a height—down a deep mine shaft, let us say—it will strike the ground vertically below its point of release. But suppose that the earth is rotating. In that case the head of the shaft will have an infinitesimally greater velocity imparted to it than the bottom, since it is further from the centre of rotation. Hence the weight, since in falling it will retain its initial eastward velocity, will land at a point very slightly to the east of that vertically below the point of release. Experiments have shown that this is so. This affords another ocular proof of the rotation of the earth from west to east.

The third experiment was also suggested by Foucault, and in theory is identical with that of the pendulum. Just as the direction of swing of the bob remains constant whatever the motion of the suspension, so the axis of rotation of a gyroscope is independent of its support. And a spinning gyroscope will be found slowly to rotate. Since, however, it cannot be the gyroscope that is rotating, it must be the earth.

It is clear, therefore, that we are not compelled to hypothecate a stellar universe rotating daily about a central and stationary earth. A spinning earth will give exactly the same appearance of rotation to the star sphere, though, necessarily, in the reverse direction; everyone is familiar with the illusion that his train, standing at a platform, is beginning to move forward when actually it is the train on the other line that is moving out of the station in the opposite direction.[1] Not only is such a conclusion a great advance upon Ptolemy's from the viewpoint of simplicity and economy of means, but it is categorically forced upon us by the experiments just described.

Effects of the terrestrial rotation

There is a further aspect of terrestrial motion which we have not yet considered: the earth, we have learnt, rotates on its axis in a period of twenty-four hours; does evidence exist of motion in any direction other than this? Since the earth undeniably feels and appears to be at rest, yet equally undeniably is in motion about its axis, we need feel neither outraged nor surprised should we discover any such evidence.

The ancients, refusing to believe that the apparently solid, unmoved and immovable earth could be in rotation, were forced to read the effects of this rotation into the external universe. Let us

[1] It is interesting to note that it always appears to be the observer's train which is in motion, whereas in celestial matters it is the observed object and not the observer's station that is judged to be moving.

reverse this process, deducting the effects of the earth's rotation from the apparent motions of the heavenly bodies, and see what is left. Let us, in fact, suppose that the earth's axial movement is arrested, and that (for simplicity) this occurs on either 21 March or 23 September, the vernal or the autumnal equinox.

Thenceforward the sun will be alternately above and below the horizon for six-monthly periods. It will proceed steadily and slowly eastwards across the sky, spending, as we have seen, one month in each of the signs of the zodiac. Only its diurnal east-west motion will have been eliminated. In the same way, the moon and planets will perform their west-east motions, the moon passing through its phases and the planets pursuing their looped or back-and-forth motions, undisturbed by daily risings in the east and settings in the west.

It will be seen that we have only eliminated the appearance of the *daily* rotation of the heavenly bodies about the earth, not their slower rotation—yearly in the case of the sun, longer for Mars, Jupiter and Saturn. The question we must now try to answer is, is *this* motion really inherent in the bodies concerned or is it the effect of another and as yet undiscovered terrestrial motion? Are these bodies really revolving about the earth in their different periods?

The aberration of starlight

The final and ineluctable answer to these questions was provided by Bradley, the then Astronomer Royal, in 1727. It is a commonplace that although a heavy downfall of rain, falling vertically on a windless day, strikes a man on the top of his head as long as he stands still, yet it strikes his face once he starts to walk or run forward. Everyone who has tilted his umbrella forward as he hurries for shelter is well aware of this fact. Again, the vertically falling rain seen from inside a stationary railway carriage does indeed appear to be falling vertically; but when the train is in motion the falling drops are seen to cross the window space diagonally from the top corner nearest the engine towards the bottom corner nearest the guard's van (Fig. 11 (ii)). And the faster the train, the further from the vertical does the direction of the downpour appear to be displaced.

A different example will help to show why it is that the apparent direction from which the rain is coming is deflected towards the direction of the observer's motion. Fig. 11 (iii) illustrates a rigid tube standing upright in our vertically-falling rain storm. A drop entering the centre of its mouth will fall axially down the tube and strike the base centrally. Now let us suppose that during the time that the drop is falling from X to the bottom of the tube, the tube

itself has moved horizontally through a distance equal to half its diameter. While the drop is falling from X towards C, C has moved away to its right and its place has been taken by D. Consequently the

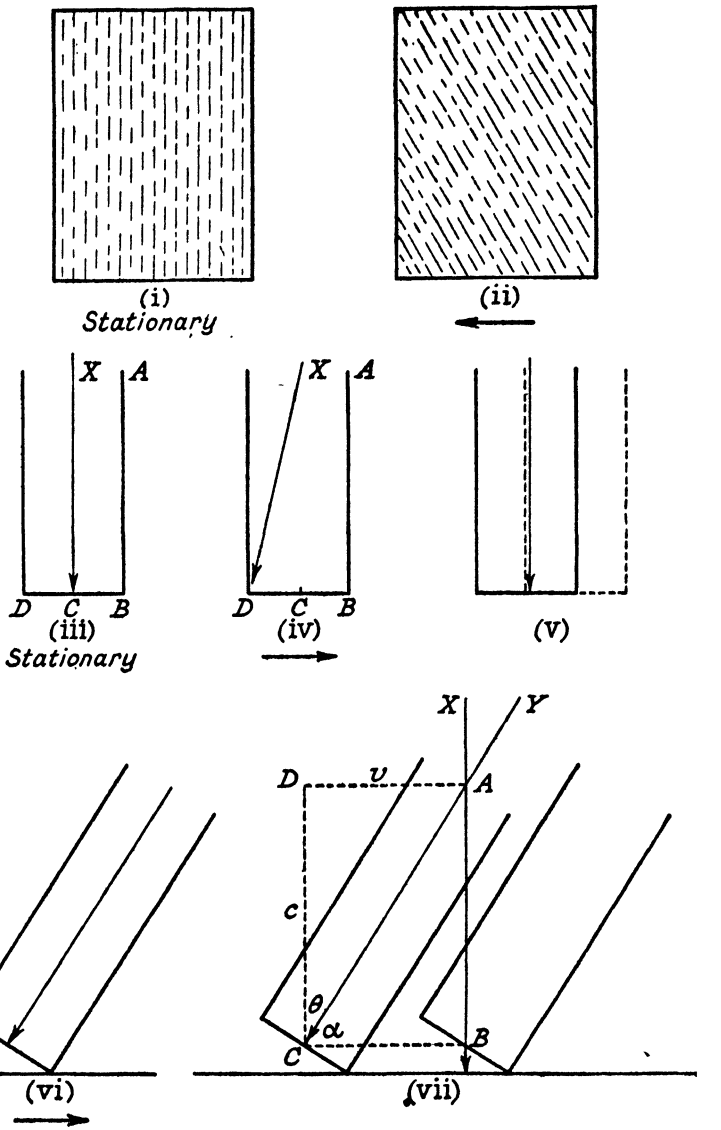


Figure 11.

drop will have fallen slantwise down the tube from X to D (Fig. 11 (iv), (v)). If we wish to prevent this slantwise motion of the drop within the tube, while retaining the tube's horizontal motion, we must tilt the tube over until its sides are parallel to XD, as in Fig. 11

(vi); in other words we must draw back the bottom of the tube by an amount equal to the tube's horizontal motion during the interval required by the drop to pass down it. This is of course what the pedestrian does when he tilts his umbrella forward to protect his face against the rain which is falling vertically relative to the ground, but slantwise relative to him.

It can now be seen that the more rapidly the tube is moved across the direction of the rainfall, the further back must the lower end of the tube be drawn in order to keep the drops travelling within it parallel to its axis. In other words, the apparent displacement of the direction from which the rain is coming depends upon the relative speeds of the tube in the horizontal plane and the rain in the vertical plane. In Fig. 11 (vii) the drops fall from A to B in the same time that the tube moves from C to B, the drop thus being kept in the axis of the tube. Clearly then, the distances AB and CB are proportionate respectively to the velocity of the rain and of the tube. If we write c for the former and v for the latter, we have

$$AB=CD=c,$$
$$CB=AD=v,$$

and α, the tilt of the tube, is given by

$$\tan\theta=\frac{v}{c}$$

showing clearly that the apparent displacement of the direction from which the rain is falling depends on the relative velocities of the tube and the rain. When c is very large compared with v, the displacement will be small, and vice versa.

The earth's orbital motion

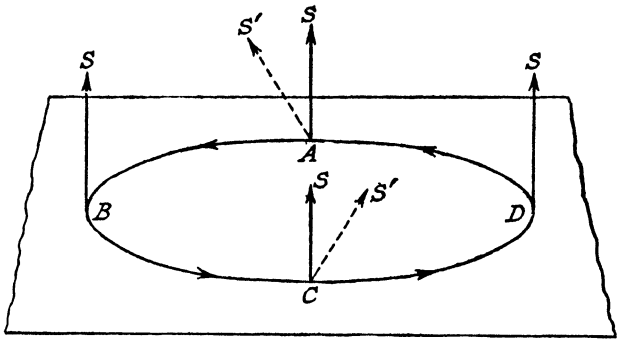
The reader may well be wondering what all this has to do with the astronomical status of the earth. The connexion is this: if for our tube we substitute a telescope, and for the rain drops a ray of light, we encounter an identical phenomenon. Just as the motion of the tube, or of the walker in the rain, makes rain coming from X appear to come from Y, so motion of the telescope towards B makes the light from, say, a star whose true direction is BX appear to come from the direction CY; in other words, the star appears to be displaced in the same direction as the observer's motion.

Such an effect can only result from the combined motions of what for the moment we will call the light rays, and of the telescope. The Danish astronomer Römer had demonstrated as early as 1675 that light travels with a finite, though very great, velocity.[1] But what of

[1] See p. 199.

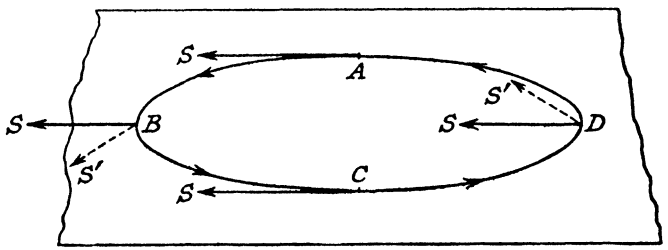
the required motion of the telescope? Since it is at rest relative to the earth, the earth itself must necessarily be moving. Let us suppose for the sake of argument that the earth is travelling in a circular orbit about the sun; then deduce from our previous experience of the tube and the umbrella what the nature of the resultant stellar displacements would be; and, finally, compare these with the observed displacements. If the two tally, we have a convincing proof of a terrestrial motion in an orbit about the sun.

Figs. 12 (i) and (ii) represent this hypothetical orbit, at the centre



(i)

All arrows lie in planes perpendicular to that of the orbit.



(ii)

All arrows lie in the plane of the orbit.

Figure 12.

of which (not shown) lies the sun; the four positions, *A*, *B*, *C* and *D* are those which the earth would occupy at three-monthly intervals. To be quite sure of our ground, let us once again refer to the results of our experiments: (i) if the motion of the telescope is inclined to the light rays, the source of the latter will appear to be displaced in the direction of the telescope's motion; (ii) if the telescope moves in the direction of the light waves, the source will suffer no displacement; if, in Fig. 11 (iii), the tube were moved vertically upward or downward, the drops would still fall parallel with its axis.

Returning to Fig. 12 (i), we will suppose that there is a star in the

direction AS at right angles to the plane of the orbit; to make the diagram clearer, the plane is drawn in, though it, of course, has no material existence. Owing to the phenomenon of aberration, the earth moving towards B, the star will appear to lie in the direction AS'; at B it will appear to be displaced outward from the page since the earth at this point is moving towards the reader; from C it will appear to lie in the direction CS', displaced toward the right; and at D it will be displaced 'into the page', since the earth is here moving directly away from the reader. The star, in fact, will in one year appear to describe a small circle on the star sphere, a minute replica of the earth's orbit.

But consider (Fig. 12 (ii)), a second star lying, not at the pole of the orbit, but in its plane. From A it lies in the direction AS, and since the earth is moving towards S, it will also appear to lie in this direction, that is, it will not be displaced. At B, the earth is approaching us 'out of the page' and the star will appear to be similarly displaced towards S'. At C, the earth is moving directly away from the star, which will again be undisplaced. Finally, at D, the star will appear to be displaced 'into the page' in the direction DS'. It has thus in the course of one year moved back and forth over a straight line from a central undisplaced position (earth at A) to maximum displacement 'out of the page' (earth at B), back to the undisplaced position (earth at C), and on to maximum displacement 'into the page' (earth at D), ending up again with no displacement (earth at A, whence it started). A moment's thought will show: (i) the nearer a star is to the pole of the orbit, the more closely will its aberrational displacement approximate to a true circle, while the nearer it is to the plane of the orbit, the more elliptical will it become, until in the plane itself it will have been flattened completely to a straight line (Fig. 13); (ii) in every case the major axis of the ellipse must be parallel to the plane of the orbit.

Now if, as we are supposing, $ABCD$ is the earth's circumsolar orbit, the intersection of the plane of this orbit produced to meet the star sphere will be the ecliptic, for in looking at the sun (i.e. in looking at a point on the ecliptic) the terrestrial observer must necessarily employ a line of sight that lies in the plane of the earth's heliocentric orbit. Thus,

from A the sun's position on the star sphere will be beyond C,
from B the sun's position on the star sphere will be beyond D,
from C the sun's position on the star sphere will be beyond A,
from D the sun's position on the star sphere will be beyond B,

and one revolution of the earth about the sun will result in the sun making a complete apparent circuit of the star sphere.

Therefore Fig. 13 (i) should represent the annual aberrational displacement of a star at the pole of the ecliptic, (ii) of a star midway between the pole and the ecliptic, and (iii) of a star lying on the ecliptic.

Bradley, with the aid of telescopic equipment denied to Kepler,

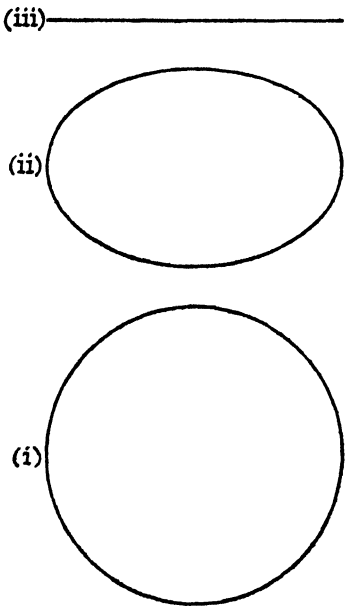


Figure 13.

showed that the stars do actually exhibit annual displacements of this character, and thus clinched for all time the heliocentric argument. The earth revolves about the sun.[1]

Effects of the earth's orbital motion

Here, then, is an entirely new (and, so far, much more promising) foundation for our speculations and deductions concerning the observational facts outlined in the last chapter. We do not have to go right back to the beginning, to the Ptolemaic hypothesis in its entirety, for the basis of our reconstruction, since we have already discovered that the earth possesses a motion of axial rotation and have reasoned how this rotation must affect the apparent motions of the heavenly bodies. Subtracting the effect of the earth's rotation from the observed solar motion, we were left with the sun moving

[1] Two further phenomena which are only explicable in terms of the heliocentric hypothesis—the parallactic motions of the stars, and certain periodic shifts in stellar spectra—will be described later. At this juncture, the aberration of starlight is sufficient to establish the earth's heliocentric motion.

steadily eastwards round the ecliptic, completing one circuit in a year. Now exactly this effect will be produced if the earth revolves about the sun in a period of one year. In Fig. 14 the inner circle represents the earth's orbit, the sun being centrally placed within it, and the outer circle the star sphere (or, more accurately, a section of it, which is the ecliptic) against which the sun is projected. When the earth is in position e_1 the sun appears to be at a point on the star sphere marked by the symbol s_1; when the earth has moved to e_2 the sun appears to be at s_2; when at e_3, at s_3. By the time the earth

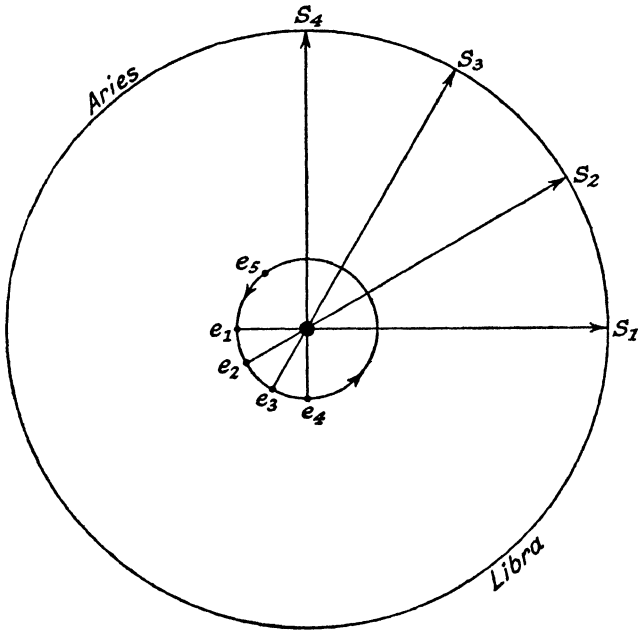


Figure 14.

has completed a quarter-revolution the sun will be at s_4, having also completed one-quarter of its entire yearly circuit of the ecliptic.

The heliocentric hypothesis and the four-minute discrepancy

One further point connected with the apparent motion of the sun must be cleared up. It will be remembered that every star culminates four minutes earlier each night. We may express this fact in a different way by saying that the sidereal day (the period required for one complete revolution of the star sphere) is slightly shorter than the solar day (the time required by the sun for one complete diurnal circuit of the heavens). This discrepancy is simply accounted for on the assumption that the earth is moving round the sun, with the result that the sun appears to be slowly advancing from west to east. In

fact this is just what we should expect to find were the earth truly revolving about the sun. Fig. 15 (which, to make the demonstration clearer, is not drawn to scale) represents the sun, the earth's orbit, and the earth in two positions in that orbit. These two positions are those that would be assumed at each end of a time interval of twenty-four sidereal hours—the period of one complete axial rotation. In the first position an observer at *a* sees the sun on his meridian at the same time that an observer at *b*, in the antipodes, sees a certain star on his. But when the earth has completed one rotation its orbital motion has changed its spatial position, and therefore the direction of the sun but not that of the incomparably more distant star. The result is that when the night-side observer, now at *b'*, sees his star

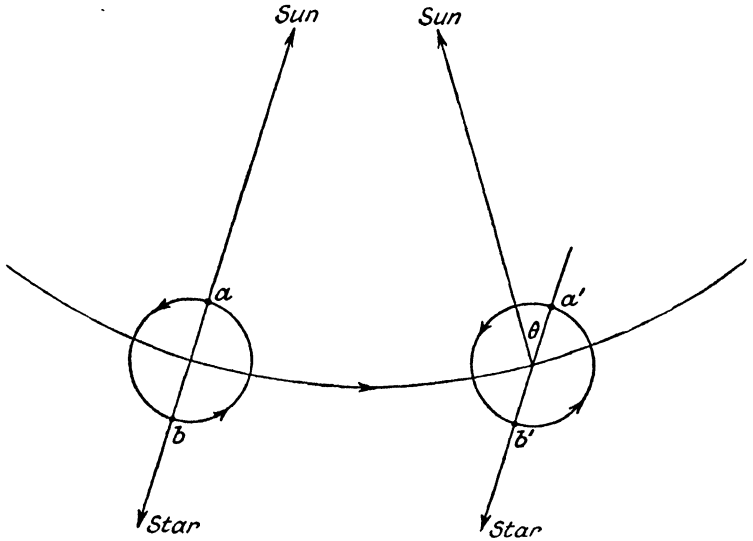


Figure 15. Explanation of difference between sidereal and solar days.

on the meridian (the figure shows the earth and the two observers at this moment) the day-side observer, now at *a'*, observes the sun still to the east of his meridian. The sun will not culminate until the earth has rotated a little further—through the angle θ, in fact. Hence the interval between successive culminations of the sun is a little longer than that between successive culminations of the star. Actually, the time required for the earth to rotate through the angle θ is about four minutes. Hence the solar day is some four minutes longer than the sidereal day, the discrepancy being caused by the earth's orbital motion.

The lunar motion

The moon appears to circle round the earth in a period of one month. Can it be that the earth is really revolving round the moon?

The answer must be negative, because if the moon is revolving steadily round the sun, and the earth round the moon, then the sun would appear to move across the sky in a series of loops, just as the epicyclic motion of the planets in Ptolemy's model gave them a looping motion. If, however, we suppose that the moon revolves about the earth, and the earth about the sun in a simple circular orbit, this difficulty vanishes. The moon, then, retains its Ptolemaic status; it is a satellite of the earth, and revolves about its primary in a period of one lunar month.

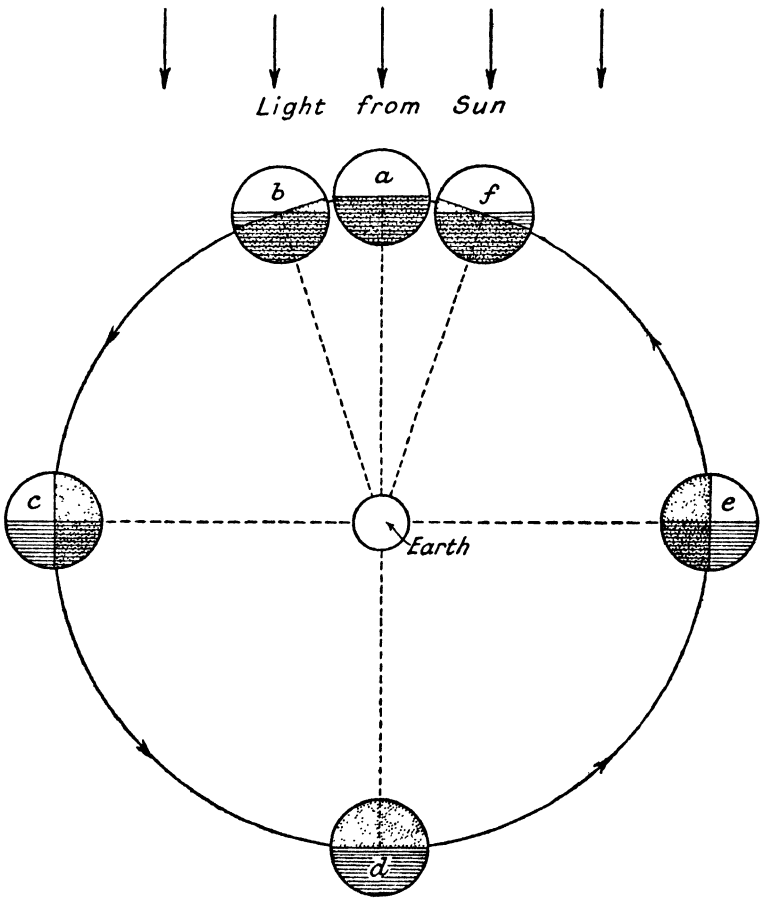


Figure 16. The earthward lunar hemisphere is stippled, the unilluminated hemisphere shaded.

Explanation of the lunar phases

In this way, too, the lunar phases are simply accounted for.[1] When the moon (Fig. 16) is in position *a* in its orbit, it is invisible; for not

[1] Though this in itself is no advance, since they were also accounted for in the Ptolemaic system.

only is it so near to the sun in the sky that it is swamped in the latter's glare, but its illuminated hemisphere is turned directly away from the earth. A few days later it has moved to *b* and is then far enough (angularly) from the sun for a narrow crescent of the illuminated hemisphere to be seen from the earth. At *c* it has moved round one-quarter of its orbit and, as can be seen from the diagram, is dichotomized as observed from the earth. A week later it has reached *d* when it will be full, its illuminated hemisphere being turned directly towards the earth: as seen from the moon, the earth and sun are both in the same direction. Three weeks after new it will be at *e* and will once more be dichotomized. But now the left-hand, or eastern, half of its disc is illuminated, whereas at the end of the first week it was the right-hand, or west, side. This, as will be remembered from our account of the monthly apparent behaviour of the moon, exactly fits the facts. Finally, towards the end of the fourth week, it is once more a narrow sickle; the east (left-hand) side being illuminated, whereas at *b* it was the west.

The heliocentric hypothesis and the outer planets

We saw in the last chapter that the planets Mars, Jupiter and Saturn make complete circuits of the star sphere. And since the periods in which they do so are all longer than that in which the sun appears to make one circuit (i.e. one year) it follows that it is possible for any of these planets to be in that part of the zodiac which is diametrically opposite to that containing the sun. That is, they may culminate at midnight. Thus it is possible for any of them to be in Aries, for instance, when the sun is in Libra (Fig. 14). But when the sun is in Libra the earth is in position e_8 approximately. It is obvious that if one of these planets is to be between the earth and that part of the star sphere opposed to the sun, then it cannot be between the earth and the sun. The orbits of Mars, Jupiter and Saturn, in fact, must lie *outside* the orbit of the earth. And since their apparent velocities across the star sphere, and hence their periods of revolution, can be used to gauge their relative distances, we may conclude that the relative positions of their orbits to that of the earth are as shown in Fig. 17.

Regarding the planets Mars, Jupiter and Saturn, therefore, our hypothesis requires that—

i. These planets (and the earth) all revolve about the sun.

ii. They revolve about the sun in orbits situated outside that of the earth.

iii. It takes longer for them to traverse these longer orbits than it does for the earth to traverse its smaller one.

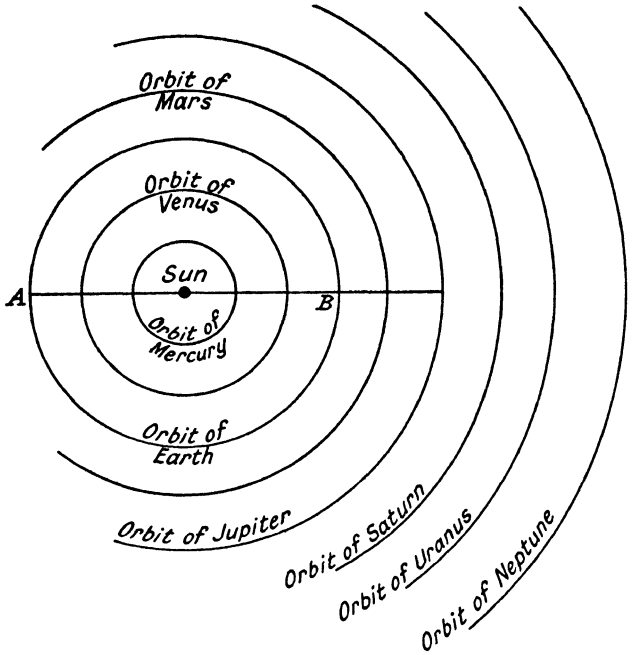


Figure 17. The Copernican plan of the Solar System (not to scale).

Assuming that this is so, can we account for their apparent looping motions as projected against the star sphere? The answer is that we can, quite simply and inevitably, and without having recourse to any complicated system of epicycles.

Fig. 18 shows the orbits of the earth and of some outer planet (Mars, for example), with the two bodies in corresponding positions in their respective orbits. When the earth is at e_1, Mars is at m_1; when at e_2, Mars is at m_2, and so on. These positions are marked on the diagram according to our maxim that the earth revolves about its orbit in a shorter time than any of the outer planets. The apparent motion of Mars upon the star sphere is clearly the resultant of its own orbital motion and of the observer's motion round the terrestrial orbit. This apparent motion is shown on the outermost circle, which represents the distant starry background against which Mars is, to a terrestrial observer, projected. It will be seen that at first Mars appears to be moving in a direct west-east direction, but with decreasing velocity: the interval M_1—M_2 is longer than M_2—M_3 which is in turn longer than M_3—M_4, and so on. At position 5 it appears to pause, and then doubles back in its tracks to position 6; this retrograde (east-west) motion continues as far as position 7 when the planet again pauses, being for a short time stationary on the star sphere. Then it resumes its direct motion, passing eastwards through

8, 9, 10 and 11 with increasing velocity. Thus the observed effect of the earth's circumsolar motion upon that of an outer planet is to make it move eastwards across the star sphere in a series of loops, the loop occurring once a year when the planet is near opposition. Here is a more satisfactory explanation than that of Ptolemy. It is

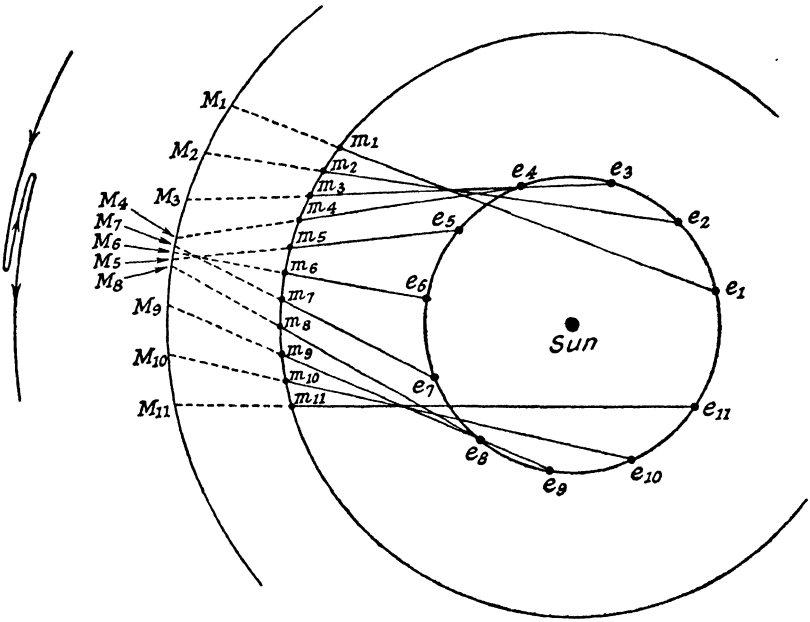


Figure 18. The Copernican explanation of the 'looping' of the outer planets.

more satisfactory because it is simpler, eliminating the necessity for complex and quite arbitrary epicyclic gearing.

The heliocentric hypothesis and the inner planets

Early in our observations we found that the planets had to be divided into two groups, the dissimilarity between the members of which being indicated by their completely dissimilar apparent motions. Since one group consists of those planets (Mars, Jupiter and Saturn) which we now know to lie further from the sun than the earth, we might make a guess at the nature of the other group and say that Venus and Mercury are nearer to the sun than the earth. But we can establish this result by better methods than hit-or-miss guessing. Three facts of observation point conclusively to its truth.

i. We have seen that if a planet revolves about the sun in an orbit more distant from the sun than that of the earth, then it must at times culminate at midnight. That is, it will be able to occupy a

position on the zodiac diametrically opposed to that of the sun. In this position it will necessarily be at an angular distance of 180° from the sun. But Venus and Mercury can never occupy this position on the meridian at midnight since the angular distances to which they can recede from the sun are strictly limited to about 48° in the case of Venus and 28° in the case of Mercury. Hence they cannot be further from the sun than the earth.

ii. From time to time both Venus and Mercury transit the sun; that is, they are observed to cross the sun's disc. When in transit, Venus is visible to the naked eye (if the eye is shielded with a dark glass) as a small black spot moving slowly across the face of the sun. Now a glance at Fig. 17 will show that under no circumstances could an outer planet ever be in a position between the earth and the sun. Hence Venus and Mercury must move in orbits situated between that of the earth and the sun itself.

iii. Fig. 19 shows that an inner planet, as seen from the earth, will swing from side to side of the sun in just the same way that an observer on the sun would see the moon do in respect of the earth. Hence, if Mercury and Venus really are moving in orbits between the earth and the sun, they must show complete cycles of phases—and not otherwise. At *a* and *b*, when the planet is said to be at greatest elongation, it will be dichotomized. At *d* it will usually be invisible, since its illuminated hemisphere will be turned away from the earth; in this position, however, it may pass directly between the earth and the sun, in which case it will be visible in transit. At *c* it will be fully illuminated, though here it will again be invisible since it occupies the same part of the star sphere as the sun and will therefore rise and set with it, never being visible in the night sky. To the naked eye neither Venus nor Mercury exhibits the phase phenomena, although their brightness depends upon what region of their orbits, relative to the earth, they are occupying. This fact of the invisibility of the phases without telescopic aid prevented Ptolemy from detecting an obvious error of his system, in which an inner planet could at no time exhibit the full phase. It was not, indeed, until nine years after the death of Tycho Brahe that Galileo observed Venus telescopically, and the phases were seen for the first time.

Regarding the relative positions of Mercury and Venus, we may confidently assert that Mercury is the nearer to the sun since its greatest elongation is only 28°, whereas Venus may recede to 48° from the sun, or nearly twice as far.

Why is the component of the outer planets' apparent motion due to the earth's real orbital motion so large? Clearly because, relative to these bodies, the earth changes its position, and therefore the

outlook of the observer, very considerably in a given space of time. But this is not so in the case of the inner planets. We have learnt from our observations of Mars, Jupiter and Saturn that the further a planet is from the sun, the more slowly it moves—or, more accurately, the more slowly it changes its position relative to the stellar background. Hence, relative to Venus and Mercury, which are nearer the sun than the earth, the earth moves through a small fraction of its orbit in a given interval—the interval required by Mercury to make one complete revolution of the sun, for instance. For this reason the effect of

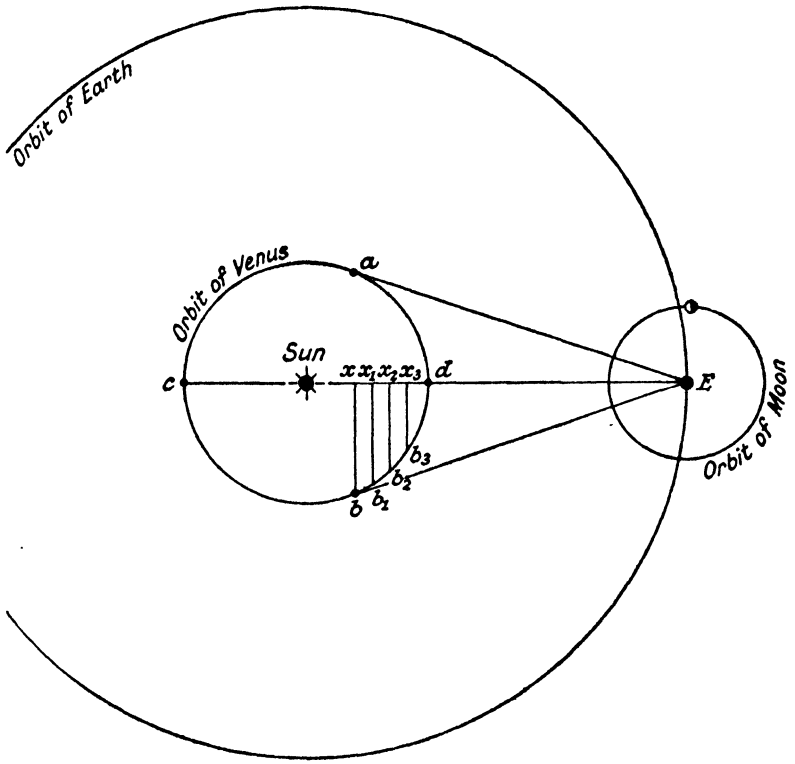


Figure 19.

the earth's motion on the observed motions of the inner planets is comparatively small: they move in nearly the same apparent paths and with nearly the same apparent velocities as they would were the earth stationary in its orbit. For the sole purpose of seeing how the heliocentric account of the relative positions and motions of the earth, Venus and Mercury, tally with the observed facts, therefore, we may neglect the earth's orbital motion.

Referring to Fig. 19, we may see how the revolution of an inner planet in its circumsolar orbit will cause it to assume the observed

successive positions of Venus and Mercury. When the planet is at *b* it is at greatest elongation east. Thus, if the line *Ec* represents the observer's meridian so that the sun is due south (midday), the planet will be as far east of the sun as it ever can be at that time of day. It will therefore be visible as an evening star above the western horizon after sunset. As it moves round its orbit and approaches *d*, it will appear to approach nearer and nearer the sun, as is shown by the fact that the distances xb, x_1b_1, x_2b_2 . . . , which measure the angular separation of the sun from the planet in the successively assumed positions b, b_1, b_2 . . . , become progressively shorter. The planet will consequently set earlier, and be visible for a shorter period, each evening. When it is at *d* the line x_nb_n will have no length at all, the angular distance of the planet from the sun being zero. It will at this time be invisible, as a bright point in the sky, but may transit the sun. (The mechanism of transits and of eclipses, which are allied phenomena, will be discussed later.)[1] Proceeding from *d* towards *a*, the planet will recede steadily further from the west side of the sun, eventually becoming visible as a 'morning star' above the eastern horizon at dawn. When at *a* itself, it is at its greatest elongation west and will then begin to swing back towards the sun until, at *c*, both it and the sun are in the same straight line from the earth; it is again invisible, rising and setting with the sun. Having passed *c*, it will once more become visible, for an increasingly long period each day, as the 'evening star.' This continues until it reaches greatest elongation east (position *b*).

Diurnal motions of the planets explained

So far, we have only considered the annual motions of the planets, but there is the discrepancy between the diurnal period of each of the planets and that of the stars, which we discovered when making the observations recorded in the last chapter. It will be clear by now that this discrepancy between the time interval separating successive culminations of a star and of any of the planets is due to two factors—the earth's orbital motion and that of the planet. When its orbital motion is carrying an outer planet eastward among the stars, with an apparent velocity that is partly determined by the earth's motion, it will require *longer* than the stars to complete one diurnal revolution. But, when retrograding (and the apparent retrogressions of the outer planets are due entirely to the earth's orbital motion) the outer planet will accomplish two successive culminations in a period slightly *shorter* than the 23h. 56m. required by the star sphere. In the same way the apparent motions of the inner planets among the stars

[1] See Chapter VII.

cause a small difference between their diurnal periods and that of the stars, and these apparent motions are compounded of the real orbital motion of the earth as well as of the real orbital motions of the planets.

The Copernican cosmology

Thus we see that the heliocentric model of the solar system gives *in general* a more satisfactory account of the observed motions of the sun, moon and planets than the geocentric. In addition, it offers an explanation of the parallactic shifts of the stars, which the Ptolemaic theory cannot (it being a Ptolemaic axiom that such shifts do not even exist), and is based upon two facts that automatically render the geocentric hypothesis nonsensical; namely, that the earth rotates on its axis and, more important, that it revolves about the sun in a period of one year. It will be essential, nevertheless, to see if the heliocentric hypothesis—once it is elaborated and refined, as it must be, to cover the *specific* motions of the planets—can avoid the hideous complexities and absence of all order or universality which characterized the Ptolemaic system.

If Copernicus can be hailed as the inspired prophet and visionary of the new heliocentric cosmology, Kepler might be regarded as the foot-slogging worker who followed him and who, with indefatigable industry though with less genius, co-ordinated and revised his inspirations and presented them in a form acceptable to the general scientific community. This is a plausible picture and one which has counterparts in most spheres of intellectual endeavour. But unfortunately a study of the historical facts impresses one with its inaccuracy. Copernicus only partially effected the break with the ecclesiastically favoured geocentric tradition, and retained one of its major prejudices. Kepler broke as much new ground as Copernicus and, in addition, changed the status of the heliocentric hypothesis from that of a plausible alternative to the Ptolemaic (if ecclesiastical authority would allow it to be judged on its merits) to that of the one and only possible explanation of the observed facts.

Copernicus, as the result of his study of observations more accurate than any that had been made hitherto, convinced himself that the sun and not the earth was the body about which the rest of the solar system revolved.[1] He also pointed out that the apparent diurnal

[1] How hazardous was the progress of science, before the advent of the printing press permitted the wide diffusion of new knowledge and speculation, is strikingly illustrated by the fact that Pythagoras had, 2,000 years earlier, suggested that the earth might revolve about the sun, and not vice versa. A disciple of his, moreover, even suggested the rotation of the earth as a possible alternative to the diurnal otation of the star sphere.

revolution of the stars could as easily be explained by assuming a real axial rotation of the earth as a real rotation of the star sphere. But he could not completely break with tradition and never succeeded in freeing himself from the grip of the dogma, handed down from the Greeks and sanctified by the Church, that the universe is 'perfect' and that since the circle is the 'perfect' figure, the planetary orbits must be circular. Yet, try as he would, he could not reconcile theory with observed facts, while working on this assumption of circular orbits, without having recourse to epicycles. It is true that he was able to reduce their number very considerably, but his submission to their inevitable introduction into his system prevented him from discovering the key secret of planetary motion. He was further forced to suppose that the circular planetary orbits were slightly eccentric—that is, not exactly centred upon the sun.

Kepler's search

About thirty years after the death of Copernicus, Kepler was born. He was obsessed with the idea that the sun, moon and planets formed a coherent system, and that their motions, orbits and distances were simply related by some general law or laws. In the meantime the Danish astronomer, Tycho Brahe, had been amassing, over a period of years, a volume of planetary observations of a scope and accuracy far surpassing those of Copernicus or of any contemporary. At the age of thirty Kepler came into possession of these records, and a close study of them convinced him of the truth of two facts: (i) Copernicus was right in placing the sun at the centre of the solar system, but (ii) his epicyclic orbits were incapable of yielding motions which agreed with the observed motions of the planets.

For years Kepler laboured at these tables. His problem was to discover a law of planetary motion describing both the shape of a planetary orbit, and also the motion of the planet within it, such that the theoretical results of its application should exactly match the observed behaviour of the planets as recorded in Tycho Brahe's tables. Once found, such a law could be checked and verified by basing predictions upon it and observing whether the planets did in fact fulfil these predictions; if they did, he might be sure that he had put his finger upon the law regulating their motions. With a perseverance perhaps unrivalled in the history of science he tried one device after another, and again and again they had to be discarded as being incapable of yielding the configurations recorded by Tycho Brahe's observations. Eventually the search led to success, and Kepler was able to formulate the three laws of planetary motion.

Kepler's laws of planetary motion

i. The orbit of each planet is an ellipse, having the sun at one of its foci. The non-mathematical reader may not have a very clear idea of what an ellipse is (many people suppose that an egg is elliptical), and the simplest way of elucidating the matter is to draw one. Rule a line across a sheet of paper and fix the sheet to a drawing board or table-top with drawing pins. Then stick two ordinary pins firmly into the sheet, piercing the line and lying about 3 inches apart. Make a loop of thick cotton or fine string and place it over the pins. Pulling the loop tight, and keeping it so with the point of a pencil, draw a free curve from that part of the transverse line on the left of the left pin to that on the right of the right pin. Then place the loop on the opposite side of the line and draw the other half of the curve. It will then be seen that an ellipse is nothing more than a foreshortened circle; it is a circle looked at, not from directly above, but aslant. Alternatively it may be regarded as a circle with two centres—the points marked by the pins—instead of one; these points are known as the foci, and if the ellipse we have drawn represents the orbit of a planet the sun will be situated at one of them. It should be noted, by experimenting with different positions of the pins but using the same sized loop, that the eccentricity of an ellipse depends upon their distance apart: the greater this distance the flatter the ellipse, the nearer together the closer the approximation to a circle. When they are as close together as they can be—that is, in the same position—the figure described is a circle. This is what is meant when it is said that the ellipse may be regarded as a circle with two centres.

ii. Kepler discovered that even if the planets move in elliptical orbits, the sun occupying one of the foci, they will not behave in the observed fashion so long as their motion is uniform; i.e. so long as they travel round their orbits with unvarying velocities. The second part of his problem was to formulate the law which states the orbital velocity of a planet at any point in its orbit. After following up a number of false trails, he eventually arrived at the solution: the motion of each planet is such that the radius vector describes equal areas in equal times. (The radius vector is the line joining the planet and the sun.) Fig. 20 represents a planetary orbit with the sun at the focus S. Within this orbit are described four triangles, all of whose areas are equal. Now, according to Kepler's second law, the radius vector will sweep out these triangles in equal times. Hence the planet will require the same period to travel from A to B as from C to D, E to F, and G to H. A general and less precise way of expressing

this law is obviously to say that the orbital velocity of a planet depends upon its distance from the sun: the nearer to the sun it is, the more rapidly it moves.

iii. But Kepler still had one further point to clear up. As we have seen, the more distant planets require a longer period in which to complete one circuit of the star sphere than those nearer the sun.

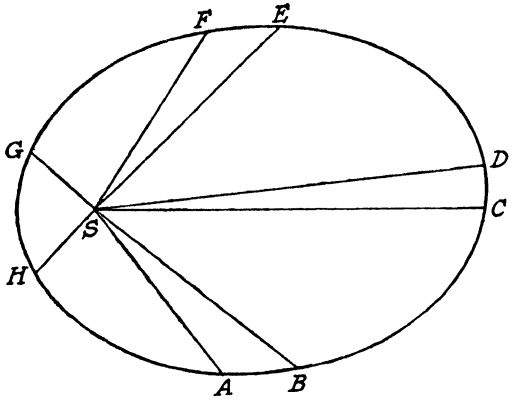


Figure 20. *S* is one focus of the ellipse. The triangles *SAB*, *SCD*, *SEF* and *SGH* are of equal area.

Kepler was convinced that this fact could be expressed in an exact form. As before, the solution persistently evaded him, but finally he discovered that there was, as he had suspected, a strict relation holding between the distance of a planet from the sun and the period in which it revolves about it. This relation may be expressed in several forms, of which this is one: if the squares of the periods in which the planets describe their respective orbits are divided by the cubes of their mean solar distances, the quotient will be the same for every planet; in other words, the squares of their periods are proportional to the cubes of their mean solar distances. Thus, if P and p are the periods of two planets, D and d their mean distances from the sun, then

$$\frac{P^2}{D^3}=\frac{p^2}{d^3}$$

and this is true for any two planets that happen to be chosen. Hence, if the periods of two planets are known, their mean solar distances in terms of each other may be calculated. To take an example, suppose that the periods of the planets are respectively 2 years and 15 years. Then

$$\frac{4}{D^3}=\frac{225}{d^3}$$

Therefore $4d^3 = 225\,D^3$

Therefore $d^3 = \frac{225}{4}\,D^3$

Therefore $d = \sqrt[3]{\frac{225}{4}}\,D$

$= 3{\cdot}8\,D$ (approximately).

That is, the mean distance from the sun of the planet whose period is fifteen years is rather less than four times that of the planet whose period is two years. If we could determine the linear solar distance of any one planet, we could, by applying the harmonic law, calculate the linear solar distances of all the other planets. For instance, if we knew that the planet whose period is two years, were 90 million miles from the sun, and the relation between its distance and that of another planet were 1 : 3·8 as already calculated, then the mean solar distance of this other planet must be 340 million miles. In the same way the linear solar distances of all the other planets could be calculated from their observed periods.

Kepler's achievement analysed

It is worth while summarizing the methods by which Kepler reached these results, for otherwise his immense achievement must appear as incomprehensible as that of a conjuror who produces rabbits from an empty hat. His first task was to determine the true motions and orbits of the planets from their apparent motions, i.e. from the observed phenomena described in the last chapter as they were recorded in Tycho's records. This in itself could have been attempted by no predecessor of Copernicus for it was only following the conception of a moving earth that the distinction between true and apparent motions could be entertained. To solve the problem Kepler had to find a way of allowing for the effects of the earth's motion and subtracting this effect from the observed movements of the planets. Before his solution of the problem can be understood, the reader must know something of the conception of sidereal and synodic periods.

Sidereal and synodic periods

Fig. 21 shows the circumsolar orbits of the earth and an outer planet. Let us suppose that at P_1 the planet is in conjunction with a certain star, Z, as seen from the sun. The planet's sidereal period is then defined as the period it would require in which to make one complete circuit of its orbit back to P_1 again.

The synodic period is defined as the interval between successive

conjunctions of the earth and the planet, also as observed from the sun. Let us see what this means. We will suppose that the earth and the planet are in conjunction; in other words, they lie on the same straight line from the sun, SE_1P_1. As the planet is moving forward

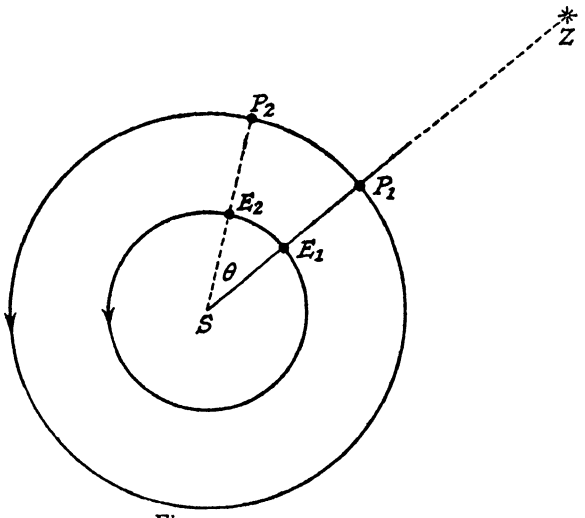


Figure 21.

like the earth, though more slowly, the latter will have to cover considerably more than one revolution before it, the planet and the sun are again in a straight line; thus, during the time required for the earth to travel completely round its orbit from E_1 to E_1 and the additional distance E_1—E_2, the planet has moved from P_1 to P_2. The time required for the planet to move this distance is its synodic period, since it is the interval between successive conjunctions of the earth and itself as seen from the sun. Furthermore, since the planet must be in opposition as seen from the earth when it and the earth are in conjunction as seen from the sun, we may define the synodic period of a planet as the interval between successive oppositions. It will be noted that the more distant, and therefore slow moving, a planet is, the smaller will the angle θ become, and the more closely will the synodic period approximate to the earth's sidereal period of 365 days. This fact is brought out in the following table:

Planet	*Synodic Period*
Mars	.. 780 days
Jupiter	.. 399
Saturn	.. 378
Uranus ..	.. 370
Neptune ..	.. 367
Pluto.... ..	.. 366·7

In the case of Venus and Mercury, it is the inner planet which does the 'catching up', and we may complete the table by writing:

Venus	584
Mercury	116

The simplest method of determining the synodic period of an inner planet is to observe two successive transits (when it passes across the sun's disc, and the sun, the earth and the planet lie in the same line) and to divide the interval between them by the number of revolutions the planet has made in the interval.[1]

Summarizing this, we may say:

Sidereal Period: time required by the planet for one complete circuit of the star sphere back to the starting point, as seen from the sun.

Synodic Period: interval between successive conjunctions with the earth, as seen from the sun.

Or, one complete revolution with reference to the line joining sun and earth.

Or, interval between successive oppositions: outer planet $\frac{\text{interval between successive transits}}{\text{number of revolutions in the interval}}$: inner planet.

It now remains to be seen how the sidereal and synodic periods of a planet are related, and how the former may be derived from the latter. Let us put

S for the planet's synodic period,
P for its sidereal period,
Y for the earth's sidereal period (one year).

Then $\frac{360}{P}$ = the angle through which the planet moves round the sun in one day.

$\frac{360}{Y}$ = the angle through which the earth moves round the sun in one day.

$\frac{360}{S}$ = the angle through which their directions as seen from the sun separate in one day.

Considering, first, an inner planet, which moves more rapidly than

[1] Although transits are similar conjunctions, they are not successive, since the inner planets do not transit the sun in every revolution, their orbits being inclined to that of the earth. To derive the period between *successive* similar conjunctions from transit observations, therefore, it is necessary to divide the interval between successive transits by the number of revolutions performed in that interval.

the earth, $\frac{360}{S}$ represents its daily gain on the earth. This gain is also given by $\frac{360}{P}-\frac{360}{Y}$, and we may therefore write

$$\frac{360}{S}=\frac{360}{P}-\frac{360}{Y}$$

or

$$\frac{1}{S}=\frac{1}{P}-\frac{1}{Y} \quad \ldots\ldots\ldots\ldots\ldots \text{(i)}$$

In the case of an outer planet, which moves more slowly than the earth, $\frac{360}{S}$ represents the earth's daily gain on the planet. $\frac{360}{Y}-\frac{360}{P}$ also gives this daily gain, and we can once more equate the two expressions:

$$\frac{360}{S}=\frac{360}{Y}-\frac{360}{P}$$

or

$$\frac{1}{S}=\frac{1}{Y}-\frac{1}{P} \quad \ldots\ldots\ldots\ldots\ldots \text{(ii)}$$

The great value of these two equations lies in the fact that by their means we can calculate the sidereal period of a planet merely from observations of its position on the star sphere. This Kepler did, using Tycho's numerous observations. He concentrated his attention on the planet Mars, and since the relative positions of sun and planets are much used in astrology, which Tycho practised in addition to his astronomical work, Kepler was provided with very full data of past oppositions of Mars. From these he was able to deduce its synodic period with reasonable accuracy, and thence, by means of equation (ii), its sidereal period.

Determination of the Martian orbit and discovery of the first two laws

Before he could use the planet's sidereal period as a stepping stone in his main task of determining the form and size of the Martian orbit, he still required to know something of the terrestrial motion, so that he could fix the true daily position of the earth. For this purpose he accepted Tycho's hypothesis, and this, though inaccurate, was a near enough first approximation to give him significant results. Thenceforward the problem was one of simple geometry.

Fig. 22 shows the orbits of the earth and of Mars. We will suppose that Mars is first observed from position E_1; after one sidereal period Mars will again be back at M, but the earth, with its more rapid motion, will have carried out more than one revolution and have

arrived at some point E_2. Mars is now reobserved from E_2. Knowledge of the earth's motion gives the values of SE_1, SE_2 and the angle E_1SE_2, the latter depending upon the interval between the two observations. The observations themselves (E_1M and E_2M) give the values of the angles SE_1M and SE_2M. These five quantities are sufficient for the solution of the quadrilateral SE_1ME_2, whence SM and the angle E_2SM can be calculated: SM is the distance of

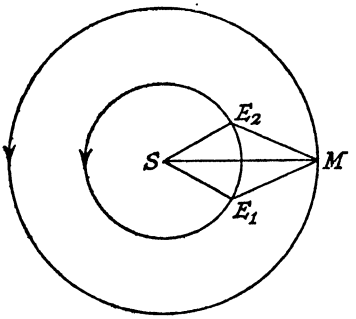


Figure 22.

Mars from the sun when at point M in its orbit, and E_2SM its direction as seen from the sun.

Thus by means of a series of paired observations, the members of each pair being separated by 687 days (the sidereal period of Mars), spread over two sidereal periods, a large number of points on the Martian orbit may be plotted and the complete orbit reconstructed, showing both its size and its form.

Once the orbit had been determined in this way, simple inspection led to the discovery of the first law; and since the times of the various observations were known, together with the solar distance of the planet at each, trial and error would eventually lead to the formulation of the second law.

Discovery of the third law

Having reached this stage, it only remained to discover the nature of the relationship which Kepler was convinced held between the planets' periods and their distances from the sun. The sidereal periods of the planets were easily deduced from Tycho's observations by means of the equations (i) and (ii), given on p. 58. Kepler still had to discover their relative distances. We have seen how easily these may be deduced from their periods by means of the harmonic law, but Kepler obviously could not avail himself of this device as he was still in the process of discovering it. An independent geometrical method is illustrated in Fig. 23, which shows the orbits of the earth

and of an inner planet. As the figure stands, the planet is at greatest elongation: the line of sight from the earth to the planet, EV, is therefore a tangent to the orbit, and the angle EVS a right angle. VES, the angular separation of the planet and the sun, is measured, whence $VSE = 180° - (90° + VES)$. These data permit the calculation of the relative lengths of SV, the solar distance of the planet, and SE, the

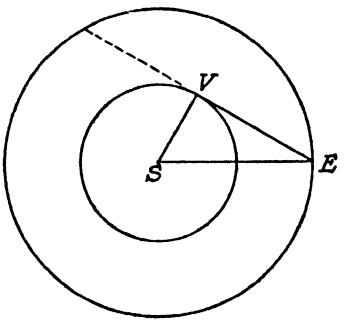


Figure 23.

solar distance of the earth. The method is essentially similar, though rather more involved, in the case of an outer planet.

Kepler was now in a position to tabulate his information as follows:

PLANET	SIDEREAL PERIOD (P)	MEAN SOLAR DISTANCE (D)
Mercury ..	0·241 years	.. 0·3871
Venus ..	0·615	.. 0·7233
Earth ..	1·000	.. 1·0000
Mars ..	1·881	.. 1·5237
Jupiter ..	11·862	.. 5·2028
Saturn ..	29·458	.. 9·5388
Uranus ..	84·015	.. 19·1910
Neptune ..	164·788	.. 30·0707
Pluto ..	247·7	.. 39·5

The final stage in the solution of the problem of planetary motion consisted of the identification of a single relationship which held between each of the first six pairs of figures in the table. (Uranus, Neptune and Pluto were not known in Kepler's day.) A laborious process of trial and error eventually revealed the relationship expressed in the third law: the square of the ratio of a planet's period to that of the earth equals the cube of the ratio of the planet's mean solar distance to that of the earth. This, then, was the climax of Kepler's life work, and of it he wrote exultantly, 'The die is cast, the

book is written, to be read either now or by posterity, I care not which; it can await its reader; has not God waited six thousand years for an observer?'

Newton's law of universal gravitation

Some fifty years after Kepler had enunciated the three empirical laws of planetary motion, Newton, who had been searching for a single general law which would cover all these, formulated his law of universal gravitation. This law explained why the planets behave in the particular manner described by Kepler, and it is consequently both more fundamental and more general than Kepler's laws. In other words, bodies revolving about the sun move in the manner described (but not explained) by Kepler, for the reason that otherwise they would be transgressing the law of gravitation. Newton's law, fundamental to the whole realm of astronomy and physical science, has been tested observationally times without number, and has never failed to square with the facts.[1] It states that every body in the universe attracts every other body with a force that is proportional to the product of their respective masses and inversely proportional to the square of the distance between them.

Those readers who do not shy away from a little simple mathematics may be interested to see how the inverse square law may be derived from Kepler's third law; this, as we have just seen, states that the squares of planetary periods are proportional to the cubes of their mean solar distances, or

$$\frac{P^2}{p^2}=\frac{D^3}{d^3}$$

Let us consider the case of a planet travelling round the sun at a velocity V in a circular orbit at a distance from the sun equal to D. (The case of an elliptical orbit, though similarly deducible, involves more complicated working.) Since the planet is moving under the influence of an inward-directed force centred in the sun, and yet does not fall into the sun, it must also be acted upon by an equal force operating in the opposite direction: this force is familiar to anyone who has whirled a conker or similar weight at the end of a piece of string.

The value of the force is given by

$$F=\frac{V^2}{D} \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(i)}$$

[1] Certain minute errors, corrected in Einstein's revised equations, excepted.

but since V, the velocity of the planet, cannot be deduced direct from observation it is desirable if possible to substitute for it a quantity, to which it stands in a known relation, which can be so derived. Such a quantity is the planet's period of revolution (P); for since velocity equals distance travelled divided by time taken, it follows that

$$V = \frac{2\pi D}{P}$$

If now we substitute this value of V in (i), we have

$$F = \frac{\left(\frac{2\pi D}{P}\right)^2}{D}$$

$$\text{or } F = 4\pi^2\left(\frac{D}{P^2}\right) \quad \ldots\ldots\ldots\ldots \text{(ii)}$$

Similarly in the case of a second planet

$$f = 4\pi^2\left(\frac{d}{p^2}\right) \quad \ldots\ldots\ldots\ldots \text{(iii)}$$

Combining (ii) and (iii),

$$\frac{F}{f} = \frac{D}{P^2}\left(\frac{p^2}{d}\right) \quad \ldots\ldots\ldots\ldots \text{(iv)}$$

Now the harmonic law (Kepler III) states that

$$\frac{P^2}{p^2} = \frac{D^3}{d^3}$$

$$\text{or } p^2 = \frac{P^2\, d^3}{D^3}$$

If now we substitute this value of p^2 in (iv) we get

$$\frac{F}{f} = \frac{D}{P^2}\left(\frac{P^2\, d^3}{D^3\, d}\right)$$

$$\text{or } \frac{F}{f} = \frac{d^2}{D^2}$$

That is, the force exerted upon the two planets is inversely proportional to the square of their distances: Newton's law of inverse squares.

Newton's modification of Kepler III

Mention must be made here of a slight modification which Newton made in the harmonic law as enunciated by Kepler. The planets move in the manner described by Kepler since they are acting under the central gravitational influence of the sun, which attracts each planet with a force inversely dependent upon the square of the

distance separating the two bodies. But another factor which enters into this attraction is the respective masses of the sun and the planet concerned: the attraction is in fact reciprocal, the planet tending to make the sun revolve round it at the same time that the sun makes it revolve about it. That the law as stated by Kepler appears to give a satisfactory account of appearances means that the mass of any planet is so small compared with that of the sun as to be, for practical purposes, negligible: this conclusion will be shown accurate when we come to determine the masses of the sun and planets. But when the masses of the two bodies are less disparate than those of the sun and a planet—for example, in the case of a double star—the mass of each body has to be taken into account, and this is what Newton's modification of Kepler III does. Where Kepler's third law states that

$$\frac{P^2}{D^3}=\frac{p^2}{d^3}$$

Newton's modified form of the equation states that

$$\frac{P^2\,(M+m_1)}{D^3}=\frac{p^2\,(M+m_2)}{d^3}$$

where M=mass of the sun, and m_1, m_2 the masses of the two planets. It can be seen that where m_1 and m_2 are negligible, as in the case of the planets, we are simply multiplying both sides of the equation by M, thus not materially affecting Kepler's equation.

The mass of the earth

While describing the physical properties of the earth in the last chapter, one in particular was not mentioned: its mass. Now that something has been learnt of the law of universal gravitation it will be well to return to this point, for not only does it illustrate very clearly the power and application of Newton's law, but also the value for the earth's mass so derived is a stepping stone to further astronomical quantities, as described in Chapter VII. First, a possible source of confusion must be cleared up. The distinction between mass and weight is usually a little difficult to grasp when encountered for the first time, but to put the matter simply we may say that the mass of a body is the quantity of matter in it, and that the spring balance measures weight while the scales measure mass. If a body just turns the scales at one pound it will always do so, no matter how strong or weak the gravitational field in which the scales are set up: the mass of the body does not vary with the forces to which it may be subjected. On the other hand, the stretch of a spring balance depends both upon the mass of body being weighed and on the gravitational

field attracting it: the weight of a body, unlike its mass, does vary

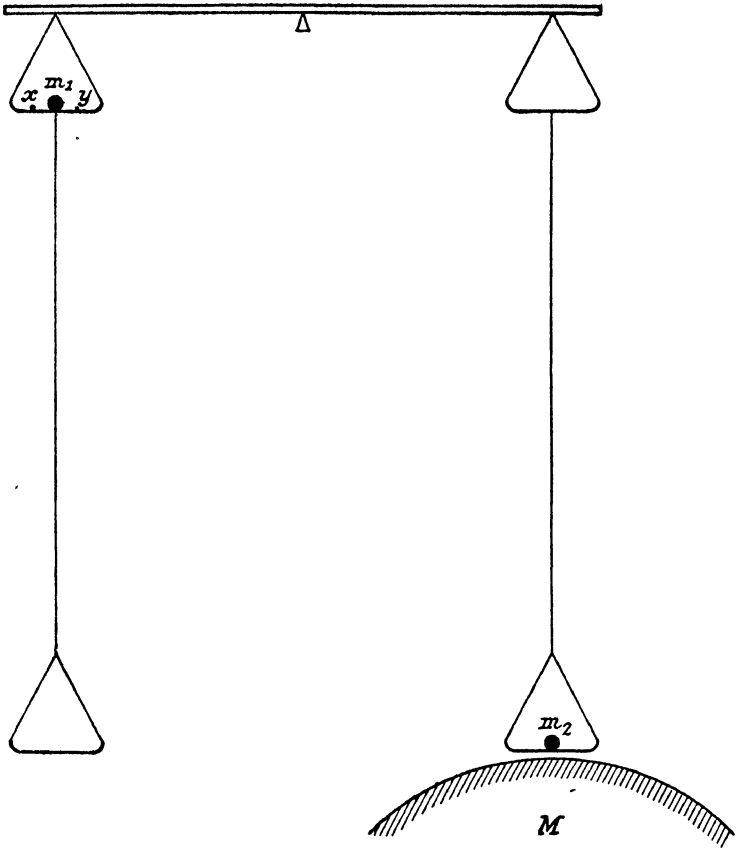


Figure 24. Joly's Balance.

with the strength of the gravitational field within which it is situated.

Expressed in mathematical terms, Newton's law states that

$$F = G.\frac{m_1.m_2}{d^2}$$

where F is the attractive force between the two bodies; m_1 and m_2 their masses; d their distance apart; and G the so-called gravitational constant (the attractive force between unit masses at unit distance apart).

Fig. 24 shows schematically the apparatus known as Joly's Balance. To the undersides of the two pans of a strong balance is suspended a second pair of pans by long wires. m_1 and m_2 are two spherical bodies of equal mass. When they occupy the same pair of pans, whether the upper or the lower, the beam will be horizontal, since they are equidistant from the earth's centre and therefore equally attracted by the earth's gravitation. But suppose that m_2 is placed in one of the

lower pans and m_1 in the opposite upper pan, as in Fig. 24. Then m_2 being nearer the earth, will be more strongly attracted by it than m_1. In other words it will weigh more, and the right hand side of the beam will be slightly depressed. This can be rectified by placing a small additional mass in the pan containing m_1; let its mass be x. When this second equilibrium has been attained a lead sphere of great mass, M, is placed close below the pan containing the mass m_2. Again, the beam of the balance will tilt down on the right, for although the lead mass is attracting m_2 the distance between the two pairs of scales is too great for it to attract m_1 appreciably. To bring the scales back into proper balance a second and even smaller mass, y, must be added to the pan already containing m_1 and x.

It is clear, now, that the attraction between M and m_2 is balancing, and therefore equal to, that between the earth and y. Applying the law of gravitation already stated, we may express this in the form:

$$\frac{G \cdot M \cdot m_2}{d^2} = \frac{G \cdot E \cdot y}{r^2}$$

where d=the distance between the centres of M and m_2,
r=the radius of the earth,
and E=the mass of the earth.

That is: $E=\frac{M.m_2.r^2}{y.d^2}$, all of which quantities are known or can be deduced. The result is approximately 5,800,000,000,000,000,000,000 or nearly six thousand million million million tons.

The imaginary case of a cloud-girt earth

At this point it may prove enlightening briefly to reconsider our knowledge of the earth's motions and physical properties from an entirely hypothetical point of view. Suppose that the earth's atmosphere, instead of being reasonably clear and transparent, were composed of dense cloud-banks similar, let us say, to those which blanket the planet Venus. Under these circumstances—with all the heavenly bodies permanently invisible, and the science of astronomy unknown—how much information would we then be able to deduce concerning the earth we inhabit? A glance through the last two chapters will show that a surprising amount of knowledge could be arrived at even under these unfavourable conditions; indeed, the earth's circumsolar motion would alone remain undetected, since relying on observations outside the earth itself—stellar observations connected with aberration or with parallactic or spectral shifts. Of the other terrestrial characteristics which we have considered, the

axial rotation could be proved by Foucault's pendulum, the gyroscope, or the eastward deviation of falling bodies from the vertical;

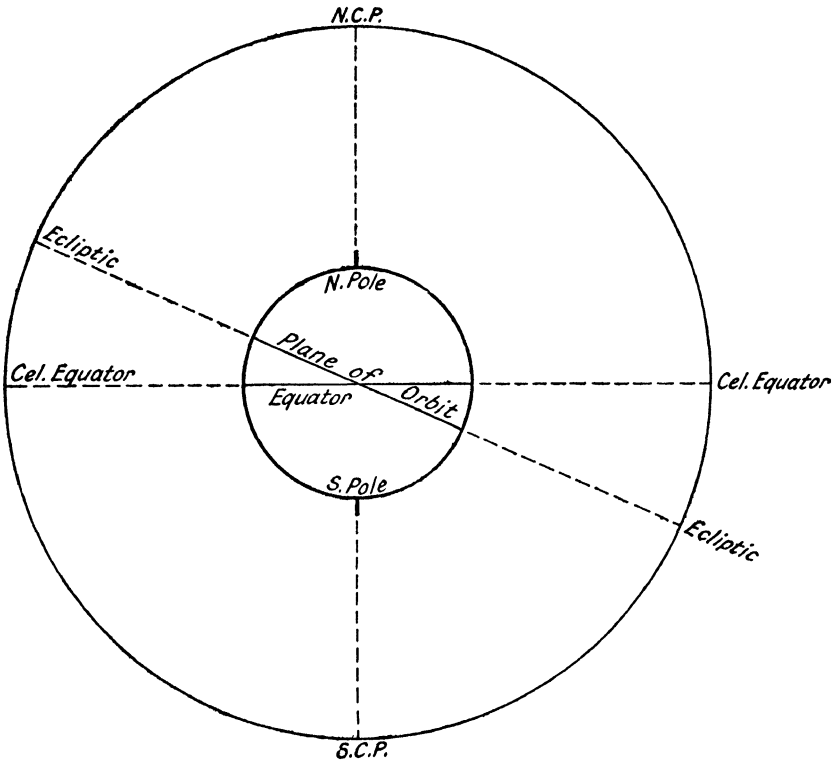


Figure 25. Both earth and star sphere are represented in median section.

certain meteorological phenomena might also hint at it though these could not furnish a rigid proof. The earth's mass could be determined with Joly's Balance, as explained on p. 65, as well as by other methods not here described. And finally, a simple calculation based on the results of the Corinth Canal experiment (p. 5) would yield the curvature of the earth's surface, and hence the size of the earth itself.

The heliocentric hypothesis and the star sphere

Continuing down the list of observations which concluded Chapter I, we return to the star sphere. One of our first discoveries was that this sphere appears to rotate between two diametrically opposed poles, just as a child's globe rotates between two pivots which must of course be opposite one another. But since then we have discovered that the earth rotates on its axis and that it is this motion which causes the apparent rotation of the star sphere in the reverse direction. The north and south celestial poles must therefore

be the points on the star sphere where the produced axis of the earth would cut that sphere. It follows from this, of course, that the celestial poles are directly overhead for observers situated at the earth's poles; also that if the earth's equatorial plane were similarly produced, it would intersect the star sphere at the celestial equator. Now since the ecliptic, which is the plane of the earth's orbit produced to the star sphere, is inclined to the celestial equator at about $23\frac{1}{2}°$, the earth's equator must be inclined to the plane of its orbit at a similar angle. The earth's axis is therefore not perpendicular to the plane of its orbit, but is inclined to it at an angle of about $66\frac{1}{2}°$. All these relations should be made clear by Fig. 25.

The meaning of the zodiac

Finally, there is the fact that all the planets move in paths lying closely adjacent to one another on the star sphere. The significant point here is that they also lie close to the yearly path of the sun, the ecliptic. Since this coincides with the plane of the earth's orbit, all the planetary orbits must lie in nearly the same plane.

III

SPANNING THE SOLAR SYSTEM

THE knowledge which we have so far arrived at has been obtained solely with the aid of the human eye, and without recourse to telescopes, spectroscopes, or other instruments. It is true that at several points during the development of the argument appeal has been made to telescopic evidence—the phases of Mercury and Venus, for instance, and the aberrational displacement of the stars—but such appeals were no more than short cuts to conclusions that could be reached without them. For Copernicus conceived his first intimations that the sun and not the earth is the ruler of the solar system before the invention of the telescope; and Kepler, although the telescope was actually being used for astronomical purposes by Galileo and others before he had formulated his three laws, worked throughout on Tycho Brahe's observations which were made with the naked eye.

Introduction of the telescope

From now onwards the naked eye must be assisted by a variety of instruments if the growth of our astronomical knowledge is to continue. At first, the telescope will only be required as a precision instrument—that is, as a means to the more accurate estimation of the positions of bodies than could be possible with the naked eye alone. The optics of the instrument need not detain us here, for even the most sceptical man of common sense will admit (i) that a telescope magnifies, and (ii), that if we see a thing with the aid of a telescope, we are justified in believing in its objective existence even though it may be invisible to the naked eye. He may not have realized, however, that the function of the telescope is twofold. In the first place, it magnifies: it increases the apparent separation of all points within its field. In the second, it increases the apparent brightness of an observed object—at any rate within limits. Thus not only do stars appear to be brighter in the telescope than without it, but stars which are otherwise too faint to be seen at all are rendered visible.

Astronomical distances: preliminary

One of the most important aspects of the astronomer's work consists of the determination of the distances of remote bodies. In this work he is fortunate in having a number of independent methods at his disposal, and, in theory at least, the discovery of the distances

of most celestial objects is comparatively simple. The practical difficulties, on the other hand, are usually great, and except in the case of the nearer bodies such as sun, moon and planets, the margin of inaccuracy is always undesirably wide; furthermore, the more distant an object is, the wider does this margin of uncertainty become[1]. The task of measuring the distance of an object 1,000 units away is more than twice as formidable as measuring that of an object 500 units away. Not only the practical difficulties, but also the effects of instrumental, personal, and systematic errors, are increased out of proportion to the increase in distance. Whereas a probable error of 0·01 per cent. in the estimated distance of the sun amounts to only a few thousand miles, an error of, say, 20 per cent. in the case of the remotest members of one stellar system may result in a margin of inaccuracy several thousand times as great as the distance from the sun to its nearest stellar neighbour.

There is a further source of inaccuracy which inevitably blurs the clarity of our distance determinations of the more remote celestial bodies, and which may even, without our suspecting it, entirely invalidate them. The basis of all these measurements is the simple trigonometrical method used by surveyors and map-makers. But this can only give a worthwhile degree of accuracy when confined to the nearer stars; for more remote objects, indirect methods must be used. These, though admittedly based upon the results of the fundamental trigonometrical method, nevertheless involve new assumptions, new sources of potential inaccuracy, and the use of analogies whose validity is less certain than it might be. At each successive extrapolation the chances of accurate results are diminished. Thus not only the mechanics of the problem, but also its principles become less accurate the further we reach out into space. Even the basic method, from which all others are ultimately derived, is not without its tacit assumptions. It has to assume, for instance, that light travels in straight lines—the so-called rectilinear propagation of light. It is only with this proviso that the trigonometrical operations are valid. And although it is within limits a safe assumption to make (and the results of its use likely to be accurate), scientists have nevertheless grown less dogmatic and sure of the infallibility of their fundamental assumptions than were their fathers in the hey-day of cocksure Victorian research, when the universe appeared to be unfolding itself as a structure built upon the principles of the naïve mechanical materialism then in vogue.

For all these reasons the astronomer counts himself most fortunate that he has independent methods of unlocking the secret of

[1] This is not true of determinations by radar.

astronomical distances. When the results obtained by two or more methods are in close mutual agreement, he may rest assured that they are accurate. For although any one method may give a wrong answer, it is stretching probability too far to suggest that a second and a third method—quite unconnected but also fallacious—should give identically the same wrong answer.

The purpose of these cautionary remarks is to dispel at the outset any misapprehensions the reader may have nourished as the result of casual readings in 'popular' scientific periodicals or even in popularized approaches to astronomy written by professional astronomers. We have a less exact knowledge of the scale of the universe, of the distribution of the stars and globular clusters, of the dimensions of our galaxy, and of the distances and real spatial distribution of the mysterious spiral nebulae, than the reader of this type of literature (with its inevitable simplifications and avoidance of constant qualifications) might be led to conclude. Or rather, there is a margin of uncertainty associated with all such statements of whose width, or even existence, he may be insufficiently aware. The greater the distance under discussion, the more important is it to bear this in mind if a reliable picture of our present knowledge is to be acquired.

Trigonometrical parallax

When a surveyor wishes to determine the distance of a not readily accessible object such as a mountain peak, he does not pace out or measure with a chain the intervening space between the peak and his place of observation. Instead, he measures a convenient baseline with the greatest accuracy (a few millimetres in 10 miles is possible), and then notes the direction of the peak from each end of it; that is, he measures the base angles of a triangle whose sides are the known baseline and the lines joining its extremities and the peak (Fig. 26). Assuming that the light reflected from the peak into the lens of his theodolite is travelling in undeviating straight lines, and that therefore his measurements of the base angles are correct to the limit imposed by the instrument itself, these data permit him to calculate the required distance trigonometrically. The great utility of this method lies in the fact that it enables the surveyor to determine the distance of a point without going there. It can be seen from the figure that the greater the distance, the smaller becomes the angle subtended by the baseline at the object. This is the reason for the exaggeration of systematic errors at great distances. Let us suppose, for example, that an instrumental defect such as backlash in the telescope bearing introduced a systematic error of 1′. If the subtended angle were 10° this error would amount to less then 0·2 per

cent. But if the peak were so distant that the subtended angle were reduced to ½°, then the error would be 1 in 30, or 3·3 per cent. We can also discern here the reason why trigonometrical parallax cannot be employed on objects whose distance is greater than a certain limiting value. For with any given baseline a limit will eventually be

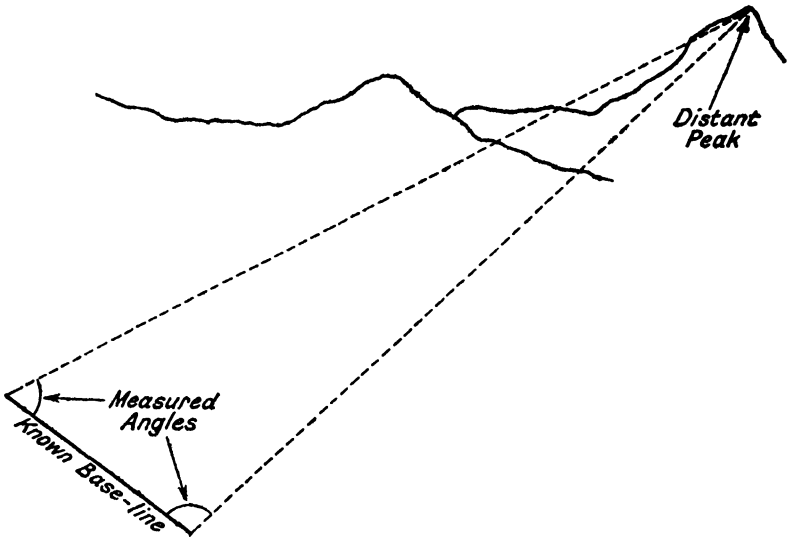


Figure 26.

reached at which the angular differences recorded at its two ends are so small as to be swamped by the estimated error, or else are too small for instrumental measurement; in the latter case the lines AB and AC may for all practical purposes be regarded as parallel. The only ways of escape from this difficulty are to increase the length of the baseline, or else to construct a series of triangles from the original baseline towards the peak via a number of intermediate points. Unfortunately the latter course is not open to the astronomer (since he cannot transport himself and his instruments out into space), and the former cannot be followed beyond a certain point.

The distance of the Moon

These difficulties do not arise, however, in the case of a body as near to the terrestrial observer as the moon, where the direct trigonometrical method based upon parallactic observations from two stations separated by an accurately determined distance, gives results correct to within 0·0005 per cent. Fig. 27 represents the earth and the moon. *M* is the moon, *WXY* the earth with its centre at *O*, and *X* and *Y* two points on its surface on or near the same meridian of longitude, and widely separated from one another. Since *OZ'*

and OZ'' are perpendicular to the horizon at X and Y respectively, Z' and Z'' must be the zeniths at these two stations. Hence the angles $Z'XM$ and $Z''YM$ are the zenith distances of the moon as observed from X and Y. These distances are measured simultaneously, and

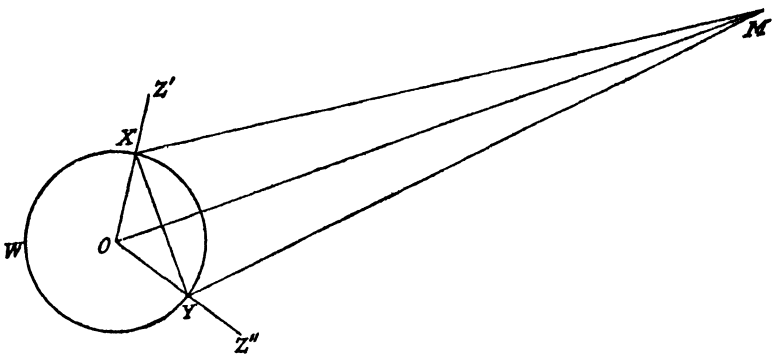


Figure 27. The parallactic determination of the moon's distance.

with the greatest degree of accuracy possible, at the two observatories. From these data the angles OXM and OYM can be derived, for

$$OXM = 180° - Z'XM$$
$$\text{and } OYM = 180° - Z''YM$$

Since the lengths of OX and OY are known (they are the earth's radius), the quadrilateral $MXOY$ can be solved trigonometrically. Hence the distance OM—that separating the moon and the earth's centre—can be deduced.

It is found in this way that the mean distance of the moon is 238,860 miles, the variation on either side of this mean value amounting to some 15,000 miles.

The form of the moon's orbit

In Chapter II we learnt it to be an inescapable fact that the moon revolves about the earth. We now know its mean distance. But since the orbit is elliptical, and not circular, its distance will steadily vary and can be determined at any given moment either by simultaneous parallactic observations, or else by constructing a 'spider' and thus determining the actual form of the orbit. This simple graphical method involves (i) the accurate determination of the distance that the moon has travelled across the star sphere from a given starting point on a large number of occasions throughout the month, (ii) the accurate measurement of its apparent diameter on each of these occasions. From a point E (Fig. 28), representing the earth, lay off a line EM_1 to represent the direction of the moon at the time of the

first observation. If at the subsequent three observations it has travelled 2°, 5° and 8° from the starting point, lay off EM_2, EM_3, EM_4 such that $\angle M_2EM_1=2°$, $\angle M_3EM_1=5°$, $\angle M_4EM_1=8°$; and so on through the entire lunation. These 'spider's legs' then represent the direction of the moon from the earth on successive occasions; and on each of these its apparent diameter was measured. Now since the moon's apparent diameter is inversely proportional to its distance—the nearer we are to an object, the larger it appears—our table of

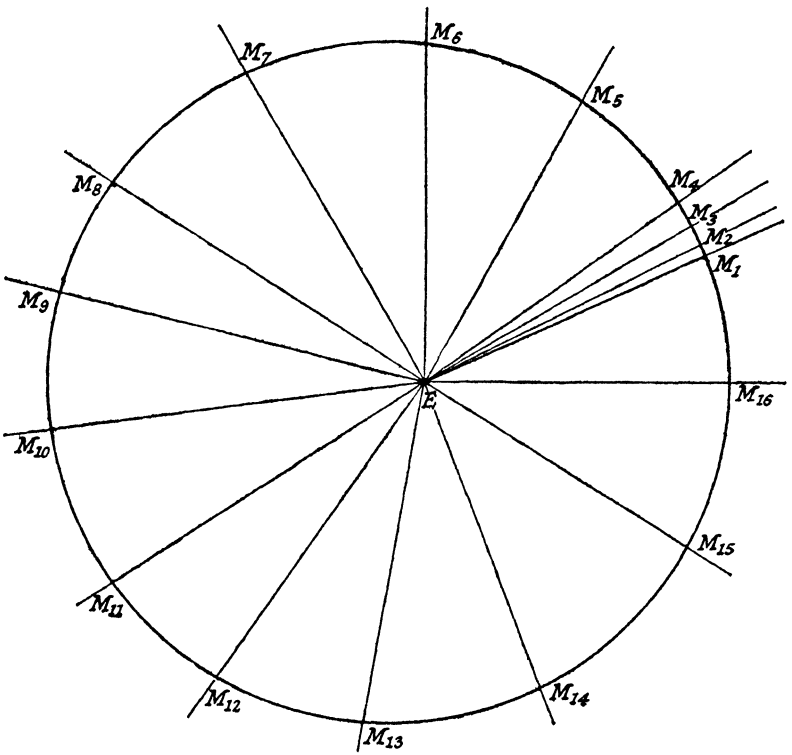


Figure 28.

angular diameters tells us the relative distance of the moon when in positions M_1, M_2, M_3 . . . If, then, on each leg of the spider we lay off a length inversely proportional to the moon's angular diameter at the corresponding time of observation, and then join all these points by a smooth curve, we have drawn the true form of the lunar orbit. And since, finally, we already know the linear length of one of the legs (the moon's mean distance, 239,000 miles) we can easily calculate the linear lengths of all the others.

In this way it is established that the moon's orbit is a nearly circular ellipse; and that, as we have been led to expect, the earth lies at one

focus while the radius vector describes equal areas in equal times; the mean orbital velocity being rather less than 2,300 m.p.h.

The lunar orbit and the ecliptic

If the moon's path among the stars is plotted on a star map it is discovered that this geocentric path is inclined to that of the sun in the same way that the latter is inclined to the celestial equator (Fig. 7). In other words, the plane of the moon's orbit is inclined to the ecliptic (or the plane of the earth's heliocentric orbit), intersecting it at diametrically opposite points known as the nodes. To discover the degree of this inclination we have only to measure the maximum angular distance of the moon from the ecliptic, for clearly θ in Fig. 7 (the angle in question) is that bounded by the two orbital planes. The inclination is about 5°.

Summary

Thus by progressive stages we have learnt the following facts about the position and motions of the moon:

It travels in a nearly circular orbit, at one focus of which lies the earth.

Its mean distance from the earth is 239,000 miles, and its mean orbital velocity some 2,300 m.p.h.

Its orbit is inclined to that of the earth about the sun at an angle of approximately 5°.

Relative solar distances

This is the first step towards constructing in imagination a scale model of the solar system. Much of the effort preliminary to taking the second step—that of establishing the exact spatial relationship of the sun to each of the planets—has already been made. In the last chapter we saw how Kepler first used a geometrical method of determining the relative solar distances of the planets, and thence deduced his much more valuable period-distance relationship. Direct observation gives the synodic period of each planet; the application of the $1/S=1/P-1/Y$ rule gives its sidereal period; and from the sidereal period Kepler III gives its heliocentric distance in terms of the earth's distance. The *relative* distances of all the planets from the sun are thus known, but until *one* linear distance is measured, our picture of the solar system must remain a map perfectly proportioned but lacking a scale. The sun is the most obvious celestial body and we might as well begin with it: for when we have determined its distance we shall have the one linear heliocentric distance which we require—that of the earth.

The sun's distance from planetary distances

This fundamental astronomical distance cannot be deduced accurately by the trigonometrical method direct, for the sun, unlike the moon, is too distant for a baseline on the earth's surface to yield a satisfactory parallax. But there is a variety of alternative methods to choose from, and the close correspondence between the results derived from them induces confidence both in their validity and their accuracy.

If the sun itself is too distant for satisfactory investigation by the direct method, some of the nearer planets are not. The nearest major planet outside the earth is Mars, and when the two planets are nearest one another the gap separating them is only about one-third of that separating the earth and the sun. An accurately measurable parallactic displacement from widely separated terrestrial viewpoints is yielded for this smaller distance, and the linear distance EP (Fig. 21) can be found. For purposes of illustration, let us suppose that it is 46,500,000 miles. We also know, from Kepler III, the relative lengths of SE (the heliocentric distance of the earth) and SP (that of the planet); suppose $SE=\frac{2}{3}SP$. Then

$$SE=\tfrac{2}{3}(SE+46{,}500{,}000) \text{ miles},$$
$$=93{,}000{,}000 \text{ miles}.$$

The nearer the planet, the greater will be its parallax and the more accurate will be the result. For this reason some of the minor planets, or asteroids,[1] have yielded better results than Mars. The asteroid Eros may approach within 14 million miles of the earth, and in 1931 approached to within 16 million; observations of Eros on this occasion have produced the most accurate determination of the sun's distance yet made: the value being 93,005,000 miles. Another method, involving the parallactic observation of a transit of Venus across the sun's disc, gives somewhat less accurate results, though providing valuable confirmation of those based on the observation of Mars and Eros.

The sun's distance from the aberration of light

An entirely independent method employs the phenomenon of the aberration of light. We have seen (p. 38) that

$$\tan\theta=\frac{v}{c}$$
$$\text{whence } v=c\,.\tan\theta,$$

where v is the earth's orbital velocity, c the velocity of light, and θ

[1] See p. 190.

the observed displacement of a stellar image situated at right angles to the direction of the observer's motion.

The story of Römer's discovery of the finite velocity of light from observations of the satellites of Jupiter will be told in Chapter VII. Two experiments devised nearly a century ago by the French physicists Fizeau and Foucault determined this velocity with a high degree of precision. Fizeau's apparatus is illustrated diagrammatically in Fig. 29. S is a source emitting a narrow beam of light; M_1 is a lightly silvered mirror which, while reflecting the beam from S towards M_2, a second mirror, yet allows the returning beam to pass through it to the observer's eye at O. A is a toothed wheel of known circumference which can be rotated at known velocities; its position is adjusted so that its teeth just intercept the beam between M_1 and M_2. When the wheel is rotated a series of light flashes will be transmitted towards M_2. At high speeds these will appear to coalesce and form a continuous beam, for the same reason (the

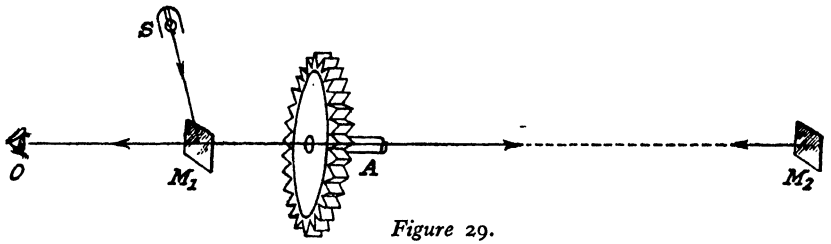


Figure 29.

persistence of vision) that the rapid succession of stationary images on a cinema screen gives rise to the illusion of smooth motion. If the speed of the wheel is steadily increased, a certain stage will be reached at which each of the teeth has moved half the distance separating it from its companion during the time required for the light to travel from A to M_2 and back to A. Hence on reaching A each flash will encounter, not the gap between teeth which it originally passed through, but one of the teeth themselves. All the returning light will be blocked by the cog teeth and it will appear to the observer at O that the light has been extinguished. The wheel velocity when this occurs, its size, the number of teeth, and the distance between the wheel and the second mirror are all the data required for the calculation of the velocity of light: the first three data will give the length of the interval required for the teeth to move round through half the distance separating them; and this is the time required for the beam to travel twice the known distance AM_2. Carried out under the most stringent experimental conditions, the modern modification of this method yields 299,776 km. per second (186,272 m.p.s.), which is the accepted value of the velocity of light.

In the equation $v=c.\tan\theta$, we therefore have only one unknown, v; for $c=299{,}776$ km./sec., and observation gives $\theta=20''{\cdot}5$, the so-called constant of aberration.

$$\begin{aligned}\text{Hence } v &= 299{,}776 \,.\, \tan\theta \\ &= 29{\cdot}8 \text{ km./sec.} \\ &= 18{\cdot}5 \text{ miles/sec.}\end{aligned}$$

Since there are 31,558,150 seconds in one sidereal year, the circumference of the earth's orbit $=18{\cdot}5\times31{,}558{,}150$ miles. Hence its radius, the distance from the earth to the sun,

$$\begin{aligned}&= \frac{18{\cdot}5\times31{,}558{,}150}{2\pi} \\ &= 92{,}900{,}000 \text{ miles,}\end{aligned}$$

which agrees well with the results of the parallactic method.

Other methods of determining the sun's distance

Other means of deriving the solar distance which should be mentioned, though they do not call for detailed descriptions, involve (i) an independent spectroscopic method of determining the earth's orbital velocity which employs the so-called Doppler principle,[1] (ii) the study of the perturbing effect of the earth's gravitation upon the motions of Venus and Eros, (iii) a somewhat similar study of the inequality in the lunar motion. All these methods permit the determination of the earth's mean solar distance with an inaccuracy not in excess of about 0·013 per cent., or 12,000 miles in 93,000,000. Since the orbit is not strictly circular the actual distance varies from midsummer to midwinter by rather more than a million miles on either side of the mean value.

The mass of the sun

Because the mass of the sun is a quantity which we shall require later, for the determination of certain stellar distances, it will be convenient at this point to see how Newton's work made its derivation possible. Since the force exerted upon one body by a second is proportional to their masses and inversely proportional to the square of their separation, it follows that the relative attractions exerted by the sun (f) and the earth (g) upon a body of unit mass situated at the earth's surface are given by

$$\frac{f}{g}=\frac{\dfrac{M}{R^2}}{\dfrac{m}{r^2}}$$

[1] See p. 153.

where M=mass of sun,
m=mass of earth,
R=distance of sun,
r=radius of earth.

Of the quantities in this equation,

M is the unknown,
r we have already discovered (Chapter I),
m we have already discovered (Chapter II),
R we have already discovered (Chapter III),
g, the force of gravity at the earth's surface, can be found from experiments with pendulums, and by a variety of other means,
f is easily calculated from Newton's laws once the earth's orbital velocity is known (Chapter III).

The value obtained for M by this and other methods is approximately 332,000 times that of the earth.

The orbits of the planets

Just as in the case of the moon, we may determine the inclination of each planet's orbit to that of the earth, thus completing our three dimensional scale model of the solar system, some of whose data are given in the subjoined table.

Planet	*Synodic period (days)*	*Sidereal period (years)*	*Mean heliocentric distance (cf. earth)*	*Mean linear heliocentric distance (million miles)*	*Mean orbital velocity (m.p.s.)*	*Orbital inclination to plane of ecliptic*
Mercury	116	0·241	0·387	36·0	29·5	7° 0′
Venus	584	0·615	0·723	67·2	21·9	3 24
Earth	—	1·000	1·000	93·0	18·5	0 0
Mars	780	1·881	1·524	141·5	15·0	1 51
Jupiter	399	11·862	5·203	483·3	8·1	1 18
Saturn	378	29·458	9·539	886·1	6·0	2 29
Uranus	370	84·015	19·191	1783	4·2	0 46
Neptune	367	164·783	30·071	2793	3·4	1 47
Pluto	366·7	247·7	39·46	3666	2·9	17 9
How derived	Observation	$\frac{1}{S}=\frac{1}{P}-\frac{1}{Y}$	Kepler III	Kepler III + one linear determination	Kepler II	Observation

The scale of the solar system

If the information contained in the fifth column is expressed in a different form it may convey a more concrete impression of the immense size of even our parochial corner of space. Light, travelling with a velocity of 186,000 m.p.s., covers a distance equal to nearly

eight round-the-world trips every second. To travel from the moon to the earth it requires rather more than $1\frac{1}{4}$ seconds; and from the sun to the earth, about $8\frac{1}{3}$ minutes. If the sun suddenly 'went out', we on earth would continue to bask in its rays for a further $8\frac{1}{3}$ minutes. The corresponding times for the other planets are given below:

Mercury	3 mins.
Venus	6 mins.
Earth	8 mins.
Mars	13 mins.
Jupiter	43 mins.
Saturn	1 h. 20 mins.
Uranus	2 h. 40 mins.
Neptune	4 h. 10 mins.
Pluto	5 h. 30 mins.

IV

BRIDGEHEAD AND BREAK-THROUGH

IF the reader looks at this page, while rapidly opening and closing alternate eyes, it will appear to him to jerk from side to side against the more distant background, whatever that may be. This simple example of parallactic shift depends upon the fact that the reader is using two viewpoints which are about two inches apart. So small a baseline gives a noticeable shift only for near objects. To produce a parallactic displacement in more distant objects, some bodily movement is required, while the two viewpoints of a body as distant as the moon must be separated by a thousand miles or more. A baseline of this magnitude also gives reasonable displacements for the nearer planets, such as Mars and Eros, but when the stars are studied with a view to determining their distances it is found that their relative positions as seen from the two stations on the earth's surface are indistinguishable, no matter what instrumental refinements are employed.

Clearly, then, the limit which is necessarily associated with every baseline is reached at distances short of even the nearest stars: we have successfully stormed the defences of our local planetary system, but it now appears that the immensely more distant and numerous agglomeration of stars is going to put up a more determined defence against the attacks of astronomers. If the scientific advance is not to be stemmed, astronomers must forge a new weapon, capable of forcing a bridgehead in these defences; and then, if necessary, new tactics to exploit and expand the break-through.

Heliocentric parallax

The weapon which finally achieved this breach had been talked about for many centuries before Bessel first used it with success: it is the simple fact that the earth is not stationary in space, but revolves about the sun. Over two millennia ago, Aristotle had derided the conception of a moving earth, quite reasonably pointing out that such motion of the terrestrial viewpoint would cause apparent displacements among the constellations as between the nearer and the more distant stars. Since it was impossible to detect any displacements of this nature, he argued that the earth must be stationary. The alternative explanation of this lack of visible parallactic displacement, an alternative which Aristotle rejected, was that possibly the

stars are so distant that the displacements, though they exist, are too small to be perceptible to the naked eye. It was this alternative that Bessel showed to be the correct one, thus driving another and by now quite superfluous nail into the coffin wherein are interred the remains of the geocentric hypothesis.

Since the earth makes one complete circuit of the sun in one year, it must occupy diametrically opposite positions in its approximately circular orbit at six-monthly intervals. And since the radius of this orbit is 93 million miles, the terrestrial observer's position in space alters by 186,000,000 miles between, say, 1 January and 1 July. It is this longer baseline which Bessel showed to be capable of yielding perceptible shifts in the case of the nearer stars; but this achievement involved the use of instruments such as neither Aristotle nor Kepler had dreamed of. When the potentialities of this baseline are exhausted, the terrestrial astronomer has come to the end of his resources so far as the determination of stellar distance by means of trigonometrical parallax is concerned: in order to be able to use this method for the measurement of the distances of stars lying beyond the reach of the 186-million-mile baseline provided by the earth's orbit, he would have to transport himself to one of the outer planets—for there astronomers, if they existed, would have the advantage of still greater orbital diameters. Fortunately, however, the breach having once been made, other and more powerful weapons are at hand to press home the advantage, and to ensure that astronomical investigation shall utilize the bridgehead established by Bessel for a full-scale break-through. These methods will be described in detail later in this chapter; for the present, let us inquire more closely into the trigonometrical method of determining a star's parallax.

Trigonometrical parallax

Fig. 30 represents (i) *X*, a sheet of glass upon which a circle has been engraved with a diamond, (ii) a small ball, *Y*, conveniently mounted, and (iii) *Z*, a screen. If an observer places his eye behind *A* and looks at the ball, its position as projected against the screen will be *a*; as the observer moves his eye round the circle towards *B*, the position of the ball against the screen will change to *b*; when the observer is looking from *C*, the apparent position of the ball will be *c*; and from *D*, at *d*. Thus the ball has traced out upon the screen an exact replica of the circle *ABCD*. The size of this circle, *abcd*, will depend upon the distance of *Y* from *X*—the greater it is, the smaller will be the circle. It is further to be noted that the apparent path of *Y* is only circular (as is the path of the observer's eye) when the line from *Y* to the centre of the circle *ABCD* is perpendicular to the sheet

of glass in which this circle lies. If the ball were moved to some position such as Y_1, situated *in* this plane, then its apparent movement against the screen (which can now be considered as the ceiling) will be a straight line: as the observer's eye travels round the circle *ABCDA*, the path of Y_1 will be from *a* to *b*, back through $c=a$ to *d*, an equal distance beyond $c=a$, and finally back to *a*. Balls situated in

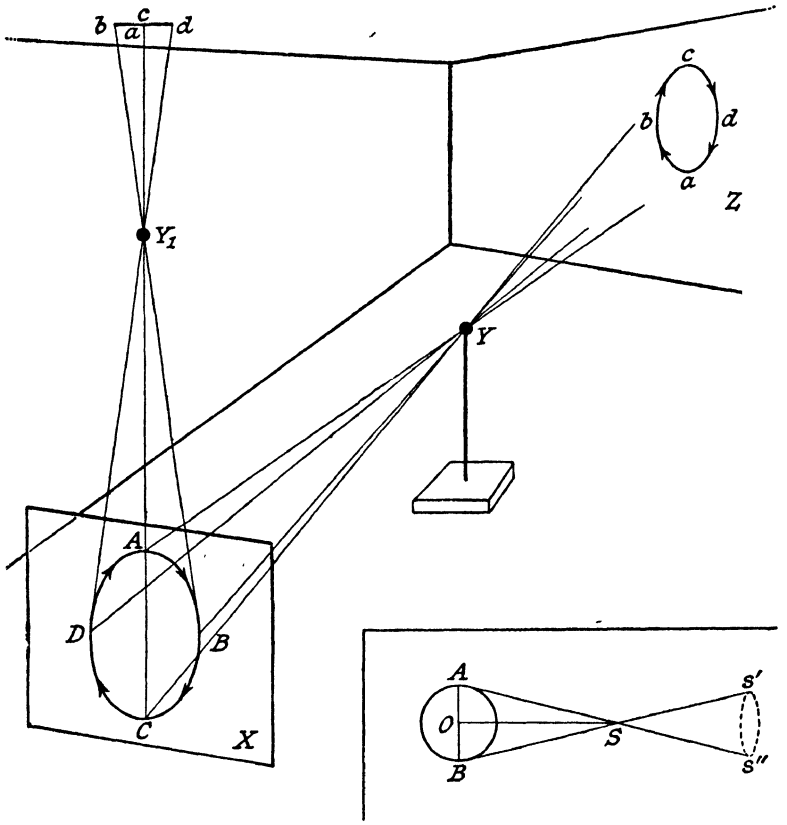


Figure 30.

positions intermediate between the projected plane *ABCD* and the perpendicular to it will describe ellipses whose eccentricity varies with their position. Finally, it is to be noted that the ball revolves with reference to its background in the same direction as the observer's eye, but 180° in advance of it.

So much by way of introduction. If we now rename *ABCD* the earth's orbit, *Y* a relatively near star, and *Z* the background of more distant stars, we have a replica of the mechanism of stellar distance determination by means of parallactic displacements resulting from the earth's orbital motion. Suppose a certain star is observed to shift

against the background of stars (which, since most of these are many times more remote, do not themselves shift appreciably) by a distance of 1″ in the course of six months. We then have $s'Ss''=ASB=1''$. Therefore at the star's distance (OS) the radius of the earth's orbit (AO) subtends an angle (ASO or BSO) of 0″·5. By trigonometry,

$$OS=\frac{93{,}000{,}000}{\tan 0''\cdot 5} \text{ miles.}$$

Or, writing D for the distance of the star (OS),
R for the radius of the earth's orbit (OA),
p for the star's measured parallax (ASO),[1]

$$D=\frac{R}{\tan p}.$$

Owing to the enormous distances of even the nearest stars, the annual parallactic displacements are excessively minute: the largest are no bigger than a halfpenny viewed from a distance of about four miles. This accounts both for the flaw in Aristotle's argument, and also for the fact that the displacements were not detected for over two hundred years after the invention of the telescope. It was not, indeed, until 1838 that Bessel detected and measured the first annual stellar parallax, that of the star known as 61 Cygni. He and subsequent workers proved observationally what we have already seen from our model, namely that, (i) the annual parallax described by a star at the pole of the ecliptic is a small circle, (ii) a star situated on the ecliptic travels back and forth along a short straight line which lies in the plane of the ecliptic, (iii) stars intermediate between these two positions describe small ellipses whose eccentricity is inversely proportional to their angular distance from the ecliptic itself. These configurations may be imitated by viewing a penny edge on, when it appears as a straight line, and then by turning it gradually over so that its visible shape becomes, first, a very flattened ellipse, and subsequently a less and less eccentric ellipse until finally, when viewed at right angles, it is circular. Observation has also shown (iv) that the parallactic orbits are described in the same direction as the earth's motion round the sun, but 180° ahead of the earth. This fact helps us to distinguish between parallactic and aberrational displacements, for it will be remembered that the latter involve a difference in position of 90°.

Another factor which has to be eliminated from the total observed displacement of a star is the component of its real space motion, if

[1] The definition of the annual heliocentric parallax of a star is: the angle subtended at it by the semidiameter of the earth's orbit. This is clearly half the *total* annual displacement.

any, at right angles to the observer's line of sight: this may be of negligible proportions, but it must nevertheless be considered. The modern method of detecting and measuring parallaxes is photographic, and was initiated by Schlesinger in 1903. Previously to that date the measurements had been made visually at the telescope, a laborious method which had yielded about sixty stellar parallaxes. Schlesinger's method, now universally adopted, is both quicker and more accurate, and also permits the measurement of parallaxes too small for detection by the older means; to-day some 3,700 trigonometrical parallaxes have been measured. The procedure is as follows. At an interval of six months two telescopic photographs are taken of the star whose parallax it is required to determine, exceptional practical precautions being observed. The relative positions of the star and the 'background' comparison stars are then measured on the two photographs with the greatest degree of accuracy possible. The movement, relative to the remoter stars of the background, thus revealed, will be compounded of parallactic displacement and actual proper motion. Consequently, a third photograph is taken after a further interval of six months. The parallactic displacement will once again be zero, while any displacement due to the star's motion will have been doubled. In this way the two may be disentangled: it may be necessary to take up to twenty photographs, covering a period of ten years, before satisfaction can be obtained that the effect of the star's motion has been eliminated.

The moment we come to mention the results of these investigations we encounter what can only be called 'astronomical' numbers: that is to say, numbers so large that they are confined in human experience to the realm of astronomy. At this point, therefore, it will be advisable to digress long enough to explain two notations which have been devised for dealing (i) with very large numbers, and (ii) with very great distances.

The index notation

The function of the first is to express unmanageably large numbers in a compact and easily written form. When a number is multiplied by itself it is said to be 'squared', and is written with a small 2 above and to the right of it; when multiplied by itself twice, to be 'cubed', when the index 2 is replaced by a 3. Thus:

$10 \times 10 = 10^2 =$ 100
$10 \times 10 \times 10 = 10^3 =$ 1,000

Similarly

$10^4 =$ 10,000
$10^5 =$ 100,000
$10^6 =$ 1,000,000

In fact, the index tells us how many ciphers follow the initial unit.

But this notation is not limited to the expression of exact powers of 10. Intermediate numbers can be expressed with equal facility, and nearly as concisely. Consider the number 15,000.

$$10{,}000 = 10 \times 10^3$$
$$20{,}000 = 20 \times 10^3$$

Therefore $$15{,}000 = 15 \times 10^3$$

By convention, however, the factor not containing the index always lies between 1 and 10. Thus we divide the 15 by 10, making it $1{\cdot}5$, and, in order not to alter the value of the whole expression, multiply the other factor by 10, making it 10^4. Thus 15,000 can be rewritten $1{\cdot}5 \times 10^4$.

The rule to remember when expanding the contracted notation into the longhand form is: move the decimal point to the right a number of places that is indicated by the index. In our example the index is 4; hence the value of the expression is

1 5,0 0 0·0
1 2 3 4

Similarly, $1{\cdot}5 \times 10^6 = 1{,}500{,}000$; $7{\cdot}542 \times 10^9 = 7{,}542{,}000{,}000$, and so on.

The very great convenience of this device will be brought home by the fact that instead of having to write, for example, a 1 followed by enough ciphers to cross the entire page from margin to margin, we may simply write 10^{69}.

Astronomical units of distance

The second notation is one of distance. When the distances of the first stars were measured it was found that they were so staggering as to render their expression in any unit as small as the mile prohibitively laborious; even the 'astronomical unit' (the earth's mean distance from the sun, or 93 million miles) is in little better case. A very much larger unit had to be discovered, and for all scientific purposes the 'parsec' is now used. This may be defined as the distance at which a star would have an annual parallax of 1″. Though an extremely convenient unit to work in, the parsec is not particularly revealing of the actual distance involved: one may form a clear mental picture of how long a mile is, and even some sort of idea of the astronomical unit, but the parsec is singularly reticent in this respect. For popular works, intended rather for the general public than for fellow astronomers, the 'light year' is accordingly more commonly used. We have already learnt something of the velocity of light, that

it takes $8\frac{1}{3}$ minutes to travel from the sun to the earth, and a further $5\frac{1}{4}$ hours before it reaches the orbit of Pluto. We may therefore say that the distance from the sun to the earth is $8\frac{1}{3}$ light minutes, and from the sun to Pluto, $5\frac{1}{2}$ light hours. In the same way, 1 light year is the distance covered by light in one year, travelling at an invariable velocity of 186,000 m.p.s. It is as many times greater than the earth's distance from the sun as 1 year is greater than 8 minutes; a quick calculation will prove that it is therefore about 63,300 astronomical units. This fact provides a further way of giving some reality to the conception of a light year, for there are 63,360 inches in one mile. So we can say that there are as many astronomical units in a light year as there are inches in a mile.

A word must now be said of the relation between light years and observed angular parallax. The distance of a star whose parallax is 1″ is 1 parsec, which is equivalent to 3·26 light years. If the parallax is 0″·1, the distance will be 10 parsecs, if 0″·01, 100 parsecs, and so on. Hence, putting D for distance, and p for parallax expressed in seconds of arc,

$$D=\frac{1}{p} \text{ parsecs,}$$

$$\text{whence } D=\frac{3\cdot26}{p} \text{ light years.}$$

Density of stars in space

To take an example, the nearest star yet discovered is a faint object in the southern constellation of the Centaur, to which the name Proxima Centauri has been given. Its parallax, the largest yet detected, is only 0″·785 (this is about the angle subtended by a plate, two feet in diameter, when viewed from a distance of fifty miles).

$$\text{Therefore } D=\frac{3\cdot26}{0\cdot785}$$
$$=4\cdot2 \text{ light years.}$$

This is equivalent to about 265,000 astronomical units.

It thus becomes clear, not only that enormous distances are involved when we come to the stellar system, but also that space is exceedingly 'empty'. The distance separating the sun from its nearest neighbour is about $4\frac{1}{2}$ light years, while its own diameter is less than $4\frac{1}{2}$ light seconds, or about one thirty-millionth of that distance. So far, 18 stars have been discovered with distances smaller than 12 light years, though there are certainly other faint neighbours, yet undetected. This figure gives a density of stars in space of about 1 in every 400 cubic light years!

Limitations of trigonometrical parallax

Finally, a word on the accuracy and scope of Schlesinger's direct photographic method of measuring stellar parallaxes. The margin of probable error varies from about 1 per cent. for the very nearest stars, to from 10 per cent. to 15 per cent. for stars at a distance of about 80 light years. Thus the distance of a star, measured as 80 light years with the greatest possible refinement, would be correctly stated as 80 ± 8 light years: it might lie anywhere between 72 and 88 light years, but is probably nearer 80 than to either of these limits. Up to about twice this distance—say, to 150 or 200 light years—tolerable results may be expected, but for still greater distances the method is practically useless: the potentialities of the longest baseline available to us have now reached exhaustion.

Thus we have, with a vengeance, established our bridgehead. But the momentum of the attack peters out at a distance of some 200 light years behind the enemy lines. Is it possible, nevertheless, to achieve a complete break-through?

Proper and radial stellar motions

It has been known for several centuries that the stars have motions of their own: that their relative positions upon the star sphere do not

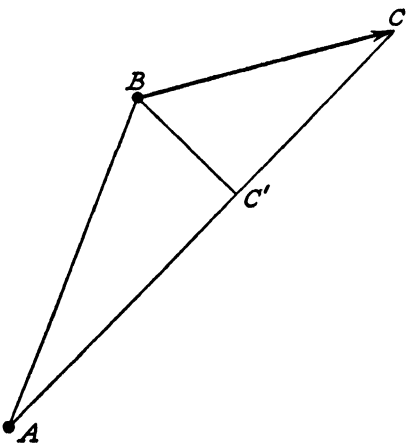


Figure 31.

appear to alter from year to year, or even from century to century, when observed with the eye alone, is solely due to their very great distances from the observer: with telescopic aid, the angular motions of a great number of stars have been detected and measured. Let us imagine that we are sitting in a captive balloon which is floating above a field (Fig. 31). Directly beneath us is a man, *A*, at a certain

distance from whom is a second man, B, who, in the space of time covered by our observations, walks to another position, C. It is clear that B has moved *across* A's field of vision by the distance BC', and has at the same time moved directly away from A by the amount $C'C$. B's space motion from B to C may therefore be resolved into two mutually perpendicular components, BC' and $C'C$, one of which lies in the observer's line of sight and the other at right angles to it. Supposing that for any reason the observer at A were unable to estimate B's true space motion, BC, he could nevertheless calculate this distance if he were able to determine BC' and $C'C$.

If we now call A a terrestrial observer and B a star whose space motion is BC: BC' is its angular motion across the star sphere, its proper motion; and $C'C$ is its motion in the line of sight, causing no displacement on the star sphere, known as the radial motion. The former, if it is large enough (and the two observations, of B and C, may have to be separated by many years before it is perceptible) can be measured with a micrometer attached to a telescope. How the latter may be deduced from an inspection of the star's spectrum will be explained in detail in Chapter VI: here let it suffice to say that it is only one of the many benefits conferred upon astronomers by the spectroscope that it allows the line-of-sight velocity of any light source to be measured directly, and with a very narrow margin of inaccuracy. Thus, provided they are above the threshold of instrumental measurement, both the proper and radial motions of a star may be arrived at: the former measured in seconds of arc per century, and the latter in miles or kilometres per second.

An idea of the minuteness of angular stellar motions upon the star sphere may be gained when it is learnt that the largest proper motion yet detected is that of a faint star (invisible to the naked eye) discovered by Barnard, the annual displacement of which amounts to $10''{\cdot}25$—roughly one two-hundredth of the angular diameter of the moon. Among the brighter stars, Arcturus has one of the largest proper motions, having moved across the star sphere a distance equal to twice the apparent diameter of the moon since the time of Ptolemy.

The possibility of a greater baseline

Since the stars are in random motion relative to the sun, the sun is clearly moving—in a certain direction and with a certain speed—relative to them. Could such motion be accurately defined, we should have to hand a new and more extensive baseline upon which to base parallactic investigations. For no matter how long we wait, the earth's orbital motion cannot carry us more than 186 million

miles (the diameter of the terrestrial orbit) from any previous position. This is the limitation imposed by the fact that the earth moves repetitively in a closed orbit. But if it could be established that the sun is proceeding in a direct course with a velocity of x m.p.s. towards a certain point on the star sphere, the terrestrial observer's spatial displacement due to this motion would increase steadily with the passage of time. If the sun's displacement after one year were insufficient to reveal a parallactic shift for a distant star, it would only be necessary to wait another year (when the shift would be doubled), or a third (when it would be trebled); this could of course be carried on indefinitely until the observer's displacement were great enough to cause a measurable shift in the star's position. Theoretically there should be no limit to the scope of this method, owing to its indefinitely extensible baseline. One drawback we shall consider in a moment, but first the problem of the sun's motion must be briefly touched upon.

The sun's motion

The reader will probably have noticed, while driving at night down a long, straight road on whose either side is a row of street lamps, an 'opening out' effect among the lamps ahead of him; and the reverse effect, that the two rows of lamps are converging upon one another, when he looks out of the rear window. At the same time, the whole panorama of lamps is slipping backwards away from the apex of his motion and towards the direction from which he came (the antapex of his motion). This is, of course, an elementary effect of perspective. The existence of a similar effect among the stars may be used to determine the direction and velocity of the sun's motion among them, relatively near stars of known distance playing the role of the street lamps. The problem is complicated, without being essentially altered, by the fact that the stars have peculiar motions of their own—it is as though the street lamps were not fixed in the pavements, but were wandering erratically about. In order to detect the 'opening out' of those stars situated in the direction of the sun's motion, and the 'closing in' of those in the opposite direction, their space motions must first be determined micrometrically and spectroscopically, and these effects eliminated from consideration: if a sufficiently large number of stars is used as the basis of the investigation, individual anomalies will tend to cancel one another out (since the motions are random), and it is possible to arrive at a sufficiently accurate result dependent solely upon the sun's motion.

Herschel, in 1783, was the first astronomer to achieve this, and his

first approximate result has been abundantly confirmed and refined by subsequent workers, notably by Boss, Wilson, Campbell and Moore. The most accurate determination yet made is that of the latter two workers: the sun is moving towards a point on the star sphere on the edge of the constellation Hercules, not far from the first magnitude star Vega, with a velocity of 12·3 m.p.s. relative to the 2,149 stars used as 'street lamps'.

Statistical parallax

In a single year, therefore, the sun's motion displaces the observer by an amount twice as great as the diameter of the earth's orbit. If, now, the stars were stationary with respect to one another (the truly 'fixed' stars of the ancients) the observed proper motion of any star would be due solely to the sun's motion, and its distance could be derived direct from its observed shift. But since the stars all have motions of their own—known as their peculiar motions—it is impossible to disentangle the component of their total observed motion due to parallax from that due to peculiar motion, without first knowing the star's distance; which is precisely what we hoped to derive. It therefore seems that our proposed new method of determining stellar distance is to be stillborn.

There is, however, a compromise way out of this difficulty. Though it is true that the sun's motion cannot be utilized to determine the distance of *individual* stars (whose peculiar motions cannot be known), it can be utilized to give us the mean or statistical parallax of a *group* of stars. For if a large enough group of stars is considered, the random peculiar motions of its members will tend to cancel one another out, and the larger the group the more nearly will this be true. The drift of the whole group towards the antapex of the sun's way can then be directly related with the mean distance of all the members of the group.

Certain characteristics of such statistical parallaxes are to be noticed:

i. they can be deduced for stars too distant for investigation by either of the methods already described;

ii. they cannot be derived for individual stars, but only as an average for a number of stars;

iii. their accuracy depends upon (*a*) the number of stars for which the motions are deduced, and (*b*) the accuracy with which the proper motions are measured. Thus optimal results would be obtained for a group consisting of a very large number of stars, all

of whose proper motions were large enough to be measured with a high degree of accuracy.

The next stage

Statistical parallax, though reaching beyond the limit at which trigonometrical parallaxes are too unreliable to be of any value, is confined to stars far short of the most distant that are known to exist; it is, moreover, not applicable to individual stars. What is required is a method of determining the intrinsic luminosities of the brightest stars, so that even at distances so great that many stars may have faded into invisibility there shall still be tell-tale beacons by whose means those distances can be discovered: this has been achieved, the stars in question being of the types known as Cepheids and Novae.

Apparent stellar brightness

Before we can learn how astronomers have developed this field of inquiry it will be necessary to digress, as briefly as may be, on the subject of the apparent and real brightness of stars. The former is measured in terms of an arbitrary scale of 'magnitudes', a magnitude being the unit of apparent brightness just as a stone is the unit of human weight or a fathom the unit of water depth. Roughly speaking, the brightest stars in the night sky are of the first magnitude, stars a little fainter, of the second magnitude, and so on till we come to those which are only just perceptible with unaided vision; these are of the sixth magnitude. The faintest which are capable of impressing their images upon photographic plates after long exposure with the world's largest telescope are of about the twenty-second magnitude. It is important to realize that 'magnitude', departing from common usage, has nothing whatever to do with size: it is solely a measure of brightness, as this appears to the terrestrial observer.

Careful visual scrutiny of different stars, notably by Pogson and the younger Herschel during the earlier half of last century, showed (i) that each magnitude is about two and a half times as bright as the one below it, (ii) an average first magnitude star is about 100 times as bright as an average sixth magnitude star. These two estimations of the relative brightness of different magnitudes differ only slightly from one another; but differ they do, for if each magnitude were exactly 2·5 times as bright as that below it, then a first magnitude star would be $(2{\cdot}5)^5$ times as bright as one of the sixth—and $(2{\cdot}5)^5$ approximately equals 95, not 100. It therefore had to be decided which was the more convenient relation on which to base the arbitrary magnitude scale. Herschel's estimate (that a numerical decrease o

six magnitudes means a hundredfold brightness increase) won. If we put x for the relative brightness of adjacent magnitudes,

$$x^5 = 100$$
$$\therefore x = \sqrt[5]{100}$$
$$= 2{\cdot}512.$$

That is to say, a star of magnitude m is 2·51 times as bright as a star of magnitude $m+1$.

Stellar luminosity

This scale of magnitudes refers only to apparent brightness, and this gives no clue to the real brightness of a star—its so-called luminosity—for the former depends not only upon the latter, but also upon the star's distance from the observer. Stellar luminosity is also measured by means of a scale of magnitudes, known as absolute magnitudes to distinguish them from apparent magnitudes. The absolute magnitude of a star may be defined as the apparent magnitude it would have were it situated at unit distance from the earth. This arbitrarily chosen 'unit distance' is 10 parsecs—i.e. that distance at which the star would have an annual parallax of 0″·1—and is equivalent to about 33 light years. It comes to the same thing to say that the absolute and apparent magnitude scales cross at a distance of 10 parsecs.

Now since the absolute magnitude of a star states its apparent magnitude were it transported to 10 parsecs from the earth, it is clear that the absolute magnitude scale is one of luminosity. Suppose that an astronomer wishes to study the luminosities of two stars: A is very faint, only visible with a telescope, while B is one of the brightest in the sky—of the first magnitude, let us say. Until he knows their respective distances he can deduce nothing about their luminosities. But once their distances are known, it is possible to calculate what their apparent brightness would be if these distances were each adjusted to the same distance of 10 parsecs. It might then be found that the bright star, B, is really fainter than the faint star, A, its relative apparent brightness depending not upon its superior luminosity but upon its smaller distance—their respective absolute magnitudes might be 5 and 2. In other words, the distance of A is less than 10 parsecs, and therefore its absolute magnitude is lower than its apparent, while B is more distant than 10 parsecs, so that its absolute magnitude is higher than its apparent. If a star is situated at 10 parsecs from the sun, its apparent and absolute magnitudes will be the same.

Conversely, if we could discover the absolute magnitude of a star,

we could immediately deduce its distance, for we should know (*a*) how bright it appears to be at the distance we wish to discover, (*b*) how bright it would appear to be if its distance were 10 parsecs, (*c*) that the intensity of illumination from any source falls off in inverse ratio to the square of its distance.[1]

Successive extrapolations

The superlative importance of being able to determine the absolute magnitudes of stars is therefore evident. Trigonometrical and statistical parallax have provided us with a mass of information concerning the distances of the nearer stars, and the problem is to discover some characteristic of these stars which shows itself to be variable with absolute magnitude. To do this is merely to take one further step in the extended extrapolation represented overleaf. At each new stage in this sequence, data are discovered which transcend that stage, thus establishing a further stage; this in turn contains not only the data from the preceding stage which made possible its determination, but also new data which transcend it and allow a third stage to be established.

In a way, the process is not unlike that whereby a convict, incarcerated in a prison cell, managed to escape from the cell, the prison and the country, although his initial equipment consisted of nothing more elaborate than a piece of wire: with this he was able to pull into his cell the key which a careless warder had dropped on the floor outside; once out of the cell he was able to steal a ladder from a storeroom, and with its aid to climb through the window of the Governor's bedroom; there he was able to change into civilian clothes, steal a keyring, and let himself out of the prison by the Governor's private wicket; once outside, he was in a position to return to his home and collect his passport and some money; and, thus provided, he boarded a steamer and sailed for South America. Neither the wire, the key, the ladder, the suit nor the passport would of itself have secured his escape: but, using each in turn to extend the range of his freedom by an amount that placed the next within his grasp, the case was quite otherwise.

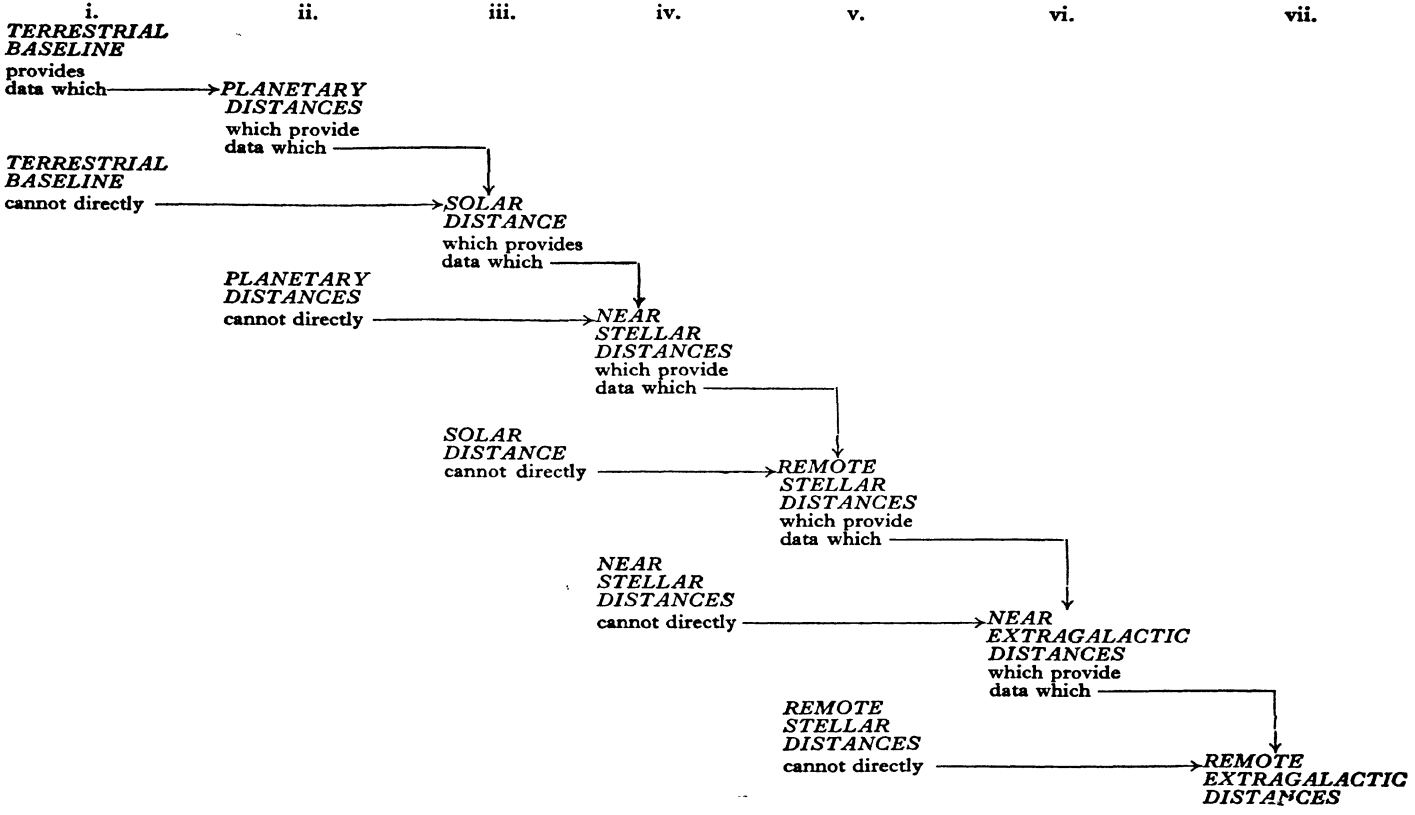
At the moment we are in the position of holding the key (statistical parallax)—which the piece of wire (trigonometrical parallax) provided

[1] It may be of interest to note that the exact form of this important relation between absolute magnitude, apparent magnitude, and distance, is:

$$M = m + 5 + 5 \log p,$$

where M = absolute magnitude, m = apparent magnitude, and p = parallax. If any two terms are known, the numerical value of the remaining unknown can be calculated at once.

Successive Extrapolations in the Determination of Astronomical Distances



for us—in our hand, and of looking round to see how it can be used to lead us to the next stage in our expansion.

Cepheids

This next stage, which transcends trigonometrical and statistical parallax while at the same time being based upon their results, is known as the period-luminosity relationship of the Cepheids. The importance of statistical parallax in establishing the much more valuable Cepheid method of determining stellar distance will appear in a moment. First, the Cepheids themselves, for a fuller account of which the reader may turn to p. 247. There are in the skies many stars whose brightness is not steady, but fluctuating: the so-called

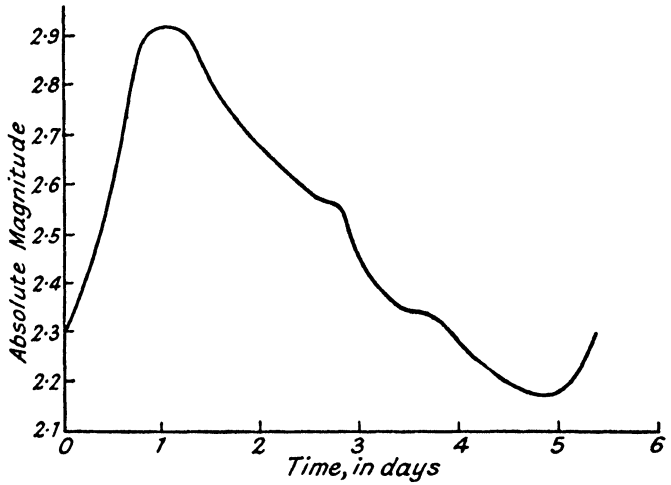


Figure 32. Light curve of δ Cephei.

variable stars. If the magnitude of one such star is measured at intervals over a given period, and the results incorporated in a curve whose axes represent magnitude and time respectively, this curve will be a 'picture' of the star's light variation, showing at a glance the manner in which the brightness increases and fades. The light curves of all variables fall into a comparatively few, easily distinguishable types, each star of a given type behaving in a similar manner to all other members of that type, and differing in obvious respects from variables of other types. The light curve of δ Cephei (the type star which has given its name to the Cepheid variables) is shown in Fig. 32; the steep rise to maximum brightness, followed by a more gentle decrease, is characteristic of all Cepheids.

The period-luminosity relationship

Situated in the southern half of the star sphere, and invisible from the latitudes of Great Britain, are two great clouds of faint stars which

to the naked eye resemble detached lumps of the Milky Way; they are known as the Greater and Lesser Magellanic Clouds. In 1908 it was discovered that the latter contained a number of Cepheids, and a most surprising connexion between their periods (i.e. the interval separating successive maxima) and their mean apparent magnitudes was found to exist: the longer the period of one of these Cepheids, the brighter it was. Since all attempts to detect parallaxes for individual stars in the cloud had proved abortive, it was clear that it must lie at a very great distance from the sun—as had already been suggested by the faintness of its stars. It was thus safe to say that its size was negligible compared with its distance, and that therefore all the stars of which it was composed were at materially the same distance from the observer. Under these circumstances, the apparent magnitudes of its stars would be directly proportional to their absolute magnitudes, and the correlation could be extended to the latter: thus, the period and luminosity of Cepheids are interrelated. But since the distance, and therefore the absolute magnitude, of no single Cepheid had at that date been determined, it was impossible to state the numerical relation between the two; in other words, the curves constructed in 1908 showing the relation between Cepheid period and mean absolute magnitude could not be calibrated. The reason for this failure was that the Cepheids are intrinsically brilliant stars and can therefore be seen at distances too great to be spanned by either the spectroscopic or trigonometrical methods: the nearest Cepheid yet discovered is situated at a distance of about 200 light years from the sun. It was not until 1917 that Shapley made the first determinations of Cepheid distances, thus enabling him to determine their absolute magnitudes and assign a zero point to the period-luminosity curve. His method was to determine statistically the average parallax of eleven Cepheids possessing unusually large proper motions. Thus a scale of absolute magnitude values was fitted to the uncalibrated curve, and thenceforth Cepheids could be used as celestial milestones wherever they were visible; and owing to their great intrinsic brilliance greater distances were measurable than had ever before been thought possible. The procedure consists simply in measuring the Cepheid's period, reading off the corresponding absolute magnitude from the curve, and comparing this with the observed mean apparent magnitude. This very powerful method has, particularly in the hands of Shapley, provided a great part of our present knowledge of the structure of the universe of stars; this application of Cepheid parallax will be returned to in the next chapter.

There is no reason for doubting the exactness of the relative values (i.e. those derived in 1908), but the absolute values of the calibrated

curve are certainly in some error, owing to doubtful fixing of the zero point, and they may even be gravely in error. This uncertainty of the zero point results from Shapley's inability to fulfil either of the conditions laid down in clause (iii) on p. 90: (*a*) the number of Cepheids used in the determination of the statistical parallax was only eleven, far too few to ensure that the derived value represented the true mean parallax, (*b*) owing to their great distances, the proper motions of the Cepheids are uniformly small, and even the largest, as selected by Shapley, are small enough for material error to enter into their measurement. The estimated inaccuracy from these causes amounts to about 0·5 magnitudes; at a later date Shapley was able to refine the original calibration by utilizing a larger number of Cepheids for the fixation of the zero point, although even now there is still a margin of uncertainty associated with all distances determined by this method. This disadvantage is to a certain extent counteracted by the great distances that can be investigated: better to have a rough idea of the distance of a very remote object than none at all. In Shapley's application of the period-luminosity law there is another source of possible error: we shall see in Chapter V that most of his results have been based upon observations of the so-called cluster variables, stars which, though being of the Cepheid type for which the relationship was worked out, nevertheless differ from true Cepheids in certain important respects—and this, also, must be borne in mind when considering Shapley's conclusions.

Novae as sounding lines

Another type of star that can be used as an approximate sounding line is the Nova.[1] Novae are stars which, for reasons still not exactly known, suddenly blaze up from comparative obscurity, increasing in brightness by many magnitudes in perhaps no more than a few days: at maximum a nova is often the most conspicuous object in the night sky. Maximum having been attained, a much lengthier and more erratic phase of fading sets in, till finally (usually not for several years) it has sunk to something like its original magnitude. Some of the nearest of these novae have been within the reach of methods of determining parallax already described, and it has been found that at maximum they not only appear abnormally bright, but are intrinsically so. Their absolute magnitudes characteristically lie between −5 and −10, the most favoured magnitudes being −6 and −7. If, therefore, a nova is discovered which is so faint as to require a large telescope to show it, the reasonable conclusion is that it is many times more distant, not only than the stars visible to the naked eye,

[1] See p. 248 for a fuller account of these objects.

but also than telescopic stars as bright as, or brighter than, itself. If we assume, as we can with reasonable safety, that its absolute magnitude lies between -5 and -10 we can derive a very rough idea of its distance; and if a number of novae have been observed which are known to occupy the same region of space, we can assume a mean absolute magnitude of -6 or -7 and achieve a considerably more definite result without sacrificing the assurance of at least approximate accuracy.

The reader may object that the Cepheid method has just been criticized for an uncertainty amounting to 0·5 magnitudes, yet we are now thinking worthy of serious discussion a method with an indeterminacy of at least several magnitudes. It must of course be admitted that individual distances derived by this method may be as many as ten times as uncertain as those based upon Cepheid observations. But where several novae in the same region of space are observed, this margin is considerably decreased. Moreover, novae have been mainly utilized in connexion with the class of objects known as extragalactic nebulae, concerning whose status there have been, as we shall see in Chapter V, two schools of thought: either they are members of our stellar system and therefore *comparatively* near the earth (their distances being of the order of 100,000 light years or less), or else they are external to it and situated at distances not much less than ten times as great. The value of the novae that have been detected in these nebulae lies in the fact that, even with the wide margin of uncertainty mentioned above, they provide a clear indication as to which of these widely different alternatives is to be preferred.

Dynamical parallax

Still another type of star can be used for determining stellar distances: the binary star. Most stars pursue their courses through space in lonely isolation, separated by many light years from their nearest neighbours. But an appreciable percentage of the stellar host are members of duple systems: that is, pairs of stars near enough to one another to be controlled by their mutual gravitation. These binary systems behave like a sun-planet system in that they travel through space together; but since the difference between the masses of two stars is many times less than that between a star and a planet, each will revolve about the other—or, more accurately, each will revolve about the centre of gravity of the system. From observations of a binary, coupled with the knowledge crystallized in the laws of Newton (in accordance with which their motions are performed), it is possible to calculate its distance. Parallaxes derived in this way are known as dynamical or hypothetical parallaxes.

Since the mathematics involved is quite simple, and the process is

difficult to fathom when described verbally, we shall run through the train of reasoning which starts from the angular separation of the members of a binary system and their period of mutual revolution, and ends with their distance from the observer.

It will be remembered that Kepler's harmonic law was shown by Newton to be strictly accurate only if the planets had no mass whatever; Newton therefore adjusted the equation to bring in the relative masses of the two bodies, planet and sun. From this modified equation it follows that for a pair of stars in mutual revolution

$$\frac{m_1+m_2}{S+E}=\frac{A^3}{P^2}$$

where m_1, m_2 are the masses of the two stars,

A, their distance apart (in astronomical units),

P, their revolution period (in years),

S, the mass of the sun,

E, that of the earth.

If A is the stars' linear separation in astronomical units, a their separation as observed (measured in seconds of arc), and p their parallax (also measured in seconds of arc), then

$$A=\frac{a}{p}.$$

We have learnt that the mass of the earth is negligible compared with that of the sun; it may therefore be omitted without materially affecting the equation. Making this omission, substituting the value of A in the above equation, and expressing m_1 and m_2 in terms of the sun's mass, we have

$$m_1+m_2=\frac{a^3}{p^3\times P^2},$$

whence, bringing p to one side,

$$p=\frac{a}{\sqrt[3]{P^2(m_1+m_2)}}.$$

Making the assumption that each star is as massive as the sun,[1] we can substitute $2S$, a known quantity, for (m_1+m_2):

$$p=\frac{a}{\sqrt[3]{P^2\times 2S}},$$

and solve for p.

[1] It is because this assumption has to be made, the true masses of the two stars being unknown, that dynamical parallax is sometimes known as hypothetical parallax. The assumption is well founded, however, for we shall see in Chapter VIII that the stars vary less among themselves as regards mass than in any other respect; the sun, moreover, is a quite 'average' star.

By the mathematical 'trick' of successive approximations—use being made of (i) a certain relationship between stellar mass and luminosity which we have not yet discussed, and (ii), the known relation between apparent and absolute magnitude, and distance—this method is capable of giving results with only a small percentage inaccuracy. Though not of wide application, it is useful in that it permits the determination of the distances of binaries which are too remote for trigonometrical investigation. The nearer binaries (those within 30 light years or so) may be dealt with as accurately by the direct method.

Group parallax

To round off this account of the means whereby man has conducted his intellectual colonization of interstellar space, we must at least mention one further method of distance determination: group parallax. Since it is of only secondary importance, and restricted in scope, it may be dealt with more summarily than those hitherto described.

Most of the readers of this book will be familiar with the group of stars, visible on winter nights, known variously as the Pleiades and the Seven Sisters. They mark Taurus, a constellation many of whose stars have been shown by a study of their proper motions to be moving in the same direction at the same speed. Further, their paths if produced forward all converge upon a single point on the star sphere. This is to be expected, since parallel lines (and the stars are moving parallel to one another) appear to meet at infinity. Now if the earth were a member of the group, it would also be moving towards this convergent; therefore the terrestrial observer's line of sight to the convergent is parallel to the stars' space motions. If, in addition to its proper motion, the radial velocity of a star in the group is known—and the spectroscope quickly provides this datum—it is possible to construct a formula linking

the star's parallax,
its space motion,
its proper motion,
its radial motion,
the angle between the observer's lines of sight to the star and to the convergent respectively.

It is to be noted that although this method gives the parallaxes of individual stars, it can only be applied to members of a group whose proper motions are large enough to be perceptible. Though

not, therefore, of extensive application, group parallax is a neat and successful method of determining certain stellar distances.

Recapitulation

To sum up:

Trigonometrical parallax,
Statistical (mean) parallax,
The period-luminosity relationship of the Cepheids,
The absolute magnitudes of novae at maxima,
Dynamical parallax,
Group parallax—

these are the more important weapons with which the modern astronomer has armed himself for his assault upon the secret of the structure of the visible universe.[1] We must now turn to the results and rewards of this assault.

[1] A seventh method, known as spectroscopic parallax, is dealt with in Chapter VIII, where the physical conceptions involved in its formulation are discussed. Though it has played an important role in the past, it is to-day tending to fade out of the picture in favour of trigonometrical and statistical parallax.

V

TO THE END OF KNOWLEDGE

WE are now in a position to inquire what information regarding the organization and extent of the visible universe is provided by the criteria of distance with which we have so far largely concerned ourselves, as well as by such additional methods of specialized application (developed from those already discussed) as we may from time to time devise.

The nature of the problem

During the sixteenth century the belief that the stars were minute points of light equidistant from the earth was shown to be false, but it was not until the end of the eighteenth century that any serious observational attempt was made to replace the outworn cosmology of a bygone age by one in closer agreement with the known facts. The problem to be solved might be stated as follows: Are the stars distributed more or less uniformly throughout the whole of space, or do they form a discrete system which occupies, possibly, only a small corner of space? If the latter, then what are the dimensions, shape and structure of this system? And, it might be added, is the system unique, or has it peers in other regions of outer space? *A priori* theorizing could give no answers to questions such as these, and the only procedure whereby the problem could be attacked with any prospect of success consisted in a careful study of the apparent distribution of the stars upon the face of the star sphere, followed by an attempt to correlate this with some sort of spatial distribution.

The Milky Way

The most noticeable feature of stellar distribution in the moonless night sky is the Milky Way. This is a faint band of light, varying in width from 45° to less than 5°, which traverses the visible half of the star sphere from horizon to horizon. A trip round the world proves that the Milky Way girdles the entire star sphere, forming a continuous, or nearly continuous, belt about the heavens. Furthermore, it is very nearly a great circle; that is to say, it divides the star sphere into two almost equal sections. One of the earliest discoveries of the telescope was that the Milky Way, or galaxy, consists of vast hordes of stars too faint and too closely crowded to be individually distinguishable by the naked eye, although their integrated light is

visible as a misty band. More detailed study shows that the richness of the Milky Way varies along its length: one area may be almost starless, while nearby the stars are so congested, so piled upon one another, that they present a 'solid' wall of light.

It is also noticeable with the naked eye, and still more so in long-exposure photographs, that the Milky Way is materially richer in one direction—roughly in the direction of the constellation Sagittarius—than in the opposite direction. Finally, although there is a colossal piling up, or concentration, of the fainter stars in the galactic regions, the brighter stars show a less marked crowding.

Structure of the Milky Way

The stellar concentration towards the galactic plane, with progressive avoidance of the regions approximating to the galactic poles, can be explained in one of three ways. We may suppose, in the first place, that the sun is surrounded by a ring of stars within which the star density is higher than elsewhere. Since this ring, the Milky Way, apparently bisects the star sphere, the sun must lie near its median plane. Since there is no reason why the ring should not be very distant, this theory can account for the observed fact that the brighter stars (in general, the nearer ones) are less restricted to the galactic regions than the fainter. According to this explanation, the apparent concentration of stars within the boundaries of the Milky Way is a reflection of a real spatial crowding: the average distance separating two adjacent stars within the ring is less than that separating adjacent stars outside it. This conception of a ring is, however, artificial to a degree, and bears no relation to any configuration of stars known to exist elsewhere within the boundaries of the observable universe.

The appearance of the Milky Way can also be reproduced without recourse to any actual concentration of stars within a certain spatial zone: the laws of perspective alone could achieve it. Suppose that the stars form a disc-shaped system similar to a biconvex lens, within all regions of which the star density is uniform. Suppose, further, that the sun lies in the median plane of this system but at some distance from its centre. Then a terrestrial observer will, in gazing out into space, be looking through a greater thickness of stars when he is facing the periphery of the disc than when he is looking towards either of its poles. In the former direction, therefore, the stars will appear to be more closely crowded together than in the latter, and the observed effect will be that of a Milky Way separating two relatively starless regions. The eccentric position of the sun within the galactic plane will, furthermore, produce the effect of

greater star density in one direction along the plane (towards its centre) than in the opposite direction, where the edge of the system is nearer.

A third possibility is a combination of the two just considered; the stars do not indeed permeate all space, but are concentrated into a roughly lenticular system, in whose median plane the star density is higher than on either side of it. This was the conclusion at which the elder Herschel arrived in the early years of the nineteenth century. He was led to this result by the simple observational method of star-counts: he divided the entire celestial sphere into a large number of equal adjoining areas, and then, using the most powerful telescope at his disposal, proceeded to count the number of stars visible in each area. Only one assumption was made—that his telescope was capable of penetrating to the confines of the star system in all directions. Were it incapable of achieving this, the derived shape would inevitably be to a greater or lesser extent misleading (as shown in

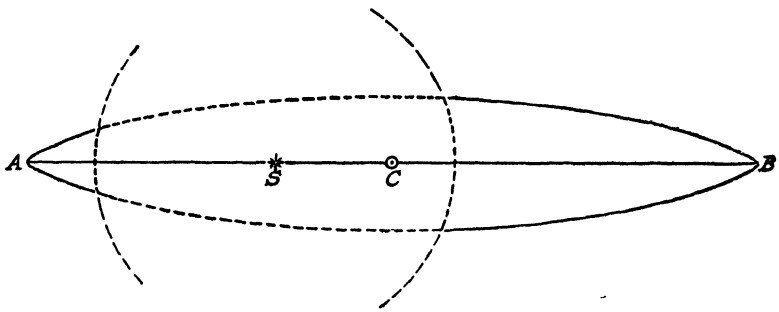


Figure 33. Star counts falsify the shape of the galaxy if the telescope is incapable of reaching to its confines. Furthermore, under these circumstances, the sun must appear to be centrally placed within the galaxy.

Fig. 33), since he regarded it as axiomatic that the distance to the edge of the system in any direction was proportional to the number of stars per unit area in that direction. At the outset of his surveys he made the further, and not entirely unconnected, assumption that the system was homogeneous; later, he was forced to conclude that this assumption was not strictly justified, and that, as is now known to be the case, the star density is not uniform throughout the system. Nevertheless, the results of his survey convinced him that the observed star crowding in the Milky Way was very largely optical.

Distribution of the brighter stars

Herschel's star counts showed clearly that there is an immense piling up of the fainter stars towards the median plane of the Milky

Way—known shortly as the galactic plane, or galactic equator—but that, as is proved by a glance at the night sky with the naked eye, the brighter stars are not similarly restricted. Indeed, what slight degree of crowding there is among the brighter stars is not about the galactic plane but about a plane inclined at some 15° or 20° to it. This second plane—known as Gould's Belt of Bright Stars, after the astronomer who first drew attention to it—is most clearly seen in winter, when the bright stars of Canis Major, Orion and Taurus, all components of the belt, are above the horizon during the night time and are conspicuously aslant the Milky Way itself. For many years this freedom of the brighter stars from galactic concentration was held to constitute a conclusive argument against the validity of Herschel's disc theory, for if the appearance of the stellar universe is due entirely to its shape and to the position of the sun within it, then it is to be expected that the brighter stars as well as the fainter would be subjected to the laws of perspective and would therefore exhibit galactic concentration.

The Local Cluster

More recent work has resolved this dilemma and has at the same time vindicated Herschel in the main outline of his universe-picture, though his assumption of approximately uniform stellar distribution is very far wide of the mark. Telescopic examination of the Milky Way shows that its structure is far from homogeneous (see Fig. 34). It consists of vast star clouds, in which the individual stars appear to be so closely crowded as to merge into an undifferentiated haze of light, interspersed with areas of lower star density or even of apparent absence of stars altogether. Also there are in galactic regions many small clusters of stars, more localized and humble editions of the vast star clouds. It has been established as the result of work on star motions that the sun itself is a member of one such cluster, or minor star cloud, all of whose other members are—compared with the faint stars of the Milky Way—close neighbours of the sun. For this reason they are to be seen in all directions about the sun, and, since the terrestrial observer is inside the cluster, are exempt from galactic concentration. In point of fact, the median plane of the Local Cluster—for, like the galaxy, it is a flattened, lens-shaped structure—is inclined at some 12° to that of the galaxy, and is closely coincident with Gould's Belt of Bright Stars, all of which are members. In general, it is true to say that all the brighter stars of the constellation figures are members of our Local Cluster. Shapley and Charlier have shown by their studies of the distribution of the brighter

blue-tinted[1] stars that these also characterize the Local Cluster and consequently show a tendency to crowd about a plane inclined at some 12° to that of the galaxy. This is true of blue stars down to about the sixth magnitude, it being statistically true to say of the brighter stars of any type that they are nearer than the fainter stars. Blue stars fainter than magnitude 7·5 congregate about the galactic plane in the normal way of faint stars, while those of intermediate brightness show a preference for planes intermediate between those of the galaxy and the Local Cluster.

Using criteria of distance already described, Charlier has discovered that the Local Cluster is a flattened spheroid with a diameter of some 2,000 light years, and a thickness, measured perpendicularly to the median plane, of perhaps 600 or 700 light years. The sun is located slightly to one side of the median plane at a distance of about 300 light years from its centre. Charlier's cluster may be the nucleus of a larger conglomeration which measures, according to Seares, some 20,000 by 7,000 light years.

Methods of plumbing the galaxy

The discovery of the Local Cluster—a minute galaxy within the Galaxy—was made possible by the study of apparent stellar distribution, and of stellar motions and distances. Owing to its comparatively small dimensions it was practicable to use distance determinations based upon individual stars, as already described. The rough outlines of the galaxy as a coherent, isolated, roughly lenticular system were traced by Herschel purely from studies of star numbers in different directions. When the problem of scale-mapping the star system arises—discovering the distances to its edges in different directions—the solution is less easy than in the case of the Local Cluster, owing to the very much greater distances involved. Even the brightest stars are invisible at distances of a galactic order (mainly owing to the existence of amorphous absorbing material within the galaxy) and although Cepheids and novae have provided a partial solution, single stars are in the final analysis of little value except for approximate statistical purposes. To probe the galaxy to

[1] Stars are classified into a number of types, each possessing distinctive spectroscopic features. These types are more fully described in Chapter VIII, after the functions of the spectroscope have been systematically set forth. Here it will be sufficient for the reader to note that the correspondence between the colours and spectral types of different stars is roughly as follows:

Type	*Type*
B: blue stars	G: yellowish-white stars
A: blue-white stars	K: yellow stars
F: white stars	M: red stars.

Figure 34. Star field in the Perseus and Taurus region. Galactic nebulosity can be seen at the top of the photograph. The streak towards the lower right-hand corner is the trail of a meteor in the earth's atmosphere. (*By courtesy of the Director of Yerkes Observatory.*)

its limits we must fall back upon the clusters and groups of stars, first evolving distance criteria for these objects.

Evidence of single stars

Before proceeding to a description of the clusters, their appearance, their classification and the distance criteria that have been elaborated for them, let us first see what light has been thrown upon the scale of the galaxy by such sounding lines as Cepheids, novae and other individual stars of great intrinsic luminosity.

Cepheids in the galactic plane have been detected in all directions to distances of about 20,000 light years, their average distance on either side of the plane being less than 500 light years. At such distances they are extremely faint and difficult of detection, and it is not to be supposed that none fainter and more distant exist. This is borne out by the evidence of other types of variable whose distance can be approximately determined from an only narrowly varying mean absolute magnitude. Such variables have extended the minimum distance to the edge of the galaxy in the direction of Sagittarius to over 30,000 light years.

This is the limit to which we are led by single stars. That even this is not the end of things is clearly indicated by figures, the results of detailed star counts, which show the numbers of stars visible at each successive magnitude level. For when star numbers are arranged in this way it is found that in the regions of the Milky Way—but not of the galactic poles—their numbers are still increasing rapidly at magnitude 21, which is about the limit of photographic detection. A star similar to the sun would have to be removed to a distance of 50,000 light years before sinking to this threshold; so far, then, the evidence indicates that within the galactic plane and at least in one direction stars are to be encountered for more than 50,000 light years.

Eccentric position of the sun

This question of direction is of some interest: even if the sun is centrally placed within the galaxy (we have seen that it must be near the central plane, i.e. near the line *AB* in Fig. 33, since the Milky Way is approximately a great circle) the edge of the system will lie at different distances in different directions. This depends upon the fact that the galaxy is not spherical but lenticular. But if, in addition, the sun is eccentrically placed within the median plane, the distance to the edge will vary with different directions even within this plane. It has been found that the most distant variables yet detected are confined to one hemisphere of the Milky Way, and congregate particularly in the constellation Sagittarius, where the star clouds are

also densest. Furthermore, the faintest novae show an identical distribution. Finally there is the evidence of the globular clusters; these will be described later in the present chapter, and it will suffice to say at this point that they likewise appear to be centred upon the Sagittarius region. We may therefore assume, at any rate until contrary evidence appears, that the galaxy is deeper in this than in the opposite direction, since the faintest objects characterize it; in other words, the sun is not as Herschel thought, situated at, or even very near, the centre of the galaxy. Fig. 33 shows how this misconception is liable to arise if the telescope employed for the star counts is incapable of reaching the edge of the galaxy in all directions. That Herschel's great reflector was able to plumb the galaxy in the direction of its poles is demonstrated by the fact that more powerful instruments show very few more stars per unit area in this direction than did his. This, however, is not true of those directions which lie in the galactic plane.

Preliminary results

These preliminary results, together with considerations based upon Shapley's studies of the globular clusters which we shall consider later, suggested 200,000 light years as a very provisional diameter for the galaxy, its thickness being perhaps one-sixth as great. It is true that this early figure has been whittled down by Trumpler, van de Kamp, Stebbins and others by something like one half, but even with these revised figures the galaxy remains a structure built to a scale immeasurably greater than that of the solar system or even of the Local Cluster. A ray of light takes 8 minutes to reach the earth from the sun; five hours later it crosses the orbit of Pluto, while four and a half years are required for it to reach the nearest star. Yet a ray of light leaving one edge of the stellar system would not reach the diametrically opposite edge for something like a thousand centuries.

Moving clusters

Stars show a marked proclivity for going about in groups; one such group, the Local Cluster, we have already mentioned. Since the sun is an internal member of the Local Cluster, and is therefore surrounded on all sides by other members, this grouping does not appear as a cluster to the terrestrial observer. In fact, as we have seen, its existence avoided detection until comparatively recent times, when astronomical methods of probing behind the mask of appearances had reached a considerable degree of finesse. This is also true of star groups which, though not actually containing the

sun, are nevertheless comparatively near to it in space. The smaller the distance of two objects from the observer, the wider is their angular separation. Hence the members of a comparatively proximate grouping appear to be widely dispersed upon the star sphere. They do not in fact look like a star cluster, and their coherent nature can only be demonstrated by proving a community of motion among them. If a number of stars—no matter how great their apparent separation—are found to be travelling in the same direction with the same velocity, the only alternative to supposing them to constitute a single, gravitationally coherent group is to appeal to coincidence, assuming them to be in reality so remote from one another that they cannot be affecting one another gravitationally to any appreciable extent. The greater the number of stars that are found to share a particular motion, the more fantastic becomes the coincidence. Though the study of proper and radial motions is necessary before the true nature of a cluster of this type can be proved, it can nevertheless sometimes be guessed at. The stars of the Plough, for instance, look as though they might form a single system, and the same may be said of the Orion stars; but before this can be shown beyond doubt to be the case, a very careful investigation of their motions must be carried out.

In the case of the nearest groupings, however, it may be quite impossible even to guess that the individual stars are gravitationally connected. The stars Sirius, β Aurigae, δ Leonis, α Coronae, and, among others, five of the stars of Ursa Major form such a cluster. Since their true nature can only be discovered through a study of the motions of their components, these clusters are known as moving clusters. Though they are described as 'comparatively near' the sun, it must be remembered that the most distant stars in the galaxy are perhaps 100,000 light years distant, so that a cluster 130 light years from the sun (as is the Taurus moving cluster of eighty known members), though inconceivably remote by terrestrial standards, is yet a next-door neighbour compared with the whole galaxy.

Open clusters

Clusters that are more distant than these moving clusters appear to be more highly concentrated; the angular separations of their components are smaller, and their cluster-nature can be perceived visually. Clusters of this category are known as open clusters, and the Pleiades is a prominent example. It is to be noted that a closer crowding of the stars within the confines of the cluster would give the same appearance as greater distance, but it is probable that the majority of the open clusters appear to be more condensed than

moving clusters because they are more distant and not because of greater spatial concentration. A moving cluster at 100 light years, in fact, would be an open cluster at 10,000. The angularly largest of the open clusters are visible to the naked eye, and in some cases, though not all, the individual stars may be seen without instrumental aid. An example of an open cluster whose individual stars are invisible to the naked eye, though its integrated light allows it to be seen as a faint misty spot, is the great double cluster in Perseus; the telescopic appearance of this object is illustrated in Fig. 35. The more distant clusters, however, are not visible to the naked eye. Telescopically they are seen to be clouds of faint stars, whose arrangement is characteristically random, the cluster as a whole having no obvious structural symmetry, although this generalization is not universally valid.

Apparent distribution of the open clusters

About 400 of these clusters are known. Their distribution upon the star sphere is characteristic and has given them their most usual name of galactic clusters, for they share with the fainter stars a strong galactic concentration. Their distribution in galactic longitude is more or less irregular and in general follows that of the fainter stars: the galactic clusters, in other words, while being confined almost exclusively to the region of the Milky Way, occur haphazardly along its length.

Distances of the open clusters

The problem of their distances and spatial distribution was attacked systematically by Trumpler some fifteen years ago, and his researches uncovered facts whose significance extends far beyond the realm of the open clusters. The first distances of the nearer clusters were discovered by determining the spectroscopic parallaxes of the involved stars. Trumpler's technique was to single out the twenty or thirty brightest stars in a cluster and to determine their spectroscopic type. As we shall see in Chapter VIII, stars of each spectroscopic type are closely similar to one another in a number of respects—mass, temperature, size, luminosity and so on. Thus a fairly precise correlation is to be established between a star's spectroscopic type and, among other things, its luminosity; hence from the type of each of Trumpler's forty-odd bright stars in each cluster its intrinsic luminosity or absolute magnitude could quickly be derived. Since, in completely transparent space, apparent brightness decreases with the square of the distance, the distances of each of the selected stars could now be calculated in a few seconds. And since each star is,

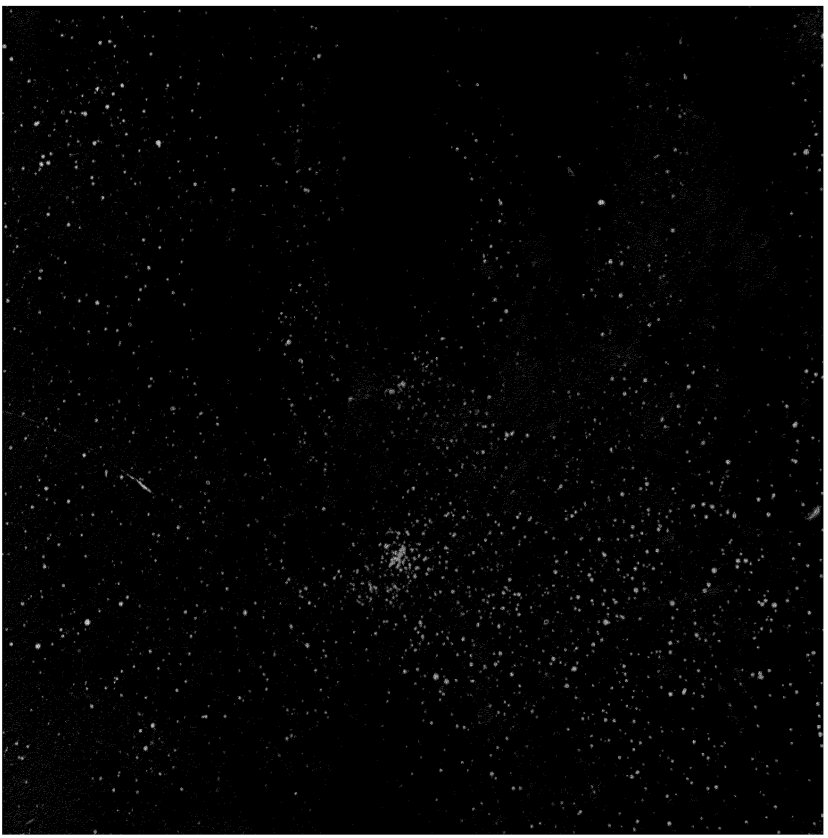


Figure 35. Galactic clusters: the double cluster in the constellation Perseus. (*Crown copyright: From an exhibit in the Science Museum, South Kensington, London.*)

Figure 36. The globular cluster M. 13. (*By courtesy of the Director of the Dominion Astrophysical Observatory.*)

within the limit of the small inaccuracies inherent in the method, situated at the same distance from the observer, it is only necessary to average the more consonant results, omitting any that are widely discrepant (probably background or foreground stars unconnected with the cluster) in order to derive the cluster's distance with a maximum probable error of not more than ± 10 per cent.

The furthest clusters however are so remote as to preclude the accurate determination of the spectroscopic types of even the brightest components, and in order to bring these particularly important clusters within his scheme Trumpler had to evolve a further distance criterion based upon observations of those clusters whose distance he had already determined. He found that the galactic clusters, although at first glance their linear dimensions varied within wide limits, were nevertheless susceptible of classification according to their structure—degree of central condensation, numbers of stars comprising the cluster, degree of symmetry, etc.—and that members of each group of this classification were much more nearly of equal linear dimensions. Thus for the most distant clusters, to which the spectroscopic method is inapplicable, it was necessary to determine by close examination to which group of the classification they belonged, to measure their angular diameter, and then to compare this with the linear diameter already derived for the group by a study of the nearer clusters whose distances had been determined by the independent spectroscopic method.

Evidence of light absorption in space

The preliminary survey of one hundred clusters yielded distances with a probable maximum inaccuracy of 10 per cent. to 20 per cent. for about three-quarters of them. From this survey an extremely interesting fact emerged, a fact whose significance has necessitated the reviewing of all previous work on the size of the galaxy and the distances of individual components of it. Knowledge of the distance and angular diameter of each cluster gave its linear diameter, and hence the average linear diameter of all the clusters included in the survey could be calculated. When this was done, and the clusters arranged in order of distance, the further fact came to light that the nearer clusters were systematically smaller than the average, while the most distant were systematically larger. It appeared that, on the scale of distances derived, the more distant a cluster is from the sun the larger it must be—an altogether fantastic proposition. Obviously some systematic error had crept into the distance estimates. After considering various possibilities, Trumpler came to the following conclusion. Interstellar space is not entirely transparent, and therefore

apparent brightness decreases more rapidly than with the square of the distance; therefore all the distance determinations were too large, the error being inappreciable at small distances, but growing to measurable proportions at greater distances—in short, the greater the distance the greater the error. It was found that the degree of light absorption in space that it was required to hypothecate, in order to remove the discrepancy between distances and diameters, amounted to about 0·8 magnitudes per 3,250 light years. In other words, a cluster which is situated at this distance from the observer appears to be 0·8 magnitudes fainter than it would if space were perfectly transparent. When this correction is neglected the derived distance will be larger than the true distance and the cluster will therefore appear to be larger than in fact it is.

Spatial distribution of the open clusters

The survey was then extended to include all of the 350-odd known open clusters; the light absorption correction was applied to the distances, and the results were correct within a probable error of 10 per cent. to 12 per cent. The following facts emerged. The galactic clusters form a flattened system similar in shape to the supposed shape of the galaxy itself, but exhibiting a degree of concentration upon their median plane even more marked than that of the fainter stars upon the galactic plane; two-thirds of all the open clusters are situated within 325 light years of the median plane of the cluster system. This plane of symmetry is nearly, though not quite, coincident with the galactic plane. The diameter of the system is something like 40,000 light years, its maximum thickness something over 3,000 light years, although the outer regions are very sparsely populated. Within it, there is a strong concentration of clusters, not only upon the median plane, but also towards the centre of this plane. The sun is situated at something over 1,000 light years from the centre of the galactic cluster system as at present defined.

The bearing of these results upon the structure and dimensions of the Milky Way itself—the star system in which the clusters are embedded, like currants in a cake—must be shelved until we have learned something of the next type of cluster, the globular cluster.

Globular clusters

The globular clusters, though having some obvious points of resemblance with the open clusters, differ from them in appearance in much the same way that the open clusters differ from the moving clusters. They resemble the more distant open clusters in their faintness and small angular diameter, but differ from all open

clusters in their extremely high central condensation and in the definiteness of their shape and structure. Fig. 36 shows a typical globular cluster and a comparison of it with Fig. 35 will immediately reveal both the resemblances and differences between the two classes of object. All globular clusters are spherical—some are very slightly oblate—whereas the open clusters have no definite shape, or, rather, a variety of more of less asymmetrical structures. The globular clusters are strongly condensed centrally—the average distance separating two adjacent stars is smaller in the centre of the cluster than in the outskirts—while a trace of central condensation is shown by only a minority of open clusters. Furthermore, the star density does not fall off steadily from the centre of a globular cluster towards its periphery, but rapidly at first and then more slowly.

These clusters are faint objects, only a few of the hundred-odd that are known being visible to the naked eye: we shall see shortly that this feature is dependent upon their great distances, for they are the most remote class of object in the galaxy. An interesting characteristic is that they are all of roughly the same size; their average diameter is about 100 light years, while that of the condensed central portion is normally little more than one-tenth of this figure. Shapley, who is responsible for the greater part of our knowledge of the globular clusters, established this fact in the following manner. Most of the clusters contain variables in great numbers, and a very careful study of these has shown that their light curves are of the typical Cepheid variety. By their means Shapley was able to determine the distances of a number of clusters; once this was done their linear size could of course be deduced from their angular size.

Apparent distribution of the globular clusters

We shall return to this point in a moment, but first something must be said of the distribution of the globular clusters about the star sphere. This distribution has two outstanding features.

In the first place, the clusters are confined almost exclusively to one hemisphere of the heavens. The pole of this hemisphere, which contains over 80 clusters, lies in the Milky Way in the Sagittarius region; in the other hemisphere there are only four. The only conclusion to be drawn from this, if the clusters are scattered through space with any degree of uniformity, is that the sun is eccentrically placed both with regard to the system of the stars and to that of the globular clusters. Fig. 37 shows how such a location of the sun within the system would result in the asymmetrical distribution of the clusters in the sky.

Secondly, although clusters are to be found on both sides of the

Milky Way, they are wholly absent from the mid-galactic regions: their distribution in galactic latitude shows a concentration in two zones lying approximately alongside the Milky Way, outside which

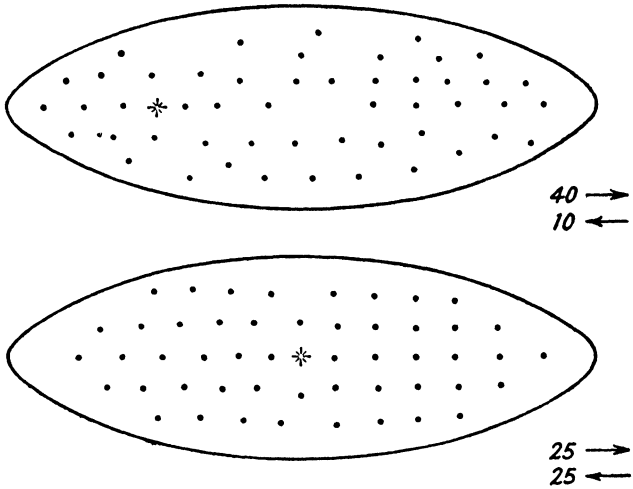


Figure 37. That more globular clusters are visible in one direction than in the opposite direction indicates an eccentric position of the sun within the system.

they occur with diminishing frequency and between which they are almost completely absent.

Distances of the globular clusters

Shapley's main objective was to discover the shape and size of this system of globular clusters, and the first step towards accomplishing it was the accurate determination of as many of their distances as possible. We have already encountered three methods that he had at his disposal. Cepheids; the fact that the clusters are all of about the same size, which size can be determined initially through Cepheid observations; and thirdly, the absolute magnitudes of the brightest stars in each cluster, for he found that the average absolute magnitude of the apparently brightest stars in a number of different clusters was much the same. It seems reasonable to assume that if this law held good for those clusters whose distances had already been determined, it would also apply to those clusters whose distances were yet unknown. For, the mean absolute magnitude of, say, the dozen brightest stars being known, it only required the sufficiently accurate determination of their mean apparent magnitude for their distance to be calculated.

Before we proceed to an account of the results of this investigation, it would be advisable to glance a little more closely at the various

assumptions inherent in the three distance criteria employed by Shapley, for upon their validity depends the validity of the results. The foundation of his series of criteria is the Cepheid method, for the subsequent discoveries of the uniformity of the size of the clusters, and of the luminosity of the brightest stars, were the result of applying this method to those clusters in which Cepheids were visible. The period-luminosity relation of these stars has been confirmed time and time again, and results based on it may be considered absolutely reliable within the small margin introduced by inexact fixing of the zero point. But the variables in the globular clusters differ from true Cepheids in one important respect, and it has been suggested that this difference vitiates the results derived by applying the period-luminosity law to them. Though the shape of their light curves is that of true Cepheids their periods are in every case very much shorter—less than one day. Despite this divergence of the cluster variables from typical galactic Cepheids, Shapley considered that the striking similarities between the two justified his assumption that the period-luminosity law held for both.

His use of the cluster variables has been attacked from a somewhat different direction by Kapteyn and van Rhijn. Cepheids are known to be giant stars. Now these workers showed that cluster variables have large proper motions, and from this they argued that they must be considerably nearer the sun than Shapley supposed. Hence, since their apparent brightness is that of a giant at Shapley's distance, they and the clusters in which they occur must be both fainter and nearer than Shapley supposed. Although it is now known that this is indeed the case, the particular argument under consideration is not to the point. Shapley retorted quite correctly that large proper motions are characteristic of variables in general, and Lindblad appears finally to have clinched the matter by showing that the spectra of at least some of the cluster variables are those of giants.

The assumptions involved by the other two criteria have already been indicated: that because some of the clusters are about the same size, they all are; and that because the mean absolute magnitude of the brightest stars in some of the clusters are the same, this uniformity applies to all clusters. Yet not only are these assumptions prima facie reasonable, but they also appear to be justified by the close concordance between the results that are yielded by different methods. The odds against this being so in a large number of cases, if one or all of the assumptions are invalid, are so great as to be unacceptable.

One further assumption that Shapley unhesitatingly made, since at that time there was no known reason why it should not be, was

that space is absolutely transparent and that therefore the brightness of a source falls off in proportion with the square of its distance. The implications of this inverse square law possibly being inoperative were not considered at the time that the distances mentioned in the next paragraph were being derived.

The nearest cluster is ω Centauri, which is also the largest (angularly) and the brightest; its distance from the sun was derived by Shapley as 22,000 light years. That of the furthest, N.G.C. 7006, is in the neighbourhood of 200,000 light years, nearly ten times as great. It will be recalled that the most distant galactic clusters, which are objects showing no preference for one hemisphere of the star vault, are distant about 20,000 light years from the sun, or just about the same distance as the nearest globular clusters.

Spatial distribution of the globular clusters

If we accept the inevitable explanation of the one-sided distribution of the globular clusters—namely, that the sun is eccentrically placed within the system—we have immediately to hand a logical explanation of the different distributions of the open and the globular clusters. For the open clusters—all of which are nearer than the nearest globular cluster—are too close to the sun for any 'piling up' effect in the direction of the galactic centre to be noticed, whereas in the case of the more distant globular clusters this is not so. Fig. 38 may help to clarify this distinction.

The second feature of the distribution of the globular clusters—their avoidance of a zone stretching for about 10° on either side of the galactic plane—at once reminds us of Trumpler's discovery that the light is absorbed by, presumably, diffuse dusty or gaseous material permeating the central plane of the Milky Way. His experience of the distribution of the open clusters led him to believe that this material constitutes a thin sheet, probably of varying density and light-absorbing powers, probably not more than 600 to 1,000 light years thick, and extending centrally through the system of galactic clusters—that is, with a minimum diameter of something like 40,000 light years. This stratum of obscuring matter would effectively mask any globular clusters lying in its plane at still greater distances, and result in the observed bilateral distribution of these distant objects.

Evidence is not lacking that this account is in the main true. According to Trumpler, the heaviest absorption would only occur within 10° of the galactic plane, the very region from which the remote globular clusters are missing. The fact that the vast majority of visible globular clusters lie outside this region explains why they are

free from the colour excesses which characterize the remoter galactic clusters within it.

Mechanism of light absorption

At this point a few words on the mechanism of light absorption must be interpolated in order to make the meaning of the last sentence plain. Material particles in the form of a rarefied cloud affect light passing through it in one of two ways, according as to whether the particles are larger or smaller than a certain critical value. This threshold is about the wavelength of light:[1] if the diameter of the particles is larger than this value, they will scatter light of all wavelengths to an equal extent, so that the source will appear to be dimmer than in fact it is; if, on the other hand, the diameters of the particles are on the whole smaller than the wavelength of light, the absorption will be selective, the degree of scattering varying inversely with λ^4 where λ is the wavelength of the incident light. Thus, blue light will be scattered more than red—just as the sun's light is scattered by the molecules of the terrestrial atmosphere, giving us a blue sky, and a red sun at sunrise and sunset—and the light from the source will appear to be redder than in fact it is. Absorption by small particles, as against absorption by large particles, changes the perceived colour of light from a source beyond them.

Now the colour of a star is expressed in the form of the colour index, or c/i. Photographic plates and the human eye are sensitive to slightly different wavelength ranges, the eye being comparatively more sensitive to red light and less sensitive to blue; hence the apparent brightness of a star as shown by a photographic plate may be greater than, less than, or the same as recorded by the eye; a red star photographs fainter, compared with a white star, than it is visually. The difference between the photographic and visual magnitudes of a star is thus solely dependent upon its colour—it is, in other words, an index of the star's colour. If we write m_p and m_v for photographic and visual magnitudes respectively, we therefore have

$$c/i = m_p - m_v$$

Since stars of different spectroscopic types are of different colours, the c/i of each type is distinctive and has been established accurately by means of nearby stars which are materially unaffected by interstellar absorption. If the light of a distant star is subjected to

[1] Anticipating the systematic description of the nature and behaviour of light given in the next chapter, it will suffice to state the following facts in elucidation of the present account: (1) light behaves as though it were a system of waves, (2) the distance separating the crests of adjacent waves is different for light of different colours, (3) the wavelength of red light is longer than that of blue.

selective absorption in transit to the terrestrial observer, its c/i will be altered, and this difference between the observed c/i of a star and the normal c/i for stars of its spectroscopic type is called its colour excess. If, then, the absorption detected by Trumpler is of this type, the stars of the more remote open clusters should have noticeable colour excesses, and the excess should be a direct function of distance. This has actually been found to be the case, adding one further buttress to the edifice raised by Trumpler.

Light absorption and the system of globular clusters

It was noted on p. 117 that the majority of globular clusters are free from colour excesses, although more distant than the open clusters, for the reason that light passing to the terrestrial observer from an object outside the 20°-wide band centred on the galactic equator would travel through only a negligible thickness of the absorbing medium. This interpretation is borne out by the fact that the few globular clusters that have been detected within 10° of the galactic plane do show the effects of the same absorption to which similarly situated open clusters are subjected. For these globular clusters are measurably fainter than their distance, as determined by their diameters, would warrant.

Shapley's estimates of the distances of the clusters lying far from the galactic plane are therefore not largely affected by the new light that has been thrown on the question of interstellar absorption. But whereas his globular cluster system was originally thought to be a flattened structure, drawn out to a diameter of fully 300,000 light years in the direction of the edges of the 'zone of avoidance', it now appears that the distances to clusters in this region of the system have to be pulled in to counteract the effects of absorption upon the original distance measurements. The result is that the globular cluster system—unlike both the galactic and the open cluster systems —is more or less spherical, with a diameter of some 80,000 to 100,000 light years. Fig. 38 gives some idea of this; both the scale and the outlines are necessarily imprecise at the present stage of our knowledge, but are likely to be fundamentally correct.

The globular clusters and the stellar system

The point at issue is whether the known galactic clusters are coextensive with the galaxy, or whether the latter is, not coextensive, but concentric with the system of globular clusters. That the globular clusters and the galactic system are not coextensive is certain, for whatever the extent of the galaxy in its median plane beyond the star fields of Sagittarius, it is certainly (as Herschel

proved nearly 150 years ago) a highly flattened structure. The further globular clusters in high galactic latitudes are without a doubt extragalactic objects—satellites of the galaxy, if we like, or flies hovering above a cake whose embedded currants are the galactic clusters.

But though the globular clusters are not coextensive with the star system, the latter will be of very different dimensions if it is (i) exclusively coextensive with the system of known open clusters or (ii) concentric with the globular cluster system. If the latter, the centre of the galaxy lies at a distance of perhaps 30,000 light years in the

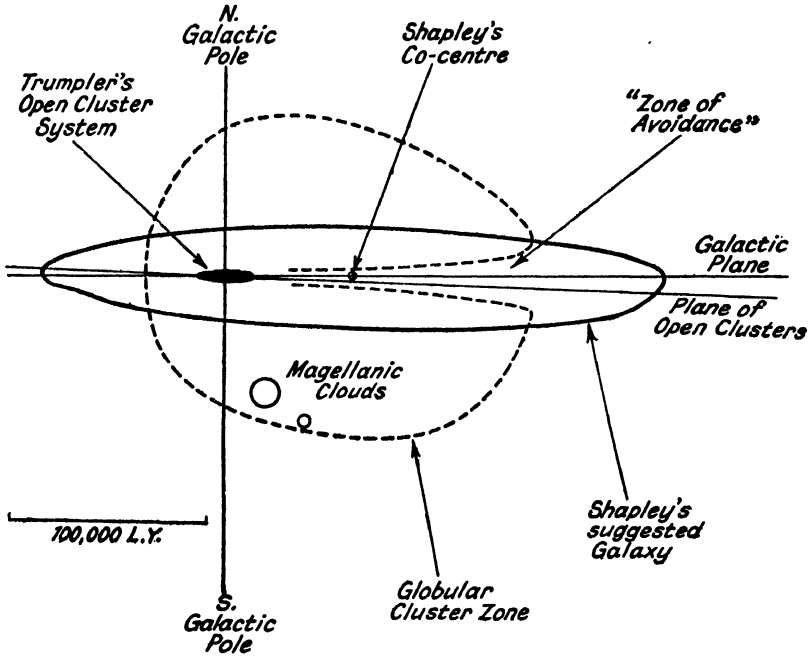


Figure 38.

direction of Sagittarius and will be of the order of 100,000 light years in diameter (see Fig. 38).[1] If, on the other hand (i) represents the true picture, then the sun is much more centrally placed in the star system, whose diameter is at the same time reduced to something like 40,000 or 50,000 light years. These opposing schools of thought are championed, respectively, by Seares and Shapley on the one hand, and by Trumpler on the other. The pros and cons may conveniently be tabulated as follows, their statement in such a forthright form being justified by the fact that at present one man's guess is as good as his neighbour's.

[1] Before correction for absorption, Shapley's figures were 300,000 and 100,000 light years for the diameter and the eccentricity of the sun respectively.

THE STELLAR AND GLOBULAR CLUSTER SYSTEMS ARE CONCENTRIC.

i. That the position of the sun within the galaxy is eccentric is suggested by

(*a*) the comparative richness of the Milky Way in the direction of Sagittarius, as against the direction 180° away in galactic longitude,

(*b*) the unilateral distribution of the globular clusters, also centred upon the Sagittarius region,

(*c*) the preference shown by the most distant types of galactic object, such as novae, for the galactic hemisphere centred on Sagittarius.

ii. Since the globular clusters are disposed with approximate symmetry about the galactic plane, it is assumed by Shapley that the two centres are coincident.

iii. This centre, in the case of the globular clusters, lies in the same direction as that suggested by the foregoing considerations for the galactic centre, at a distance of very roughly 30,000 to 50,000 light years.

THE STELLAR AND OPEN CLUSTER SYSTEMS ARE COEXTENSIVE

i. More significant than the roughly symmetrical distribution of globular clusters on either side of the galactic equator, is the entirely unstellar distribution of these objects. Whereas the stars are concentrated into a flattened, pancake-like system, the clusters occupy a more or less spherical volume of space.

ii. The distribution of the galactic clusters, on the other hand, is pre-eminently stellar in character. The concentration of the open clusters upon the galactic plane (or one very close to it) is, in fact, even more marked than that of the fainter stars.

iii. The closeness of the correspondence between the two systems is indicated by the closeness of the correspondence of their respective median planes.

iv. Mere inspection shows how intimate is the association of the galactic clusters and the star clouds of the Milky Way.

v. Other stellar systems are known, and are shortly to come under discussion. Shapley's estimate of the size of our own stellar system makes it about five times as large as these similar star systems. One of the nearest of these 'island universes' is about 40,000 light years in diameter, an approximate figure which agrees well with Trumpler's figure for the diameter of the open cluster/galactic system. That this nebula (M. 31) is known to be exceptionally large, accentuates the discrepancy in Shapley's figures,

Figure 39. Filamentous galactic nebula in Cygnus. (*By courtesy of the Director of Mt. Wilson Observatory.*)

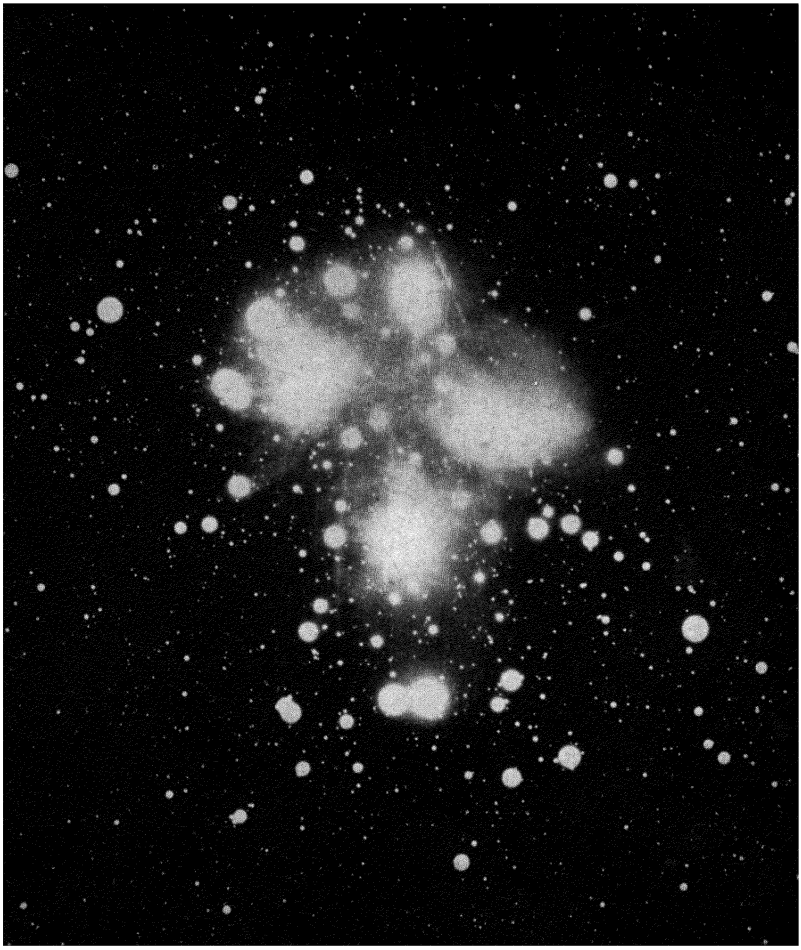


Figure 40. Nebula surrounding the stars of the Pleiades. (*From a photograph by Dr. Max Wolf in* Modern Cosmologies, *by Victor Macpherson. Oxford: The Clarendon Press.*)

although, as we shall learn later, not too much reliance must be placed on this analogy.

The realm of the extragalactic nebulae

The objects referred to in the last paragraph are the extragalactic nebulae, or so-called 'island universes'. They are exceedingly numerous, the positions of some 20,000 having been specifically determined, this perhaps representing one-tenth of the number whose images are recorded on existing plates. Hubble, the foremost authority in this branch of astronomy, has estimated that were the galaxy perfectly transparent, something like 100,000,000 extragalactic nebulae would be within long-exposure photographic range of the Mt. Wilson 100-inch reflector—within, that is to say, a distance of about 5×10^8 light years. Thus the observable region of space is a sphere centred upon the earth, whose present diameter[1] is of the order of one thousand million light years (cf. the distance to the remotest galactic objects—probably not much more than 100,000 light years).

These myriad, faint, enigmatic objects have been the subject of of energetic speculation for the better part of two centuries, and the idea that they may be stellar systems similar in general outline to our own—i.e. true 'extragalactic' objects—is far from new. In the time of Kant (mid-eighteenth century) the conception of 'island universes' was already considerably developed and bandied about as a philosophical speculation. The conception remained in this category of human ideas until little more than twenty years ago, when the distances of some of the nearer nebulae were at last determined, and the theory of external systems elevated from the plane of speculative fancy to that of ascertained fact.

Apparent distribution of the extragalactic nebulae

But before we can describe in any detail the methods whereby this most important advance was achieved, something must first be said of the distribution of the extragalactic nebulae. Their disposition upon the star sphere is even at first glance peculiar, and unlike that of any class of object which we have yet studied.

Briefly, the spiral and elliptical nebulae[2] are orientated upon the Milky Way, a fact that was for many years held to indicate that they were at any rate to some extent dependent upon the galaxy, and therefore probably not extragalactic; from the Milky Way itself they are completely absent; they are of rare occurrence for some distance

[1] The construction of the 200-inch reflector at Mt. Palomar was suspended for the duration of the war.

[2] The two main categories into which the extragalactic nebulae fall. Elliptical nebulae are illustrated in Fig. 68; and spirals in Figs. 69, 70, and 71.

on either side of it; and it is only in the regions nearer the galactic poles than the equator that they are encountered in large numbers. Thus they reverse the usual procedure as exemplified by the galactic clusters, faint novae and distant variables, and star clouds, all of which congregate about the galactic plane and are conspicuously absent from the polar regions. Before the first extragalactic distances were incontrovertibly established this aspect of their distribution was the usual rejoinder of adherents of the 'extragalactic school' to the intragalactic argument just mentioned. Although the extragalactic objects of all types share with the globular clusters an avoidance of the galactic plane, they differ from them in occurring with equal frequency in all galactic longitudes. Thus in the matter of distance, the globular clusters might be expected to represent a sort of half-way house between galactic objects on the one hand and the extragalactic objects on the other.

Methodical and painstaking surveys have filled in the details of this rough outline. Three distinct zones may be distinguished. The first is an irregular strip, from 10° to 40° wide, circumscribing the star sphere; the median line of this strip is the galactic equator. This zone is one of complete avoidance. The second zone, or rather pair of zones, consists of a fringe along either side of zone 1, and is one of partial avoidance. The last pair of zones consists of the rest of the star sphere: two polar caps extending from the galactic poles down to the edges of zone 2. It is in this third zone that the spirals and elliptical nebulae are found in their myriads, but even here their distribution is not precisely uniform: area for area, progressively larger numbers of nebulae are encountered the further one moves the direction of observation from the lower galactic latitudes towards the poles. The precise nature of this numerical increase is peculiar and its significance will be discussed shortly.

Effect of interstellar absorption on apparent distribution of the nebulae

Having determined the apparent distribution of the extragalactic nebulae it now remains to account for its peculiarities in terms of three-dimensional or space distribution. The clue to this translation has already been placed in our hands by Trumpler. Not only is the piling up of star clouds and diffuse nebulosity in the plane of the galaxy apparent from any photograph of the region, but the spectroscope and the investigation into the linear diameters of the open clusters both indicate the existence of an absorbing layer, probably somewhat thin, extending throughout the galactic plane. It is this galactic absorption, then, which is responsible for the first zone in the apparent distribution of the extragalactic nebulae: the zone of total

avoidance is rather one of 'total obscuration', the galaxy itself hiding from our eyes those extragalactic objects that lie beyond it in the extension of its median plane.

The second zone, the fringe along either side of the 'zone of avoidance', derives from the gradual thinning of the absorbing layer along its limits. But what of the third zone, that extending from the second towards the poles? Since obscuration, rather than a graded spatial distribution, is the cause of the apparent anomalies of distribution within the first and second zones, it is natural to look for a similar cause here. And that this is correct is shown by the fact that the number of nebulae visible in successive galactic latitudes from the edge of zone 2 up to the poles themselves, increases in the same manner that the visibility of faint stars increases as they move from the horizon towards the zenith in their diurnal revolution of the star sphere. In other words, the particular increase in numbers visible in increasingly higher latitudes is of such a nature as indicates a progressively smaller degree of absorption by some gaseous or dusty medium, the relative degree of absorption in different latitudes being proportional to the respective lengths of the light paths through the absorbing medium. Not only, therefore, is there a heterogeneous and relatively dense layer of absorbing matter in the galactic plane, but also a tenuous absorbing cloud probably coextensive with the whole star system.

Hubble has estimated that the optical thickness of the galactic absorbing layer is about half a magnitude, and when allowance is made for this absorption, the apparent dependence of extragalactic distribution upon galactic latitude vanishes. The large-scale distribution of the nebulae is seen to be uniform in different galactic latitudes, and at the same time no appreciable systematic variation of numbers with galactic longitude is to be detected. Furthermore, the distribution of the nebulae throughout the two polar caps is, within the limit of error, identical, indicating that the sun is approximately centrally placed relative to the minor axis of the absorbing layer, and that this layer is indeed coincident with the galactic plane.

Large-scale spatial distribution of the nebulae

This important result—that, when the effects of galactic absorption are allowed for, the large-scale distribution of extragalactic objects is the same in all directions—was reached by means of surveys of nebula-numbers down to a limiting magnitude of 20 in over 1,000 sample areas distributed over the whole star sphere. But in addition to a 'surface' survey of this sort, it is possible to carry out

surveys in depth. Granted the assumption, certainly justified, that *in general* a faint nebula is more distant than a bright one, it is possible to plot their numbers in successive concentric spheres all centred upon the earth and each associated with a particular limiting magnitude; simply to count the numbers of nebulae brighter than each limiting magnitude is to provide us with invaluable information regarding their distribution in depth.

When this is done, the same spatial homogeneity is encountered. The whole of the observable region—some five hundred million million million million cubic light years—is uniformly and impartially populated with nebulae. Nor is there any indication whatever that the present limits of vision, the outer edge of the observable region, are anywhere near the limits of the system of extragalactic objects, for with the numerical increase in magnitude their numbers increase steadily, and (when certain corrections for the red-shifts are made)[1] exhibit no falling off as the limits of the observable region are approached. This is in striking contrast with the stars; we know that, in the direction of the galactic poles, existent instruments can probe almost, if not quite, to the confines of the galaxy, for fewer of the faintest stars are visible than would be expected were there no thinning out of stars in these distant regions. Since the extragalactic nebulae show no such thinning out, it follows that with numerically increasing magnitude the nebulae catch up with the stars in number. From magnitude 1 down to magnitude 21 there are at the galactic poles more stars than nebulae, area for area, though the numerical superiority of the former decreases with each successive magnitude. At magnitude 21·5, however, stars and nebulae are equally numerous. This happens to be the limiting magnitude of the Mt. Wilson 100-inch reflector, but there is no doubt that when the 200-inch instrument is completed it will show that, at lower magnitudes than this, the extragalactic nebulae are actually more numerous than the stars.

Whereas present equipment can reach to the limits of the galactic system in the direction of the poles, the system of extragalactic nebulae stretches out beyond our reach; there is no observational indication that they do not 'go on for ever'. It is worth noting that this clear distinction between the extragalactic nebulae and the stars, in the matter of their distribution in depth, is a conclusive demonstration of their true extragalactic status, and remains so independently of any linear distance determinations.

[1] Red-shifts, a spectroscopic phenomenon which will be discussed later, tend to make the more distant nebulae appear fainter than they really are; hence they appear to be more distant than they are, which in turn produces a spurious effect of thinning out.

Small-scale spatial distribution of the nebulae

Although the large-scale spatial distribution of the extragalactic nebulae is uniform in all directions and at all depths throughout the observable region—their average separation one from another being 2×10^6 light years—their small-scale distribution is highly irregular. Apart from single nebulae and small groups consisting of several members, there is also a number of giant clusters each containing some 500 nebulae.

Of the former the 'local group', consisting of 13 known members, all within about 10^6 light years of the galaxy, may be taken as tolerably representative. The largest member of this sub-family of 'extragalactic' systems is the galaxy itself. The Magellanic clouds (see Fig. 41) both on account of their nearness, small size, and unusual structure, may probably be regarded as extragalactic 'satellites' of the galaxy rather than extragalactic nebulae in the accepted sense of the term. Their distances from the sun are about 85,000 and 95,000 light years respectively; and their diameters 12,000 and 6,000 light years. The giant M.31 of Andromeda is to northern terrestrial eyes the most striking member of the group. Distant some 760,000 light years, it also has two satellite galaxies, M.32 and N.G.C.205 (see Fig. 42). M.33, a late spiral, is the most distant of the 'local group'. It was to this group of nebulae that the initial Cepheid observations were confined.

The giant clusters are of infrequent occurrence, only about twenty having been definitely established so far. The largest and nearest is situated in the constellation Virgo at a mean distance of some 7×10^6 light years; that of the most remote yet studied is about $2{\cdot}4 \times 10^8$ light years. The proof by distance determinations that these are true spatial clusterings is reflected in the fact that the members of any one cluster are mostly of the same type: in one, for instance, spirals predominate, in another the globular and early elliptical types. But the diversity of types within any given cluster is considerable enough for the interesting fact to be established that there is a continuous increase in size from the early globular nebulae to the late-type spirals. It is not even required to know the linear distance of a cluster of nebulae in order to establish this fact, for all its constituent nebulae are at substantially the same distance from the observer and consequently linear diameters are directly proportional to apparent diameters.

Distances of the extragalactic nebulae

The ground is now cleared for a discussion of the linear distances of the extragalactic nebulae. The problem has been attacked and

solved from a number of angles, and their extragalactic nature is now established beyond possibility of reversal—less by any single and conclusive proof than by the uniform consistency of results derived by independent methods.

It was not until the early twenties of the present century that individual stars were finally and incontrovertibly detected in an extragalactic nebula; before that time—indeed, as long ago as 1889—there had been indications of the stellar nature of certain regions in the larger spirals. Photography had effected partial resolutions but whether the quasi-stellar points whose images were imprinted upon the plates were stars or unresolved clusters, or possibly some form of stellar progenitor, was not certain. Again, novae had been found in some nebulae, but the evidence of these was to a certain extent contradictory and the significance of the observations difficult to evaluate, owing primarily to incomplete knowledge of these objects. The status of the 'extragalactic' nebulae was, in fact, a matter for speculation, but, on the incomplete data then available, for nothing more.

Evidence of individual stars

In 1923, however, the whole complexion of the controversy was changed. The Mt. Wilson 100-inch reflector not only carried resolution of M.31 further than had been achieved previously, but for the first time established that the questionably stellar points were indeed stars: for some of them exhibited Cepheid variation. As so often happens in the progress of science, the breaching of a long unassailable barrier was quickly followed up by a spurt of new developments in the same direction, and by the end of the decade forty Cepheids and well over eighty novae had been discovered in the Andromeda nebula, while Cepheids had also been detected in other spirals. The maximum magnitude of the first Cepheid in M.31 was only 18·2; yet its period of about one month indicated an absolute magnitude of −4, or 7,000 times that of the sun. For a star of this luminosity to be reduced to an apparent magnitude of 18·2 it must be removed to a distance of some 900,000 light years from the observer. At such a distance it would be well beyond the furthest frontier of the galactic system. On the evidence of a single star, abundantly confirmed later, the Andromeda spiral was thus shown to be truly extragalactic: a hypothetical observer on an imaginary planet revolving about a star in the nebula M.31 would call our stellar system an extragalactic nebula—probably of late spiral type.

Parallel investigations confirmed this general order of distance. The brightest stars visible in the nebulae to which these early

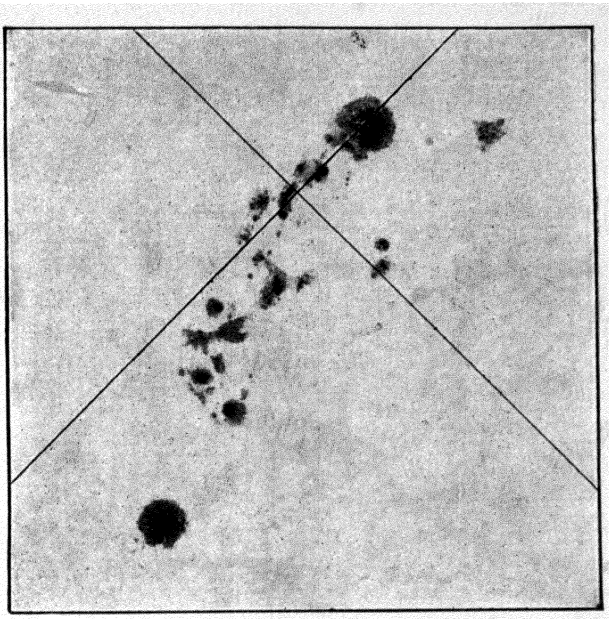


Figure 58. The Nebecula Major, a sub-system of the Galaxy. (*From* General Astronomy *by Sir Harold Spencer Jones. London: Edward Arnold & Co.*)

ERRATA

The block facing p. 215 should face p. 126 and the caption should read, 'Figure 41. The Nebecula Major, a sub-system of the Galaxy.'

The block facing p. 126 should face p. 215 and the caption should read, 'Figure 58. Sun spots photographed in integrated light, showing details of umbrae and penumbrae.'

Sidgwick : The Heavens Above.

sequence, yet its luminosity remains tolerably constant at all stages from globular to late spiral. Thus a new method of calibration is established—a correlation between distance and the nebula's total apparent brightness. The permitted range of luminosity among the extragalactic nebulae was quickly discovered to be small: the luminosities of half the near nebulae whose distances had been established by observations of involved stars lay between one-half and twice the mean value of $8{\cdot}5 \times 10^7$ times that of the sun. With this new weapon, the distance of any visible nebula may be estimated with a probable error not in excess of 20 per cent. to 30 per cent.—an incredible achievement, as will readily be acknowledged when it is reflected that the frontier of the visible universe is something like 500 million light years distant.

The red-shifts as a distance criterion

It has already been remarked that inner consistency is the main justification for confidence in the scale of distances revealed by the extragalactic nebulae. The correctness of the extrapolation from the Cepheid and other stellar determinations of the nearer nebulae to the brightness determinations of all nebulae within the observable region is strikingly confirmed by the so-called 'distance-velocity' relation, or, more prudently, the red-shifts. This phenomenon—the happy hunting ground of popular science writers in the Press, and the subject of much loose and illegitimate speculation by the scientifically-minded man in the street—remains one of the foremost enigmas of the fundamental pattern of the universe as at present discerned.

The reader is already familiar with the fact that radial velocities—i.e. velocities in the observer's line of sight—are revealed and measured by the spectroscope. When the source is approaching the observer, the wavelength of the radiation is shortened, and the whole spectrum shifted towards the blue end; when receding from the observer, the wavelength is lengthened, and the shift is consequently towards the red.[1] In the former case, the radial motion of the source is said to be negative, and in the latter positive.

The measurement of the radial velocity of an extragalactic nebula was first accomplished by Slipher in 1912; the nebula was the great Andromeda spiral, and the radial velocity derived was -190 m.p.s., i.e. a velocity of approach. This was followed by results for other of the nearer nebulae, and by the time that the first distances of the spirals were being determined, some forty radial velocities had

[1] The mechanism of velocity shifts will be more circumstantially described in the next chapter.

been measured. A few of these, which had been derived for the brighter and therefore presumably the nearer spirals, had been negative, but as the determinations accumulated, positive radial velocities quickly predominated. Some of these were of a high order of magnitude, strikingly contrasted with the velocities of various galactic objects with which astronomers had, at the time, been solely familiar: another clear indication of the galactic independence of the spirals. The limiting velocities of the group of 45 known in 1925 were -190 and $+1{,}125$ m.p.s., the mean being $+375$ m.p.s. At this time, too little reliable knowledge of the distance criteria had come to light for any correlation of distance with radial velocity to be attempted, but by the end of the decade the startling discovery had been made that the size of the red-shift of the spectrum of any nebula was dependent upon that nebula's distance: the more distant the nebula, the greater the shift.

The relation between distance and red-shift is direct and linear. As the distance increases by millions of light years, so the velocity of recession increases by hundreds of miles per second. More precisely, every increase of 10^6 light years adds 101 m.p.s. to the positive radial velocity. We may express this relation in the form

$$v = 101d$$

where v is the velocity in m.p.s., as indicated by the red-shift, and d is the nebula's distance in 10^6 light years. At once we are furnished with an additional distance criterion which provides an independent check on those already derived. The distances as derived from the red-shifts, combined with the apparent magnitudes of the nebulae, led to their intrinsic luminosities by a simple calculation, brightness decreasing with the square of the distance. Intrinsic luminosities of a number of nebulae had also been calculated on the basis of distance determinations by means of involved stars. And it was found that there was a good correspondence between the two sets of results. The validity of both lines of approach to the problem of extragalactic distances was thereby strengthened, for it would be quite beyond the bounds of probability that an identical numerical error was involved in each.

Measurement of the red-shifts of the fainter, more distant nebulae was carried on, notably by Humason, until the velocities reported had reached such fantastic values as to cast doubt upon the initial assumption that the red-shifts were velocity shifts—a point to which we shall return when discussing the nature of the extragalactic nebulae. By 1935 Humason had collected some 150 new red-shifts; those of the remotest nebulae studied, distant about 240 million light

years, corresponded with positive radial velocities of 26,000 m.p.s., or one-seventh of the speed of light! Humason confirmed the linearity of the 'distance-velocity' relation over this greatly extended range, and although practical difficulties prevent the measurement of the red-shifts of the remotest known nebulae, there is no doubt whatever that the relation holds good for them also. Assuming this to be the case, they must be receding with velocities of the order of 50,000 m.p.s., while it is at a distance only three times as far as we can at present explore that the velocities reach that of light.

Recapitulation

In brief outline, the foregoing account represents the extent of our present knowledge relating to the dimensions of the extragalactic universe, and of our galaxy as a member of that universe. From naked-eye observations of the diurnal and annual behaviour of the star vault we have, by a series of over-lapping bounds, reached a limit—500 million light years distant—at which our knowledge is finally arrested.

The initial step of this journey was to establish the true figure of the earth. This necessary preliminary was followed by the disentanglement of terrestrial from extraterrestrial motions, and the discovery that the veil of delusive appearance that distorts our vision of reality is largely the creation of the earth's own motions—its axial rotation and circumsolar revolution, neither of which is directly apprehended as such by the senses.

Man's progress in cosmological understanding was then traced from Ptolemy, to Copernicus and Kepler, and on to Newton; it was shown that Kepler's laws of planetary motion were corollaries of the single law of universal gravitation.

The true spatial arrangement of the various members of the solar system was thus established, but it was still a map without a scale. The first object to be surveyed in was the moon, by means of trigonometrical parallax from widely separated stations on the earth's surface. To determine the sun's much greater distance, however, required new principles of distance determination. Of these, the most important is the prior determination by trigonometrical parallax of the distance of a near planet such as Eros, whence the solar distances of all the other planets (including the fundamental earth-sun distance) can be derived by the harmonic law.

The leap from the solar to the stellar system involved yet another search for new and more powerful distance criteria, interstellar distances being incomparably greater than those between planets. The successive extrapolations, by means of which further and

Figure 42. M. 31, a prominent spiral nebula in Andromeda; its distance is 760,000 light years. The two other extragalactic nebulae shown, M. 32 and N.G.C. 205, lie within 35,000 light years of M. 31. (*From* The Realm of the Nebulae, *by Edwin Hubble. Yale University Press.*)

further reaches of the stellar universe are brought within the astronomer's grasp, were then described; this led to a discussion of the stellar system as a whole, and of its structure and size as revealed, in particular, by studies of the clusters.

Again a gigantic leap had to be taken—from the realm of the stars to that of the extragalactic nebulae. And once again, the discussion was restricted to the problem of their spatial distribution, including their distances.

But of the nature of these various bodies—planets, stars and nebulae—next to nothing has been said. The picture that has been drawn is no more than a geometrical diagram. It is now necessary, therefore, to retrace our steps and to reconsider the heavenly bodies, not merely as stepping stones across the observable region, but as things in themselves.

Part II

QUALITY—THE NATURE OF THINGS

VI

ENTRY OF THE SPECTROSCOPE

THE man in the street will believe the evidence of his eyes—little realizing what a fallible criterion of the truth this may be—provided that such evidence is, so to speak, direct, and not in need of interpretation. Thus he does not regard the astronomer as transgressing the bounds of reason when he states that the existence of mountains on the moon is proved by the fact of their being visible with a telescope. A photograph of the moon, taken with the aid of a telescope, requires no more interpretation than does 'common-sense' everyday experience: there are the mountains, plain and clear for anyone to see.

But presented with a spectrogram of the sun—a photograph taken with a spectroscope in addition to a telescope—and told that it constitutes conclusive proof that (i) the sun is rotating on its axis in a period of some four weeks, (ii) that its surface temperature is in the neighbourhood of 6,000° C., (iii) that it is in early life, (iv) that it is possessed of an atmosphere whose temperature is lower than that of the radiant surface, and (v) that this atmosphere contains calcium, iron, hydrogen and nearly sixty other elements, all of which are also found on the earth—told all this, and our 'common-sense' man may well be sceptical. For the spectrogram seems to be—and, without proper interpretation, is—totally unconnected with any of these facts; this will be appreciated from Fig. 44. A considerable degree of interpretation of the spectrogram is necessary before this information emerges: the fact of the sun's surface temperature does not stand out from the spectrogram in the same obvious way that the fact of the moon's mountainous surface stands out from a photograph of the moon.

It is the ultimate purpose of this chapter to show that such interpretation of spectroscopic data is valid: that, for example, the existence of solar calcium is as indisputably proved by the spectroscope as is the existence of lunar mountains by the telescope.

The role of light

The connecting link between any celestial body and the astronomer's instruments is light. A body is visible when light is reflected from it to the observer's eye, as in the case of the moon and planets, or radiated from it as in the case of the sun. In the latter

case we know, incidentally, that the source of the light must be hot, since a body has to reach a tolerably high temperature before it incandesces; we know this of the stars also, although their great distances render their heat imperceptible without the intervention of delicate instruments.

It must be our first concern to discover what we can about this radiation, by whose agency all our knowledge of the heavenly bodies has been derived—to learn why some bodies radiate light while others do not, what form the radiation takes, the manner in which different forms of radiation are related to one another, and, most important of all, what changes are actually occurring within a substance while it is radiating—in other words, the physical operations underlying the process.

In order to tackle this last point, the crux of a vast amount of the physical research of the past few decades, we have to readjust our mental focus from the gigantic cosmic distances of the last chapter to those most minute, ultimate particles of which the material universe is constructed. Whereas we have up to now been dealing with distances measurable in thousands or millions of light years, we have to turn to the realm of the ultramicroscopic, in which one hundred thousandth of a centimetre is a handy unit of measurement.

The action of the prism

A few simple experiments will serve to make clear some of the fundamental characteristics of the behavour of light. The scene of

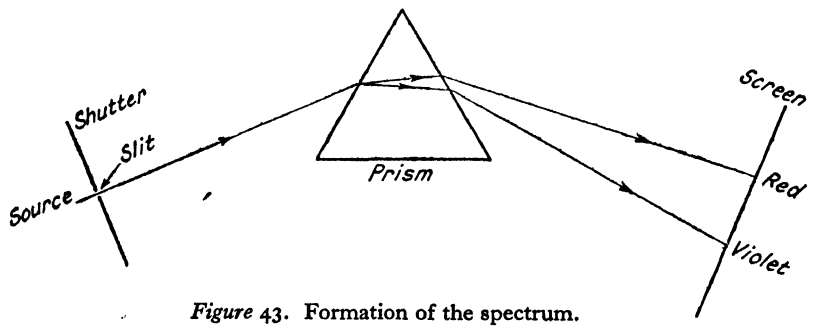


Figure 43. Formation of the spectrum.

the first of these is a darkened room; a small chink left in the shutters admits a narrow beam of sunlight. If this beam is intercepted by a white screen it will be observed that there is a thin line of colourless light upon the screen's surface. Now interpose a glass prism between the screen and the shutter (Fig. 43), in such a position that the beam falls obliquely upon one of its faces. The appearance of the line of light upon the screen is immediately changed. The narrow streak of

white light is drawn out laterally to form an elongated band of colour. It is of the same height as the original streak but many times wider; from one end to the other of this band there is a steady gradation of colour from red, through orange, yellow, green and blue to violet.

Since the prism is only acting upon the light that is already there—upon the beam of white light—and is not adding colours of its own, it follows that the lights of different colour now separated upon the screen were originally constituents of the white beam. The prism's action on this beam is merely to disentangle the various components which, when blended, constitute white light. The inference is that white light is polychromatic, i.e. that it is composed of many different colours.

It will be seen from Fig. 43 that the prism bends or refracts the incident polychromatic beam, and, furthermore, that it refracts the violet constituents through a larger angle than the red. It is this property of refracting lights of different colours in varying degrees that enables the prism to form the coloured band known as the spectrum.

Corpuscular and wave theories of light

This experiment, which proves the polychromatic nature of white light, was first performed by Newton in 1665. In addition to refraction, he studied other aspects of the behaviour of light, such as reflection, dispersion, diffraction and interference. These experiments need not be described here, but their upshot was that in order to explain their results Newton evolved the corpuscular theory of light. This theory conceives a visible body as emitting or reflecting a stream of material particles, these, when they impinge upon the retina, causing the sensation of light. Further work showed, however, that this explanation is incapable of explaining all the modes of the behaviour of light; the alternative theory that light is a wave phenomenon was proposed by Huyghens, Young, and Fresnel. During the nineteenth century Clerk Maxwell elucidated the theoretical reasons for believing light to be an electromagnetic phenomenon, as subsequently demonstrated experimentally by the German physicist, Hertz. Instead of regarding light as a stream of particles emanating from the source of the radiation, we must look upon it as a system of waves radiating away from the source in ever-widening circles.

Wavelength and frequency

Suppose that in the centre of a pond there is a source of disturbance which breaks the surface of the water at regular intervals: it

might be a succession of pebbles dropped into the water at one-second intervals. Each time a pebble breaks the surface, a single wave travels out from the source towards the banks. A cork floating on the pond somewhere between the bank and the centre of the wave system will bob up and down once a second. Now, the point where the pebbles break the surface is analogous with a source of visible radiation—the sun, for example, or the filament of a lamp—in which something is happening which sets the light waves in motion. These travel outward through the three dimensions of space just as the waves of the pond spread out across its two dimensional surface. Finally, the effect of these waves on the cork represents the effect of the light waves upon the retina of the human eye.

If the fall of the pebbles is manipulated in such a way that the surface of the pond is disturbed at regular intervals, it will be found that the distance separating adjacent wave-crests or wave-troughs is equal in all regions of the system. This distance is known as the wavelength of the particular wave system or radiation. If, instead of the pebbles being dropped at one-second intervals, they are dropped at intervals of half a second, it will be discovered (i) that the wavelength has been halved, and (ii) that the frequency has been doubled. This means that twice the number of waves now occupy a given distance and therefore twice as many pass a given point in unit time, since the velocity of the waves is invariable; hence the cork will oscillate twice as rapidly as it did when the release of pebbles was once per second. Thus the greater the wavelength the smaller the frequency, and vice versa. This relationship may be expressed in more definite form by saying that the wavelength is inversely proportional to the frequency:

$$v \propto \frac{1}{\lambda},$$

$$\text{or } v = \frac{c}{\lambda}$$

where v=the frequency of the wave system,
λ=its wavelength,
c=the velocity of the wave system.

The electromagnetic gamut.

By means of an instrument known as the interferometer it is possible to measure the wavelengths of different radiations very accurately indeed. It is found that the only difference between red light and violet light is one of wavelength (and with it, of course frequency). Red light has a longer wavelength than violet, but they

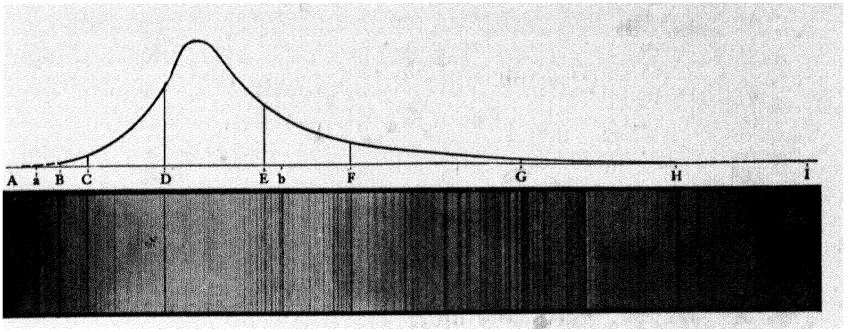


Figure 44. The solar spectrum: Fraunhofer's original map. (*By courtesy of the Director of the Science Museum, South Kensington, London.*)

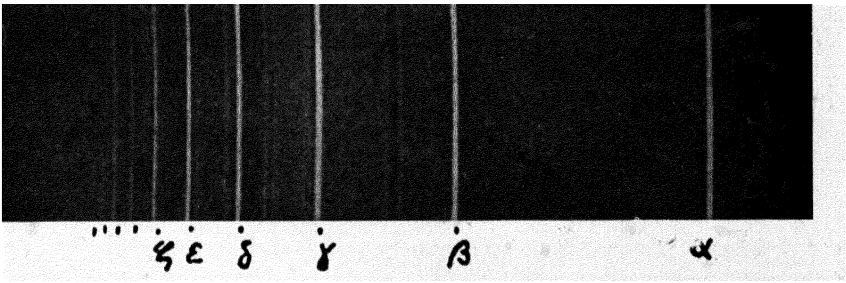


Figure 45. The Balmer series in the emission spectrum of hydrogen. (*By courtesy of the Director of the Department of Physics, Imperial College, London.*)

are both examples of the same type of radiation. They travel with the same velocity as one another (this, we have seen, is 186,000 m.p.s.), just as the velocity with which the waves crossed the pond was the same when the wavelength was long (frequency of 1) and when it was short (frequency of 2). The distances concerned are very minute, the wavelength of red light being 0·00008 cm. and that of violet light 0·00004 cm.

These wavelengths represent the limits within which radiation of this type, known as electromagnetic radiation, can be detected with the human eye. But there are electromagnetic radiations whose wavelengths lie beyond both these limits, and these are detectable by other means. If the wavelength is decreased past 0·00004 cm. it becomes invisible to the naked eye but leaves its imprint upon the photographic plate; such radiation is known as ultra-violet. If it is increased past 0·00008 cm. it is found that the radiation is heat, and can be detected by the human skin and by thermometers and thermo-couples. If the wavelength is lengthened still further the radiation is capable of affecting resonant electric circuits; this radiation, with wavelengths from 10 cm. to about 25,000 metres, is utilized in radio transmission. At the other end of the scale, a shortening beyond the ultra-violet produces, among others, X-radiation and the very hard γ-radiation associated with radioactive decay.

This vast range of radiation is of the same nature throughout, although the human organism uses different senses to perceive different sections of it or even has to fall back upon instruments for its detection. These different sections are only distinguished from one another by difference of wavelength: we might almost say that radio, or Hertzian, waves are nothing more than light with a very long wavelength. It should be noted, too, that as in the pond experiment the velocities of all sections of the electromagnetic gamut are the same: cosmic rays, X-rays, ultra-violet, visible, and infra-red radiation, and radio waves all travel with a uniform velocity *in vacuo* of 186,000 miles per second.

Returning to the action of the prism upon white light—that section of the electromagnetic range whose wavelengths lie between 0·00004 and 0·00008 cm.—we are now in a position to say that this action consists in deflecting the various constituent radiations through angles that depend solely upon their wavelengths. Thus to any section of the spectrum a definite wavelength can be assigned.

The atomic structure of matter

That matter is not in the final analysis continuous, but consists of discrete particles—protons and electrons, atoms and molecules—is

to-day a commonplace item in the mental landscape of every well-read man in the street. The ultimate constituents of matter are minute positive and negative charges of electricity, associated into atomic systems, the clearest picture of which is that presented by the Danish physicist Bohr. The atom of every elementary substance is conceived as consisting of a positively charged nucleus, around which revolve a number of negatively charged particles, known as electrons. The atoms of different substances differ one from another only in the number of these orbital electrons; but in the neutral atom of any substance the total negative charge of the electrons is exactly balanced by the residual positive charge on the nucleus. The simplest atom is that of hydrogen, which consists of a single proton for nucleus, and a single orbital electron. Although the proton and the electron carry identical but opposite charges, the greater part of the atom's mass lies in the nucleus. Laboratory experiments indicate the following data for the hydrogen atom:

Relative masses	
Nucleus (1 proton): orbital electron	1800 : 1
Diameter of nucleus	10^{-13} cm.
Diameter of electron's normal orbit . .	10^{-8} cm.

Thus, were it possible to remove all the electrons from a quantity of hydrogen gas, its mass would not be materially affected.

The atom may be likened, somewhat loosely, to a miniature solar system, with a central sun (the positively charged nucleus) and a number (one, in the case of hydrogen) of revolving planets, the electrons. But the distinctive feature of the Bohr model, namely, that the electrons may leap from orbit to orbit, must be shelved for the moment.

Stationary states and resonance potentials

Experiments performed by Goucher, Hertz and a number of other workers, have revealed some interesting and highly significant facts about the behaviour of atoms, and these must now be described briefly.

The apparatus used in these experiments consists in principle of (i) a box containing hydrogen, mercury vapour, or other gas, (ii) a device for projecting a stream of electrons into the gas, and (iii) a collector to receive and measure the energy of the electrons after passage through the gas. A further device permits the kinetic energy of the electrons to be varied at will. With apparatus of this type it was found that when the electrons were projected into the gas with low initial energy, their energy at the collector was only slightly

reduced: the difference between initial and final kinetic energies being accounted for by the collisions with atoms of the gas which the electrons had sustained during passage. In the same way, a billiard ball will be travelling more slowly after impact with a second ball than before it; a part of its initial energy of motion having been utilized to move the impacted ball.

But as the energy of the electron stream is raised, a point will be reached when the electron current t the collector falls sharply. In the case of magnesium vapour this occurs when the electrons are injected with an energy of 2·7 electron-volts. The same phenomenon occurs again when the energy of the electrons is further raised to 4·4 E.V.

It thus appears that when an electron carrying exactly 2·7 E.V. suffers collision with a magnesium atom, the latter is capable of absorbing the whole of the electron's kinetic energy. Returning to the analogy of the billiard balls: whereas, before, a small proportion of the moving ball's energy was used to move the second ball, now the whole of the moving ball's energy is used, not to make the second ball bounce away from it, but to effect some alteration within the ball itself. If the electron possesses only 2·5 E.V., this particular type of impact (called 'inelastic') cannot take place; any impacts which do occur are of the elastic type, the atom and electron merely bouncing apart with a small transfer of kinetic energy. If the electron's kinetic energy is between 2·7 and 4·4 electron-volts, any impacts will be inelastic, the electron being deprived of exactly 2·7 E.V. and retaining the rest. Thus an impact between an atom of magnesium vapour and an electron carrying 4·0 E.V. will result in the atom's internal energy being stepped up by 2·7 E.V., the kinetic energy of the electron accordingly being reduced to 1·3 E.V. If the kinetic energy of the electrons is raised slightly higher, to 4·4 E.V., a second type of inelastic impact will transfer all the electron's energy to the atom. The magnesium atom is therefore capable of absorbing into itself energy in exact amounts of 2·7 and 4·4 E.V., and in no other amounts. We may therefore say that the atom of magnesium is capable of existing in three distinct energy states—normal, and 2·7 and 4·4 E.V. above normal—and that it can exist in no states with energy-values other than these.

Such states are known as 'stationary states' of the atom, and the equivalent energy-values as 'resonance potentials'. The discovery of these different atomic states was a decisive advance, and an important step towards the solution of other major problems of modern physics. While each element has at least several critical energy-values, all the resonance potentials of every element are different: no single potential

is to be found duplicated among the very large number possessed by the complete list of 92 elements.

The mechanism of emission

The next point to be clarified is the connexion, if any, between the stationary states of an atom, each associated with a definite and characteristic resonance potential, and its power of emitting radiation. Let us consider the simplest atom of all, though what is said of the hydrogen atom applies *mutatis mutandis* to the atom of any other element. The spectrum of glowing hydrogen is shown in Fig. 45, and at even a cursory glance the regular disposition of its lines is apparent. Since the action of the prism is to separate radiations in order of wavelength, it follows that the hydrogen is only radiating a selected number of wavelengths. It is as though a radio transmitter were broadcasting on, say, wavelengths of 500, 700, 800 and 850 metres; a receiving set tuned in to any wavelength other than these (corresponding to the dark sections of the hydrogen spectrum between the bright lines) would receive no signals from the station. Hydrogen, then, only emits radiation of certain, definite wavelengths, and from the ordered arrangement of the spectral lines it is clear that these wavelengths are related in some simple mathematical manner. This series of lines, converging upon the point of the spectrum corresponding to a wavelength of λ3646, is known as the Balmer series, after the Swiss physicist who made it his particular study and who discovered the relationship linking the different lines of the series. He showed that the wavelengths of all the lines in this series in the hydrogen spectrum could be expressed by the formula

$$\lambda = 3646\frac{n^2}{n^2-4}$$

where n may be 3, 4, 5, 6, etc. Thus if we put $n=3$

$$\lambda = 3646 \cdot \frac{9}{5}$$
$$= 6563$$

which is the wavelength of the line designated $H\alpha$. Other convergent series in the hydrogen spectrum occur in the ultra-violet and infra-red regions; these are known respectively as the Lyman and Paschen series, and are of astronomical interest.

Bohr's most pregnant suggestion was that the hydrogen electron can revolve about the nucleus in any one of a number of different orbits, the radii of which are proportional to the terms in the series 1, 4, 9, 16, 25, etc. If it is moving in the innermost orbit—when the atom is

said to be normal—a definite amount of energy must be absorbed by the atom to enable it to jump to one of the outer orbits, in which latter state the atom is said to be excited. The larger the orbit, the greater the energy required to effect the transition. If the atom absorbs one of these exact 'packets' of energy, the electron will leap to a larger orbit.

Such a state of affairs reminds one irresistibly of the stationary states, whose main characteristics we have already examined: we may say that atoms in different stationary states are simply atoms whose electrons are revolving in different orbits, and the larger the orbit, the higher will be the internal energy of the atom.

Quanta

Now an atom always tends towards the normal state, and this transition from excited to normal occurs very rapidly indeed. Consequently an atom which has been 'boosted' to one of the higher energy states (having absorbed from an external source a packet of energy corresponding to one of its resonance potentials, this energy being utilized by the electron in jumping to a larger orbit) will tend to relapse almost instantaneously to its normal state. When this occurs, the same amount of energy will be released. Thus an atom cannot get rid of internal energy in the form of a constant stream, but only in the form of discrete units. The larger the orbit in which the electron of the excited atom was moving, the higher the potential of this excited state, and the larger will be the packet of energy which is released on the atom relapsing to the normal state. This 'packet' of energy is known as a 'quantum', and with each quantum is associated a definite wavelength. The relation between the energy and the wavelength of a quantum is given by a formula derived by Max Planck, who in 1900 published his revolutionary quantum theory, which envisages radiation as a succession of isolated quanta, or 'energy bullets', rather than as a steady stream of energy:

$$e = hv$$

where e = the energy of the quantum,

v = its frequency (equal, as we have seen, to $1/\lambda$),

h = a constant, known as Planck's Constant, whose value is $6{\cdot}5 \times 10^{-27}$.

When, therefore, an atom relapses from a higher to a lower stationary state, a quantum in the form of a flash of radiation is emitted by the atom; and the wavelength of this 'radiation bullet', or photon, depends upon the distance between the two orbits. When the hydrogen electron jumps from the 3rd to the 2nd orbit, $H\alpha$ light is

emitted; when from the 4th to the 2nd, $H\beta$; from the 5th to the 2nd, $H\gamma$, and so on.

The permitted orbits in Bohr's model are just such that radiation of the wavelengths of the Balmer series will be emitted by transitions between them. Thus the phenomenon of the radiation of selected wavelengths by the hydrogen atom is intimately linked with the atom's stationary states, whose existence was first established in the laboratory as the result of work on resonance potentials.

Radiation, according to the quantum mechanics of Planck, can only occur in precise units of e, $2e$, $3e$, $4e$, etc., and not under any circumstances in fractions of e. Since the energy of a quantum is directly proportional to its frequency, it follows that quanta of short wavelength possess greater energy than those of long.

Within the last fifteen years the early quantum theory, together with details of Bohr's atom model, have come under considerable revision and, to an extent, supersession. Regarding the latter, it remains firmly established that atoms can and do exist in various stationary states, and that they pass from one to another of these in the process of emitting radiation; it is doubtful, however, if Bohr's actual picture of oscillating electrons corresponds with any objective reality. The original quantum theory has come under graver suspicion, and although it was undoubtedly a very great advance on any previous conception, it has been tentatively replaced by the matrix mechanics of Heisenberg and the wave mechanics of Schrödinger. These are only susceptible of discussion in mathematical terms of an advanced order, and in an account of this nature the work of Bohr and Planck may be accepted as a reasonably satisfactory account of the theoretical background.

The mechanism of absorption

There is one further aspect of this picture of the mechanism of radiation which must be described, since upon it depend some of the most important and spectacular achievements of astronomical spectroscopy.

We have seen how an atom, in slipping from one stationary state to another and lower state emits a quantum of radiation appropriate to the transition in question. In this way we have accounted for the formation of bright lines, known as emission lines, in the spectrum of a gas such as hydrogen. But in order to pass from the normal to the excited state which must precede emission, the atom has to absorb an identical quantum from some external source; and this preliminary stage of excitation also leaves its signature upon the spectrum. Mixed radiation falling upon the atom will be deprived of certain

wavelengths—those whose energy, as given by Planck's equation, are capable of raising it to the higher energy states—and the spectrum of this radiation will accordingly be crossed by dark lines at certain points, indicating that these particular wavelengths are lacking. Since these wavelengths will be identical with those emitted by the hydrogen atoms when they relapse from the excited to the normal state, the dark 'absorption' lines will occupy precisely those positions in the spectrum which would be occupied by the emission lines of hydrogen. In other words, a Balmer series of *dark* lines will be superimposed upon the bright continuous spectrum. Now since, as we have already seen, all the resonance potentials of all the elements have distinctive values, both absorption and emission spectra of each of the ninety-two elements are different and distinguishable in regard to every single line. This fact is the basis of qualitative spectroscopy, and has proved of incalculable value in providing us with information regarding the physical constitution of the sun, stars and nebulae.

Kirchhoff's experiments

From this, necessarily somewhat sketchy, survey of the theoretical background, we may now turn to the brilliant empirical work of Kirchhoff, which was brought to its culmination in 1859. At that time the theoretical background was lacking, though this in no way detracted from the practical value of his results: fortunately one does not have to know the technical details of the locksmith's art in order to be able to use a key for unlocking a door. Nevertheless, the significance of Kirchhoff's observations has only been appreciated in comparatively recent years.

The apparatus used in these experiments is almost the same as that required for Newton's demonstration of the polychromatic nature of white light. On a bench in a completely darkened room is placed a prism; to the right of it stands the screen on which the spectrum is to be cast (actually, a small telescope for direct viewing of the spectrum); to the left of the prism stands another screen, in which is cut a fine slit. The relative positions of the prism and the two screens are adjusted so that the spectrum of a light source placed behind the slit is cast upon the right-hand screen. If a lump of iron is placed behind the slit and heated to incandescence, a continuous spectrum is formed: it consists of a band of colour, red at one end, violet at the other, with orange, yellow, green and blue between. Clearly, the source is emitting radiation of all wavelengths within the visible range. No matter what solid is used as source—it may be copper, lime, tin, zinc or what you will—the same continuous spectrum is produced. Furthermore, liquids give identical continuous

spectra, as also do gases when subjected to sufficiently great pressures; thus mercury vapour has been made to yield a continuous spectrum when subjected to a pressure of some 20,000 pounds per square inch in the laboratory.

Kirchhoff therefore formulated his first law, which states that *any incandescent solid, liquid or dense gas emits radiation of all visible wavelengths*; in other words, their spectra are continuous and mutually indistinguishable, no matter what the chemical constitution of the source may be.

A volatilized solid is now placed behind the slit. If, for example, the spectrum of sodium is to be studied, a colourless Bunsen flame is placed behind the screen and a few grains of common salt (which contains sodium) are blown into it; alternatively a piece of sodium may be held in the flame on the end of a platinum wire. In either case the flame is coloured a livid yellow, and a spectrum quite unlike that in the last experiment is produced. Instead of a continuous polychromatic background there is now nothing but a pair of fine, closely adjacent, bright yellow lines on a dark background; nothing else is visible. This clearly indicates that no radiation of wavelengths other than these two is being emitted; the lines are yellow because they lie in the yellow region of the (invisible) continuous spectrum, i.e. they have the wavelengths of yellow light. This can easily be demonstrated by marking their positions on the screen, and then, taking care not to alter the relative positions of the prism and the two screens, replacing the sodium flame by a solid source. It will then be found that the marks indicating the original positions of the sodium lines lie in the yellow region. No matter how often the experiment is repeated, the two lines are always produced and always in exactly the same positions, so long as the source contains sodium vapour: in other words, sodium in a volatilized state always emits radiation of these two wavelengths and never emits radiation of any other wavelengths. Therefore, if we find these lines in the spectrum of a star, for example, we know beyond a shadow of doubt that the visible portions of that star consist partially of sodium. Experiments with numerous other elements prove that the bright-line emission spectrum of each is invariable and unique. This, the basis of qualitative spectrum analysis, reflects the more fundamental fact that in the whole range of resonance potentials for the ninety-two elements no potential occurs more than once.

Kirchhoff also demonstrated that incandescent gases likewise yield bright-line spectra when they are not under pressure, and that the spectrum of every gaseous element is unique. When the source consists of a gaseous or volatilized compound—whose ultimate

particles are molecules, or closely knit groups of atoms—the bright lines are extremely numerous and crowded into groups which appear as bright bands. These complex band spectra are, like the line spectra of elementary substances (whose ultimate particles are single atoms), unique for each compound.

Kirchhoff's second law therefore states that *the spectrum of a glowing rarefied gas or volatilized solid consists of bright lines only; if the source is molecular, the lines are more numerous and crowded together into bands.*

Kirchhoff next tried combining the two foregoing experiments. He placed an incandescent lump of some solid behind the slit, and between it and the slit itself interposed a sodium flame. The result was, to him, somewhat startling, for although a composite spectrum was formed, the sodium lines were not bright but dark; they stood out against the yellow region of the continuous spectrum as two fine black lines. If the source of the continuous radiation is removed or screened, the dark lines are immediately replaced by bright ones in precisely the same positions; the dark lines, in fact, become bright with the removal of the continuous background. Since the two vertical strips of the continuous spectrum that underlie the dark lines would be bright were the sodium vapour not interposed between the prism and the incandescent solid, it follows that this vapour is absorbing from the polychromatic radiation passing through it just those two wavelengths which it itself emits; we have already described the physical counterpart of this act of absorption. The significant point in this experiment is that the sodium vapour is at a lower temperature than the source of the mixed radiation producing the continuous spectrum, and it is only under this condition that it absorbs its own wavelengths. When it is hotter than the incandescent solid the sodium lines appear bright once more, and are superimposed upon a less bright continuous background. The same phenomenon is observed if the volatilized solid is replaced by a gas. The third law may therefore be stated as follows: *a gas will absorb from mixed radiation passing through it exactly those wavelengths that it itself is capable of emitting, provided that the source of the radiation is at a higher temperature than itself.*

Forbidden lines

A further class of spectral lines must now engage our attention—the so-called 'forbidden' lines. This is somewhat of a misnomer, for under exceptional circumstances forbidden lines do make their appearance in astronomical as well as in laboratory spectra. The forbidden lines of an element may, however, be defined as those lines

which do not appear in the normal spectrum, but which make their appearance when the radiating material exists under peculiar and uncommon physical conditions. Examples of such conditions are, in the laboratory, intense electric fields; and in the astronomical sphere, exceptionally low densities such as involve widely spaced atoms among

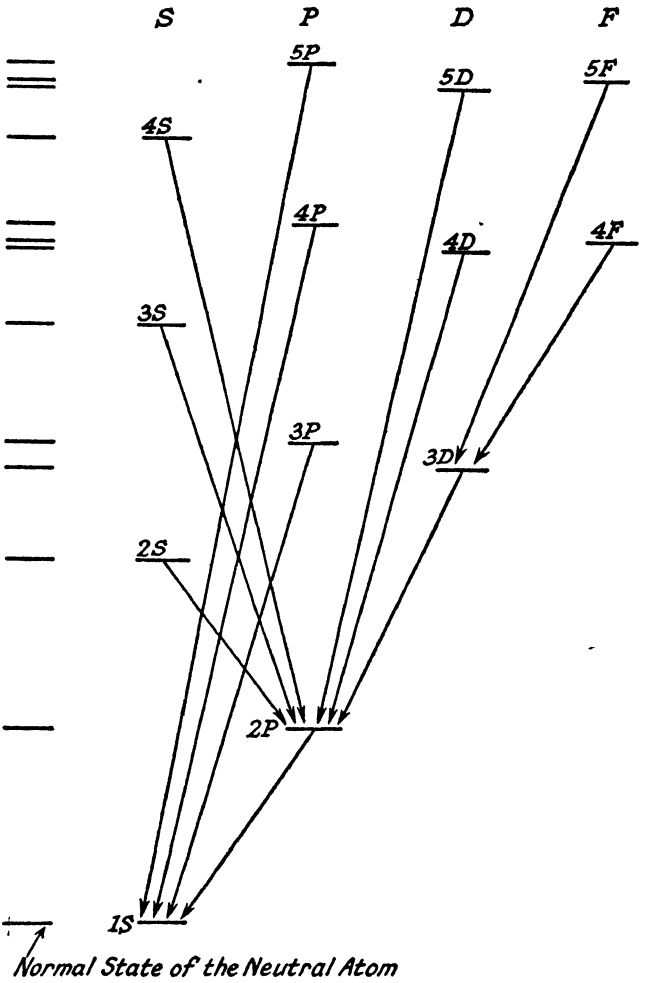


Figure 46. (Modified from Darrow) Normal transitions within the sodium atom.

which collisions will occur only at very much longer intervals than obtain in gases at even the lowest experimentally produced pressures.

The conception of stationary states goes far towards an elucidation of the forbidden lines, and the mechanism of their production. Fig. 46 shows in diagrammatic form some of the stationary states of sodium. These are shown on the left in a single column of ascending

energy-values. To the right they are sorted out into four separate columns, S, P, D and F. From a study of the sodium spectrum it is possible to determine what transitions are responsible for the individual emission lines, and each such transition is shown in the diagram by an arrow linking the two stationary states concerned: for example, three lines in the spectrum result from transitions between the states 2S, 3S, 4S and 2P. It will be observed that the states in columns S, P, D and F have been chosen so that normal transitions—i.e. those producing lines observed in the normal spectrum—occur only between states in adjacent columns. That this is not entirely arbitrary, but does in fact reflect some significant facet of the structure of the atom itself, is shown by the ordered disposition of the levels in each of these columns.

This principle of selection is, however, violated by certain lines, which result from transitions between states in the same column or between states separated by one or more columns. These are the forbidden lines. Viewed in this way they can be seen to depend upon transitions within the atom which are 'forbidden' under normal conditions, but which become permissible under certain special laboratory conditions, as well as under some celestial conditions (such as ultra-low density) which cannot be paralleled in the laboratory.

Even at this stage, then, the spectroscope has vastly increased our powers of investigating bodies which are not directly accessible. Merely by inspecting the spectrum of, say, a star, we can tell (i) if it, or part of it, consists of a solid, a liquid or a dense gas, (ii) if it consists wholly or partially of vaporized solids or liquids, or of rarefied gases, and (iii) what the chemical constitution of these is. And this is only a beginning.

Temperature from spectra

In the first place, it will be remembered that Kirchhoff's third experiment gave information regarding the relative temperatures of the two sources of a composite spectrum: by inspection only, we can tell which of the two is hotter.

Bright-line and absorption spectra give us detailed and absolutely accurate information regarding the chemical composition of the source, but continuous spectra (those of glowing solids and dense gases) give us no information of this sort, since whatever the source may be they are always the same continuous band of colour. Nevertheless, continuous spectra can give valuable information regarding the physical conditions of the source. A careful inspection of a continuous spectrum will reveal the fact that it is not uniformly bright all

along its length: there is a definite zone of maximum intensity. And it is found by experiment that the higher the temperature of the source, the nearer to the violet end of the spectrum is this zone situated; the wavelength of maximum intensity varies inversely with the temperature. By noting the approximate wavelength at which the intensity is greatest, for a number of different source temperatures (which can be varied at will in the laboratory), the physicist Wien formulated the law which bears his name, namely that

$$\lambda_{max}=\frac{b}{T}$$

where λ_{max}=wavelength of the zone of maximum intensity,
T =the Absolute temperature,
b =a constant, whose numerical value need not concern us.

By means of Wien's law it is possible to arrive at an estimate of the sun's surface temperature, or that of a star, merely by direct examination of its spectrum. In Chapter VIII we shall see the application of this law to the problem of the sun's temperature.

The study of an incandescent body's radiation may reveal its temperature by two further methods, the first of which is the employment of the Stefan-Boltzmann law. This states that a body's total radiation in all wavelengths increases with temperature in a particular manner. In exact terms

$$E=\sigma T^4$$

where E=rate of emission of energy,
T=Absolute temperature,
σ=a constant whose value, once again, need not concern us here.

A numerical example of the application of this law will be given in Chapter VIII, when the sun's temperature is under discussion; it will there be seen that the method is of the utmost theoretical simplicity, the main difficulty in its practical application being the sufficiently accurate determination of E.

The third method whereby a study of the radiation from a body[1] enables its temperature to be deduced need not detain us beyond bare mention. It involves the use of a law formulated by Planck in 1900. Planck's equation embodies a more general law of radiation

[1] Accurately speaking, the 'body' to which all these laws of radiation apply is what the physicist terms a 'black body', i.e. a perfect radiator (and therefore absorber, also) in all wavelengths. Since the sun and stars are not black bodies the laws cannot be applied to them with 100 per cent. accuracy, but fortunately their deviation from this theoretical standard is not large enough to cause serious inaccuracy.

than either Wien's or the Stefan-Boltzmann law—these may, in fact, be regarded as special cases of Planck's law—and has been abundantly verified in practice.

The approximate temperature of the source may be determined spectroscopically in other ways. The line spectrum of an element (whether bright or dark—emission or absorption) consists of a series of lines, whose number may vary from only one or two to several hundred, which are relatively widely spaced; widely enough, at any rate, to show as separate and individual lines. But the spectra of compounds, as we have seen, consist of one or more bands which high magnification and wide prismatic dispersion reveal to consist of innumerable lines crowded closely together. Compounds can be identified by their spectra as surely as elements, and are at the same time easily distinguishable from them. Now all chemical compounds are dissociated by high temperatures—i.e. are split into their constituent elements—and the dissociation temperature of each compound is only variable within narrow limits, which can be discovered experimentally. Thus if, for instance, the spectrum of a particular region of the sun shows the banded spectrum of a certain compound, it is certain that the temperature of that region is lower than the compound's dissociation temperature as determined in the laboratory. This, we shall see, is the method whereby the temperature of sunspots has been measured.

Still another means of determining the temperature of the source from the appearance of the spectrum is available to the astronomer. Although the line spectrum of each element is unique, it does itself undergo characteristic modifications with varying temperature. Thus the temperature, not only of an incandescent solid or dense gas, but also of a volatilized solid or rarefied gas may be determined spectroscopically. In the laboratory three ranges of temperature in particular are available for study: the temperature of the flame, such as the Bunsen burner (about 2,000° C.), that of the electric arc (about 3,600° C.), and that of the high tension discharge (up to about 20,000° C.). Experiments carried out at these temperatures have proved that the number and strength of the spectral lines of each element differ slightly in the flame, arc and spark spectra. Thus if the spectrum of a star bears 'enhanced' calcium lines—i.e. lines that can only be produced in the laboratory with the aid of the electric spark—we can assign a certain minimum temperature to that part of the star containing the calcium. In point of fact this technique is capable of giving more detailed results than the foregoing account might lead the reader to suppose; for the gradual emergence and subsequent strengthening of the enhanced lines is a direct function of increasing

temperature, and a close correlation between the two has been established.

Ionization

From what we have already learnt of the structure of the atom and the mechanism of radiation, this appearance of new lines with increased temperature (also facilitated by reduced pressure) would suggest that some radical alteration or structural rearrangement is taking place within the atom itself. This is in fact the case, as has been shown by an extension of the work on resonance potentials already described. As, at each successive potential, the electron is forced out to a more and more remote orbit, a stage is eventually reached after which more energetic bombardment of the gaseous atoms will result in one or more of the outer electrons being knocked out of the atomic structure altogether. The atom will then have a residual positive charge, and can no longer be termed a neutral atom. In this condition it is said to be ionized (singly, doubly, trebly, etc., according to the number of electrons it has lost), the process being known as ionization, and the partially stripped atom as an ion.

The degree of ionization, as reflected in the spectrum, is, then, another way in which temperature may be determined by the spectroscope.

Pressure and density from spectra

Two other important physical conditions of a source, which may be millions of miles distant, can be measured spectroscopically. The first of these is pressure, which modifies the spectrum in certain recognizable ways. The most important of these modifications has already been noted: a gas under great pressure behaves spectroscopically like a solid, giving a continuous spectrum, while a rarefied gas gives the characteristic line spectrum. But line spectra themselves are modified by pressure, just as they are by temperature. In general, the effect of decreasing the pressure of a gas or vaporized solid is the fading and finally, with great rarefaction, the disappearance of all but the strongest lines. Conversely, high pressures tend to widen the lines of the normal spectrum, with the possibility of adjacent lines merging if sufficiently great pressures are applied, and sometimes to shift them slightly towards the red. Reduction of the quantity of a radiating gas likewise produces the gradual disappearance of all but the strongest spectral lines; the very fine lines that remain even under great rarefaction of the source are called *raies ultimes*, and a consideration of the number of these ultimate lines relative to the other lines of the normal spectrum allows a fairly close

estimate of the source density to be made. In the same way, the pressure-shifts and the widening of the lines can be related to an approximate quantitative scale.

Magnetic fields and spectra

It has been proved by laboratory experiments that if the source of a radiation is placed between the poles of a strong electromagnet a curious spectroscopic effect results. In a relatively weak field the spectral lines are thickened, and if the field is strong enough they are split into pairs and triplets. The extent of this doubling and trebling of single lines allows the strength of the field to be calculated with considerable accuracy. One application of this effect, known as the Zeeman effect, will be described when we deal with that section of our knowledge of the sun which has been gained spectroscopically.

Radial motion and spectra

One further application of the spectroscope must be mentioned before we can proceed to a more detailed description of the instrument in its astronomical form. Probably everyone has at one time or another noticed the effect that motion has upon the pitch of a sound. An observer standing on a station platform will notice that the whistle of an express train passing through the station not only increases in volume as the train approaches, but also rises in pitch; once it has passed and is receding from him, the pitch drops again. The reason for this alteration of pitch is that while the train is approaching the observer it is all the time catching up the sound impulses which it is emitting; they are therefore crowded closer together than they would be were the whistle stationary. Since they are closer together, more of them pass a given point in a second, and more therefore impinge upon the observer's tympanum in that time. Since the pitch of a note is determined by the frequency of its impulses (a small tuning fork gives a higher note than a large one because it vibrates more rapidly), the pitch of the whistle will appear to rise. In the same way, when the source of the sound is travelling away from the observer the distance between consecutive impulses is lengthened, and a lower note is heard.

This can be demonstrated quite simply:

Suppose that S is the sounding body, that it is emitting n impulses per second, and that in one second they travel the distance from S to

O, where there is an observer. The space between *S* and *O* will therefore be occupied by *n* impulses. Now suppose that *S* is in motion towards *O*, and that in one second it covers the distance *Ss*. The distance *sO* will now be occupied by *n* impulses; hence they will be closer together (*sO* being shorter than *SO*); hence a greater number will pass *O* in a second; hence he will hear a note of a higher pitch than that emitted by *S* when stationary. If *S* is moving away from *O* the reverse effect will obviously result.

Sound and light affect different human senses because they are phenomena of a fundamentally different character; sound waves, to cite one example, require a medium (normally the air) in which to travel, whereas light and other electromagnetic radiations are propagated *in vacuo*. Nevertheless, the term 'light waves' may be substituted for 'sound impulses' in the foregoing account without invalidating the argument, and it is found that analogous alterations in the wavelength and frequency of a system of electromagnetic radiation are caused by movement of the source relative to the observer. If *S*, now a source emitting light, is moving towards *O*, more waves will pass *O* per second than would be the case if *S* were stationary. Hence—since wavelength and frequency are inversely proportional to one another—the wavelength of the radiation is shortened. Let us suppose, to take an example, that the radiation is monochromatic and that its spectrum consists of a single line in the red; the wavelength of the line can be accurately determined with the interferometer. If the source of the radiation is moving towards the observer, the wavelength will be shortened, with the result that the line will be displaced towards the violet, or short wave, end of the spectrum; were the source receding from the observer the displacement would be towards the red. In either case the amount of the displacement would depend directly upon the velocity of this motion in the line of sight. The size of the displacement, or the amount by which the wavelength has been altered, will be given by the simple relation

$$\Delta\lambda = \frac{v \cdot \lambda}{c}$$

where $\Delta\lambda$ = difference between displaced and normal wavelength of the line,
v = line-of-sight velocity of the source,
c = velocity of light,
λ = the wavelength of the undisplaced line.

$\Delta\lambda$ can be accurately measured, and since the other two terms in the equation are known the velocity of the source can quickly be calculated.

As an example of how this works, suppose that the $H\beta$ line in the blue, whose wavelength[1] is $\lambda 4861$, is observed in a certain spectrum to be shifted to $\lambda 4860$. Then $\lambda = 4861$

$$\Delta\lambda = 1.$$

Substituting these values in the equation, we have

$$1 = \frac{v}{186{,}000} \times 4861$$

$$\therefore \quad v = \frac{186{,}000}{4861}$$

$$= 38\tfrac{1}{4} \text{ m.p.s.}$$

Summary of spectroscopic data

We may now tabulate the varied information that the spectrum of, say, a star gives us about the visible regions of that star:

i. Its physical nature: whether it is liquid, solid, volatilized solid, dense gas, rarefied gas, or a combination of these.

ii. Its chemical composition, and whether compounds as well as elements occur.

iii. The relative temperatures of the component sources of a compound spectrum, e.g. a continuous spectrum with superimposed absorption lines.

iv. Its actual temperature, ascertained by independent methods.

v. Its density.

vi. The presence or absence of a magnetic field.

vii. Its line-of-sight velocity.

The spectroscope

The spectroscope, as an astronomical instrument, is essentially the same as the apparatus on the work bench in the experiments already outlined. The three components, or their equivalents, are combined in a single instrument which can be attached to a telescope. The construction of the prismatic spectroscope is shown diagrammatically in Fig. 47. The tube *a* is known as the collimator, whose function it is to prepare the light of the observed object for the prism. It screws into the drawtube of the telescope at *f*, and at *g* there is a fine slit through which a narrow beam of light passes to the lens *e*. The light emerges from *e* as a parallel beam which falls upon one face of the prism. The emergent, refracted beam then encounters the object lens of the view telescope, *c*, the spectrum being viewed

[1] The adopted unit of wavelength is the angstrom (Å), so named in honour of Ångström, a physicist who did much fundamental spectroscopic work. 1 angstrom $= 10^{-10}$ metres. The $H\alpha$ line at λ 6563, for example, has a wavelength of 0·00006563 cms., or 6563 angstroms, a much less cumbersome expression.

direct instead of being cast on a screen as before. This lens focuses the spectrum upon the eyepiece *d*, where it is magnified and observed. The view telescope is pivoted at the centre of the instrument so that it may be directed at different regions of the spectrum.

This is the basic type; two modifications must be mentioned, but

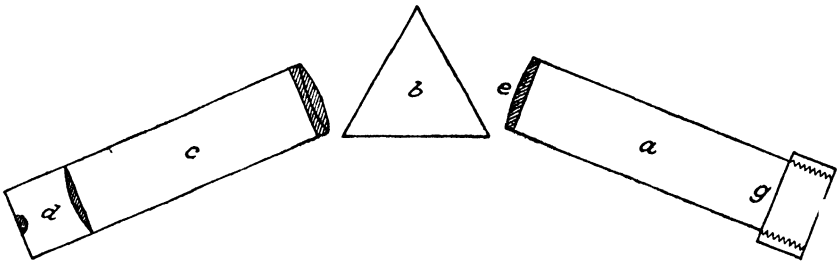


Figure 47. The prismatic spectroscope.

need not be described in detail. To secure wider dispersion, thus increasing the length of the spectrum and securing greater separation of closely adjacent lines, a train of several prisms may replace the single prism shown in the figure. Alternatively, an entirely different principle may be utilized, the formation of the spectrum resulting from the diffraction instead of from the refraction of the incident light. The position of the prism or train of prisms is now occupied by a small sheet of glass or metal upon which are engraved many thousands of fine lines to the inch. When a mixed radiation is reflected from such a surface the different wavelengths are sorted out, as with a prism. A crude example of a diffraction grating can be provided by a gramophone record. If the disc is held up slantwise against the light from a window, it will be seen that the band of light reflected across it contains some of the spectral colours. The dullness of these colours and the poor quality of the spectrum are due to the coarseness of the grating, the disc having comparatively large and few ridges to the inch.

It is usually more convenient to photograph spectra than to study them visually, for not only is a permanent record made in this way, but the measuring of the positions of the lines and the comparison of unknown lines with spectra obtained in the laboratory are greatly facilitated. For this purpose a photographic plate is exposed in the focal plane of the lens of the view telescope, where normally the eyepiece would be.

VII

THE MOON AND THE PLANETS

THE moon is the nearest body to the earth which is comparable with the planets in size: its diameter of 2,165 miles is rather more than one-quarter of the earth's. We have learnt in an earlier chapter that it revolves about the earth in a somewhat eccentric orbit, completing one circuit in about twenty-seven days at a mean distance of 238,860 miles.

The moon's orbit and mass

Its mass, which is less than one-eightieth of that of the earth, may be determined by means of an interesting application of Newton's law of universal gravitation. Up to this point we have always envisaged the earth as describing a perfectly elliptical orbit about the sun, and the moon as revolving about the earth. This is not strictly true, for the earth and moon constitute a single gravitational unit, and it is the centre of gravity of this system which describes the elliptical circumsolar orbit; the earth and moon each describes an orbit about the centre of gravity whose size is inversely proportional to the body's mass. The phenomenon is somewhat similar to that of a pair of exhibition dancers performing one of those athletic manoeuvres in which the man turns rapidly over the same spot, with his partner, held at arms' length, flying through the air round him. If such an exhibition is watched carefully it will be observed that it is an incorrect description to say that the man is rotating on his axis, while his partner revolves round him: in fact, both partners are revolving about a point between them, but this point is much nearer the man than the woman.

If, to employ more accurate terms, M in Fig. 48 represents the

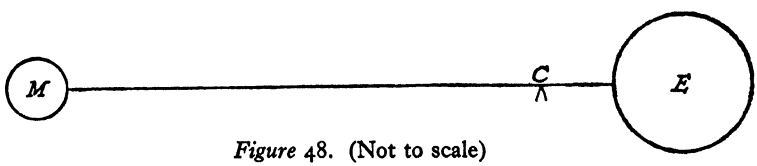


Figure 48. (Not to scale)

moon, and E the earth, the centre of gravity of the system would be at some point C such that were the two bodies connected by a rigid, weightless rod, they would balance at this point. If we write

M_e for the mass of the earth,

M_m for the mass of the moon,

D_e for the distance of the earth's centre from the centre of gravity,

D_m for the distance of the moon's centre from the centre of gravity,

then,

$$M_e \times D_e = M_m \times D_m.$$

Though small, the movement of the earth about the centre of gravity of the two bodies is large enough to cause observable parallactic displacements of the nearer planets such as Mars or Eros when in the vicinity of the earth. Measurement of these displacements allows the distance of the earth from the centre of gravity to be determined,[1] from which M_m can be calculated by means of the equation just given.

The lunar surface

The naked eye shows that the moon's surface is patched and mottled, and observations extended over a period of time prove that the same lunar hemisphere is always turned towards the earth, a fact which must be held responsible for the anonymous housemaid's memorable lines:

> O Moon, when I gaze on thy beautiful face,
> Careering along through the boundaries of space,
> The thought has often come into my mind
> If I ever shall see thy glorious behind.

Telescopically the moon is a wonderful sight, especially when a comparatively narrow crescent. At this time the sun's light falls obliquely upon the whole of the visible surface, and throws its innumerable ridges, craters, valleys and mountains into sharp relief against their jet-black and unfathomable shadows. At full it is a much less spectacular object, because there is then no terminator on the disc, and the terminator is, as we have seen, the line of sunrise or sunset where the illumination is necessarily most oblique. The shadows which gave an appearance of such stark grandeur to the crescent moon are absent, and the general effect is of flatness, lack of relief, and unalleviated glare.

The maria

The most obvious lunar surface features are the dark plains, or maria, which are easily visible to the naked eye. They are seen

[1] The parallactic displacement of the sun, for example, amounts to about 12″. At the sun's distance of $9{\cdot}3 \times 10^7$ miles, 12″ is subtended by 5,760 miles. Hence the radius of the earth's orbit about the centre of gravity of the earth-moon system is 2,880 miles.

telescopically to be vast smooth, or nearly smooth, plains, sprinkled with minute craters, low ridges, and occasionally mountain peaks and larger crater rings. The name 'mare' is a misnomer, since these flat areas are not seas and almost certainly never held water. Our knowledge of the physical conditions prevailing upon the moon proves that they cannot be areas of water, but at this point it will be a good enough demonstration of this fact to point out that the sun is never observed to be reflected in the maria, as it would be were they seas. It will be noticed at first glance that the maria are approximately circular and that they are largely confined to the northern hemisphere, the southern being much rougher and more broken up, consisting largely of crowded and often contiguous crater rings and walled plains. The largest of the maria, Mare Imbrium, has an area of about 350,000 square miles.

The crateriform objects

The most numerous type of lunar surface formation is the so-called crater. The name 'crater' is perhaps unfortunate, since they cannot be likened to the volcanic craters of this planet, as regards either origin or appearance. They vary in size from small, wall-less pits a few hundred feet across, to huge plains, 150 miles in diameter and rimmed with mountain walls many thousands of feet in height. The larger craters frequently have a central mountain mass and a floor bearing much detail in the way of craterlets, rings, ridges and isolated mountain peaks. In general, the smaller a crater is, the more regular are its shape and ramparts, but apart from this and such minor details as the presence or absence of a central mountain, they differ very little among themselves. One almost universal characteristic is that the floor of the crater is at a lower level than the surrounding terrain; thus the drop from the summit of the rampart is greater within the crater than on the outer side.

The lunar mountains

Like the earth, the moon possesses several great mountain ranges and masses; these, when seen under oblique illumination (when near the terminator, that is) are perhaps the most arresting features of the telescopic view of the moon. The height of a lunar mountain or other feature may be determined quite easily by measuring the angular length of its shadow. From this datum its height can be calculated since the moon's distance, and hence the linear length of the shadow, is known, and the only other factor in the problem, the angle of illumination, may be derived from the mountain's distance from the terminator at the time the observation was made. An

alternative method is to determine its distance from the terminator at the moment when the sun's light first catches its summit, making it shine like a star from the darkness beyond the terminator.

In this way it has been found that the highest of the lunar mountains are comparable with the earth's most notable examples. But taking into account the size of the body on which they are situated they far outstrip any terrestrial mountains, since the diameter of the moon is only about one quarter that of the earth. The Leibnitz Mountains, situated on the moon's southern limb, rise in places to 30,000 feet, the equivalent of well over 100,000 feet on the earth.[1] For the most part, however, the great lunar ranges are confined to the northern hemisphere where they are commonly the bulwarks separating adjacent maria.

The bright rays

The maria, the crateriform objects and the mountain ranges are the most important of the lunar surface formations. The telescope reveals in addition a multitude of minor features such as ridges, valleys, faults and clefts. These are all topographical or 'geological' features, objects of rock and soil; but perhaps the most interesting and certainly the most baffling of all the objects visible on the moon are the bright streaks, because they have no terrestrial counterpart. These rays, which occur most prolifically in the southern hemisphere, are well illustrated in Fig. 49. They are apparently surface markings only, for they cast no shadows, and are invisible under low illumination, when all objects in relief cast the longest shadows; on the contrary, they are characteristically a feature of the full moon. At this time they contribute appreciably to the glare already referred to and even render invisible whole ring formations which happen to lie on the territory which they cross. For the most part they are not distributed haphazard over the whole lunar surface but are grouped into definite systems which usually radiate in all directions from a central crater. One of the most important of these systems is that connected with the crater Tycho; the longer of the rays from this centre cover hundreds of miles. Besides their shadowlessness, their invisibility under a low sun, and their grouping into separate systems, a notable characteristic of the bright rays is their inflexibility. They proceed for hundreds of miles across broken, mountainous and crater-strewn terrain without suffering the slightest deviation or interruption from these formations. When a ray encounters a large

[1] It must be remembered, however, when comparing the heights of lunar and terrestrial mountains, that whereas the latter are measured from sea level, the former are measured from a plane of reference that is roughly equivalent (on the earth) with the ocean bed.

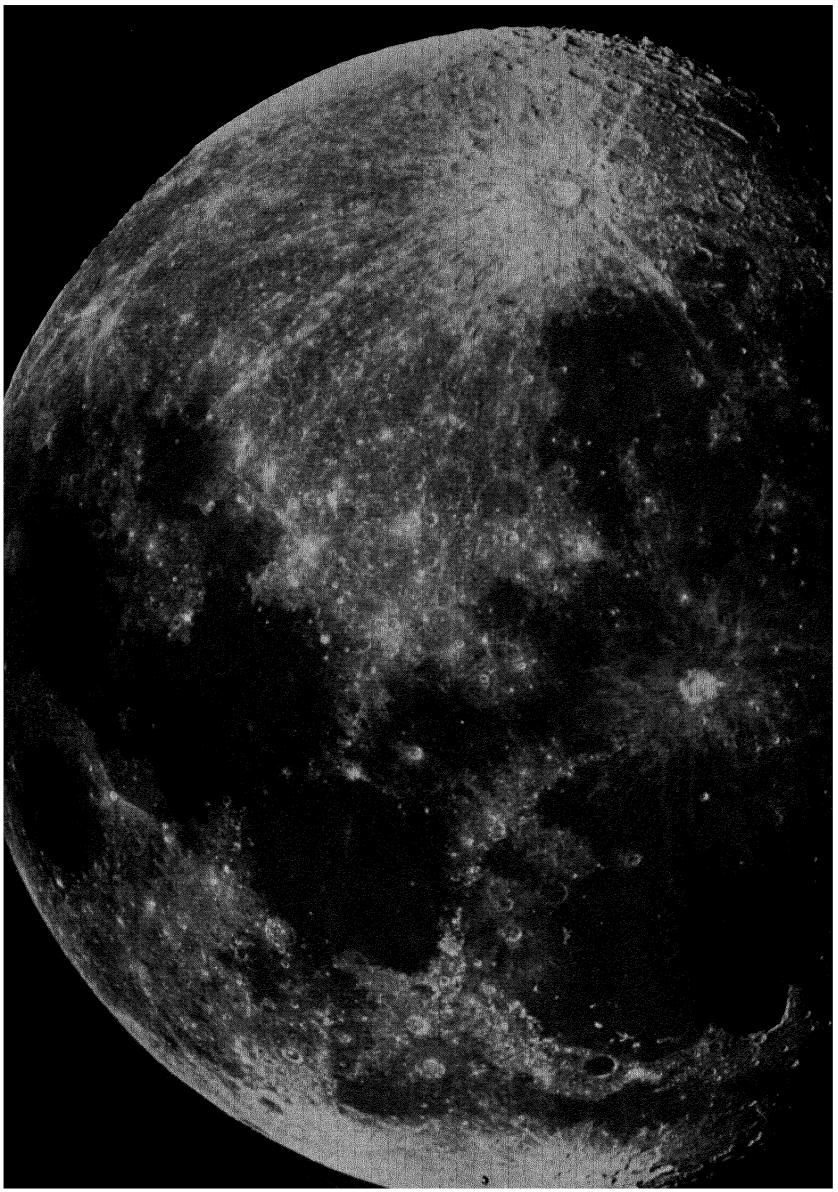


Figure 49. The twelve-and-a-half day moon, showing the two major ray systems, Tycho (upper) and Copernicus (right centre). (*By courtesy of the Director of the Paris Observatory.*)

crater ring or mountain range it is not deflected, but merely stains the whole formation with its whiteness. In only about one known instance does another formation cause any divagation of a bright streak.

The markings and formations of the lunar surface present two major problems, as yet virtually unsolved. The first concerns these bright streaks. What is their nature, and (when that has been answered) what was their origin? If the first question could be answered conclusively, the answer to the second would probably follow. But even their nature is veiled in mystery, chiefly because of the lack of any terrestrial analogy. Since the systems radiating from central craters bear a superficial resemblance to the splashes formed when a viscous object or a squirt of liquid is projected at a hard surface, or when a solid object is thrown at a viscid or liquid one, it has been suggested by Würdemann that the streaks may be the 'splashes' of meteors that impacted with the lunar surface when it was still in a semi-liquid state. Again, it has been suggested that they may mark cracks in the lunar surface which strike out radially from the central craters and through which some white substance at a past epoch oozed up from the interior. The drawback of these theories is that they cannot explain the observed fact that in no instance does a streak throw a shadow under a low sun—that they are, in fact, flush with the surface.

Perhaps the mosy ingenious and attractive theory is that developed by Stewart and Buell, and supported experimentally at Princeton Observatory. It has been known for a number of years that moonlight is polarized in a manner that can be most nearly matched among terrestrial materials by powdered pumice. If the surface of the moon does largely consist of this volcanic material, it is probable that it is in a broken and shattered condition, owing to the rapidly fluctuating extremes of temperature to which it is subjected. Stewart and Buell suggested that the bright rays consist of more finely powdered (and therefore more highly reflective) material which has been expelled from the central craters, either by volcanic action or as the result of meteoric impact. This fine dust would settle between the coarser grains of the surrounding lunar surface, and in this position would only be visible under a high sun; under oblique illumination it would be hidden from view in the multitudinous small shadows of the larger particles of which the surface consists.

The moon's past history

No instrument is more powerful in the hands of scientists than mathematics. Newton generalized the laws of Kepler, showing the

nature of the 'why' that underlay his 'how', and all subsequent work has gone to ratify the universality of the law of gravitation. By its means the behaviour of bodies in space can be calculated for almost any set of conditions. The physical constitution and behaviour of matter, investigated in the laboratory by means of experiments which the reader can find described in any textbook of physics, also allow of mathematical expression and investigation. Thus it was that G. H. Darwin was able to proceed from premises which are generally believed to correspond most closely with the past condition of the solar system to a mathematical demonstration that the moon was very probably 'born' of the earth.

Before describing the results of his investigation a word must be said of the currently accepted picture of the earth's origin; it is well to bear in mind, however, that this account of the birth of planets is entirely lacking in anything approaching proof. At some epoch in the past a star is said to have swung out of the depths of space, passed comparatively near to the sun and once more receded into the unknown. As the distance separating the two bodies—sun and star—grew steadily smaller their increasing gravitational attraction raised gigantic tides in their plastic bodies; for a similar reason the moon to-day raises tides in the earth's oceans. The solar tides increased in volume as the two bodies drew nearer to each other, until a point was reached when equilibrium could no longer be maintained, and a jet of solar material was drawn out towards the star. From this liquid or gaseous filament the planets later condensed.

It follows from this hypothesis that at some period in the past—the study of radioactive minerals in the earth's crust suggests some two thousand million years ago—the earth was a plastic body with a temperature of several thousand degrees; it also follows that its rotation period was very much shorter than it is to-day. It was from this stage in the earth's supposed early history that Darwin began his deductions. He proved that this body would contract and that the contraction would be accompanied by a steady acceleration of its rotation. Then came the crucial stage of the demonstration in which it was proved that if this acceleration increased beyond a point when the earth was rotating once in about three hours, then it would disrupt with the formation of two bodies of unequal size. At first these would be in contact and in rapid revolution about their common centre of gravity. But just as the moon now raises tides in the terrestrial oceans, and as the earth's two 'parents' are supposed to have raised tides in one another, so each of the two bodies would cause enormous tides in the semi-plastic body of the other. It was shown that the immediate effect of these tides would be to put a

brake upon their axial rotations: and since the angular momentum of the whole system must be conserved, this would result in the two bodies moving further apart. The moon would continue to recede from the earth until continued cooling resulted in the cessation of the tides and the assumption of their present relative positions. Observational confirmation of a single point in this hypothesis is possible. Darwin proved that not only would the tides affect the lunar orbit, but that they would also decelerate the moon's rotation until its axial rotation and its circumterrestrial revolution occupied the same period. As can be observed, this actually is so; the moon rotates on its axis in the same period as it revolves, and therefore always turns the same hemisphere to the earth. Furthermore, the oceanic tides must have a similar retarding effect on the earth's rotation, and though the effect is slight it is large enough for detection: the day is steadily but very slowly lengthening.

Thus we have sound cause to believe—it is generally regarded as proven—that at some past epoch the moon was a plastic body at a very much higher temperature, and a very much smaller distance from the earth, than at present. These facts have been made use of in one of the many attempts that have been made to solve the second great problem propounded by the moon's surface features: that of the origin of the maria and crater rings.

Origin of lunar 'craters': meteoric theory

Only two of the more interesting suggested explanations of the existence of lunar craters will be discussed here; even so, the treatment must be unsatisfactorily brief. The first supposes that the moon was once subjected to a prolonged meteoric bombardment. At this epoch it is envisaged as having lost whatever atmosphere it may once have possessed, but as having developed, through cooling, a solid crust. At each point of impact between a meteoric fragment and the lunar surface an intense temperature would have been instantaneously generated by the translation of the meteor's kinetic energy into heat energy. The resultant explosion, given a meteor of considerable mass, would have made the atom bomb look like a cap-pistol. Assuming meteors of dimensions comparable with the asteroids,[1] there is no difficulty in envisaging the formation of even the greatest of the walled plains by this means.

Other points in the theory's favour are the satisfactory explanation it provides of the astonishing circularity of formations often 150 miles in diameter, and its provision of a mechanism for the expulsion

[1] See p. 190.

of jets of pulverized material called for by recent suggested explanations of the bright rays.

Two of the commoner objections that have been levelled against the meteoric theory are easily disposed of: that the earth also should be pockmarked with craters, and that since not all of the impacts would be in a vertical plane a large proportion of the craters should be elliptical. The former neglects the fact that the earth's atmosphere provides a protective screen which the moon lacks, and also that erosion and sedimentation are terrestrial smoothing agencies not to be found on the moon. The latter makes the false assumption that an oblique impact must cause an elliptical crater: if the impact were explosive, as it would be, the crater would always be circular. This fact is well known to readers who have had experience of aerial or artillery bombardment.

It is, however, a valid criticism of the meteoric theory that it cannot account for the maria, and it seems probable that these do represent areas of subsidence such as are envisaged in the seismic theory.

Origin of lunar 'craters': seismic theory

The seismic or tidal theory proceeds from entirely different premises. According to this hypothesis, the first age of crater formation occurred when the moon had a thin solid crust over a still molten interior which was scoured by gigantic tides raised by the comparatively near earth. The progressive cooling of the moon would result in the contraction of this crust, strain being set up in it, and the final collapse of points which were either weaker or thinner than the surrounding surface. Each time the internal tide swept past such a rupture the molten material would be forced up through it, enlarging it and leaving a cooling deposit round its edge. Since each vent-hole would be continuously growing, the older craters would be larger than the more recent. Finally, the crust, cooling gradually throughout this period of crater formation, would grow sufficiently thick to prevent the periodic uprushes. The maria are represented as areas of subsidence and secondary crater formation at a later epoch. The fact that they are darker than the areas of initial crater formation is explained by supposing that the less dense materials which would float to the molten surface were of a lighter hue than the denser substrata. The subsidence theory of the maria receives support from the observed fact that they contain many reduced and partially melted ring plains of a light colour, relics of the primary stage which escaped complete destruction at the time of the formation of the maria.

Against the seismic theory it may well be urged that such geometrically circular formations could not invariably have resulted from the 'eating away' of the periphery of the original vent or subsidence.

It seems reasonably safe to believe that for the true explanation of the origin of the craters and maria we must look to one of these two theories, or more probably to some modification or combination of them. On the other hand, it would be invidious in a book of this nature to support one against the other: all that can be said is that the case for neither is yet proved, and that the debate continues.

The lunar atmosphere

Perhaps the most striking feature of the lunar landscape, as seen telescopically, is its unrelieved starkness. Shadows are dead black, and land and surface markings hidden from the direct rays of the sun by the terminator are completely invisible. The definition is as sharp as that of a steel engraving, and half tones, grey shadows, blurred details and soft distances are conspicuously absent. This alone would lead one to suspect that the moon has no atmosphere, for these are characteristically atmospheric effects. Still more important, objects near the limb are seen with a stark clarity which is in no way inferior to that of objects near the centre of the disc. Such a dimming would inevitably result from the existence of any but an extremely rare atmosphere.

General appearances, then, lead one to suspect that the moon has no atmosphere. But this is not enough, and more specific observations can be found to confirm this conclusion. Several times a month, on the average, the moon passes in front of a tolerably bright star. Because the moon's motion is eastward the star disappears at the eastern limb and reappears at the western. If the moon had no atmosphere, the disappearance (and the reappearance) of the star would be instantaneous—at one moment the star would appear to be perched on the moon's limb and the next it would be gone. But when light passes from one medium to another it is bent or refracted: everyone is familiar with the effect that a stick piercing the surface of a sheet of water is bent. If the moon had an appreciable atmosphere, therefore, the star's light would be similarly refracted in passing through it, with the result that it would appear to slow down as it approached the limb. Now this effect has never been observed, the disappearance and reappearance of occulted stars always being instantaneous and their motion uniform. Hence it is concluded that the moon has no *appreciable* atmosphere.

The observation of the occulted star cannot tell us that the moon

has *no* atmosphere: on this point it can give us no information. All we are justified in concluding is that if there is an atmosphere at all, it is below a certain threshold density at which the occultation phenomenon referred to would be sufficiently marked to be noticeable with our most accurate instruments. This threshold is 10^{-4} of the terrestrial; the atmospheric pressure at the surface of the moon cannot therefore be greater than one ten-thousandth of that at the surface of the earth, or 0·076 mm. of mercury. Rare though such an atmosphere would be, it would not be non-existent.

The second pointer which indicates that the moon has not a dense atmosphere is the absence of lunar twilight. On earth darkness does not fall the instant that the sun is obscured by the horizon; higher levels of the atmosphere are still lit by the direct light of the sun for an hour or more after sunset, and the molecules of this atmosphere reflect and scatter the light, thus preserving the indirect illumination known as twilight. An observer looking at the earth from outer space would not see a sharply defined terminator separating night from day, such as we see on the moon, but a zone of appreciable width which is neither day nor night, but an insensible shading into each. It has been estimated that a lunar atmosphere with a density exceeding 10^{-5} that of the earth would cause an observable twilight effect in the cusps.

The spectroscope clinches the matter. Since the moon shines by reflected sunlight, its spectrum is a faint replica of the solar spectrum. But should the moon possess an atmosphere, it would be betrayed spectroscopically by the presence of absorption bands, provided its density is sufficiently high. For the solar light when reflected from the moon's surface would have to pass twice through its atmosphere, once before reflection and once after. Yet the lunar spectrum is identical with that of the sun, and no lines that might be attributed to lunar atmospheric absorption can be detected. There are non-solar lines, it is true, but these have all been established as telluric, i.e. caused by absorption in the earth's atmosphere. The lunar atmosphere, then, if it exists at all, does not exceed one ten-thousandth of our own in density.

The lunar temperature

From the point of view of life, the next most important environmental factor is temperature. The earth's atmosphere contains considerable quantities of water vapour, and acts as a blanket wrapped round the day-warmed hemisphere which retards the radiation of its heat into the night sky; it is well known that a clear, starlit night is colder than one that is overcast and cloudy. But whether or not

the moon does possess an atmosphere 10^{-4} or 10^{-5} as dense as our own, an atmosphere of lower density than this would certainly be incapable of performing this blanketing function. Hence the lunar surface must not only get very much hotter than the earth's during the day, but also become very much colder during the night. Still more so, since the lunar days and nights are equal in length to fourteen terrestrial ones.

The measurement of the moon's temperature would be impossible were it necessary to do so directly. The measurement of very minute units of heat presents considerable practical difficulties. It is fortunate, therefore, that heat can be converted into electricity, for minute electric currents can be detected and measured more easily than minute units of heat. The galvanometer (an instrument that performs this operation) is a very much more delicate and sensitive piece of apparatus than the thermometer. If two pieces of metal —one, say, of iron and the other of copper—are soldered together and then heated, an electric current will flow between them. It has been discovered that the metals which give the strongest current for a given temperature are antimony and bismuth. A number of these antimony-bismuth units, known as thermocouples, are joined in series and connected with a highly sensitive galvanometer. Then the light and other radiation from the moon is allowed to pass through a telescope of large aperture and is focused on the bolometer; the galvanometer is read, and a simple calculation (based on data derived from laboratory experiment) permits the temperature of the source to be deduced. It is found that the temperature of the night side of the moon is very low indeed, and different workers are almost unanimous in their estimate of a temperature near to absolute zero ($-273°$ C.). This is the absolute cold, the complete lack of heat, which characterises interstellar space.

Unfortunately, estimates of the temperature of the sunlit side that have been made by different workers and at different times do not agree so closely. But the most reliable recent work on the subject makes it probable that the temperature at the centre of the sunlit hemisphere is some 30° above the boiling point of water (130° C.); it falls off rapidly during the long afternoon, freezing point being reached shortly before sunset. During even the few hours that the sun is obscured at lunar eclipse, the temperature may fall to the neighbourhood of $-120°$ C.

Thus the moon, judged by terrestrial and human standards, is not a hospitable place. It has an atmosphere which is at best almost non-existent, and its temperature fluctuates between approximately 270° of frost and that of superheated steam. It is hardly surprising,

therefore, that the moon has generally been regarded as a dead world. Within the last fifty years, however, some doubt has been cast upon this easy assumption, notably by W. H. Pickering. Even before this time it had been known that the face of the moon is not entirely changeless, though such minor changes as had been noticed were consistent with a 'topographical' explanation—that is to say, were probably in the nature of landslides, subsidences and the like.

Topographical change on the lunar surface

The most famous case of such lunar change is that which occurred to the crater Linné. This is a small, wall-less crater situated on the flat expanse of Mare Serenitatis. It first figures in the map of Riccioli, constructed in 1651, where it appears as a deep crater some $4\frac{1}{2}$ miles in diameter. At the end of the eighteenth century Schröter described it as a doubtful crater consisting of a small, brilliantly white spot. Not only do these two accounts conflict with one another, but had the crater been no larger during the seventeenth and eighteenth centuries than it is to-day neither Riccioli nor Schröter would have been able to detect it with their relatively primitive instruments. In 1823 Löhrmann described it in terms reminiscent of Riccioli, and in 1831 Mädler, as the result of several observations, stated that it was a deep, bright and distinct crater 6 miles in diameter. Twelve years later Schmidt confirmed this general description after having observed the crater carefully on eight separate occasions; he estimated its depth to be at least 1,000 feet and its diameter something between five and a half and seven miles. Up to this time, then, different observers' estimations of the size of Linnê had varied widely, but, if we ignore Schröter's observation, no change of structure had been definitely established.

In October 1866 Schmidt re-observed Linnê and noticed that its appearance had altered strikingly since he had last observed it twenty-three years previously: where before there had been a deep and distinct crater there was now a featureless white patch. In the following month he announced that the crater had disappeared. After several more months had elapsed he announced that a minute crater-pit, not more than a quarter of a mile in diameter, was just visible on the white patch that had replaced the vanished Linné. By 1868 this crater had increased in diameter to about one and a half miles. To-day Linné appears to be a small crater about one mile in diameter.

It may be regarded as established that Linnê is now smaller than it was during the period 1651–1866. Its disappearance and what looked like its replacement by a white cloud in the latter year may possibly

be explained as the subsidence and collapse of the original deep crater, though the subsequent formation and gradual growth of the present crater is puzzling.

Other types of lunar surface change

The floor of the crater Plato has also furnished some well-attested examples of lunar surface change. This crater floor is of a dark tint, smooth and featureless but for a number of minute crater-pits interconnected by a system of faint rays. Independent observers at different times have reported the disappearance of some of these pits, and it is safe to say that their visibility is variable. In addition to these unpredictable disappearances, the floor of Plato itself is subject to a periodic variation of brightness, the depth of tint being dependent on the time of lunar day (terrestrial month). This is impossible to explain by recourse to any 'topographical' theory of lunar change, and leads us on to the work of Pickering and his followers which within the last half century has strongly suggested—although the suggestion has not been universally or even widely accepted—that some primitive form of vegetable life may still carry on a precarious existence on the moon's surface.

The partly bright areas

The first discovery produced by Pickering's intensive study of the moon was that of the new type of surface feature which he named the partly bright areas. These are small, light coloured areas which most usually occur on broken ground—crater floors and ramparts, and the upper slopes of mountain ranges are common sites. What gives these partly bright areas their outstanding interest is the fact that their brightness varies with the lunar day. Some twenty-four (terrestrial) hours after sunrise they begin to fade and decrease in size, and by the middle of the lunar afternoon they may have disappeared completely. Presumably they increase in size again during the long lunar night, for when they are next visible soon after the following sunrise their size and intensity is once more maximal. The peculiar nature of this variation—the size and brightness of the area being inversely proportional to the sun's altitude—suggested to Pickering that they might possibly be deposits of snow or, more probably, hoarfrost. As we have seen, other considerations indicate that the moon has long been without any trace of water; nevertheless, no more satisfactory explanation of these partly bright areas has yet been proposed.

The variable spots

The next feature to be discovered—also by Pickering—which hinted that the picture of a dead and waterless moon might need some

revision was the variable spots. These spots first appear some hours after sunrise and increase in size and depth of colour to a maximum some hours before sunset, when a rapid change in the opposite direction sets in; by the time the sun sets, and the terminator reaches the spot, its colour is normal once again, and as a spot it no longer exists. Variable spots are not found within about 30° of the poles, and the nearer a spot is to the equator the more rapid is its cycle of changes. What are these spots, and what is the significance of their monthly changes? Pickering asked whether they might not be patches of some kind of lunar vegetation. This explanation, though once again at variance with earlier ideas concerning the nature of the moon's surface and of the conditions prevailing upon it, gives a satisfactory account of the phenomena that it sets out to explain.

But the revolutionary implications and disregard of theoretical considerations which characterize Pickering's theories have resulted in their somewhat unenthusiastic reception by the astronomical fraternity.

Mechanism of eclipses

The moon, in revolving about the earth, traces out a great circle upon the star sphere. Similarly the sun describes in the course of a year the great circle known as the ecliptic. The two points where these circles intersect are called nodes: the ascending node where the moon crosses from the south to the north of the ecliptic, and the descending node where it passes back again from north to south. When the sun and moon are at opposite nodes simultaneously, the earth will lie directly between them, since the nodes are 180° apart; in this case a lunar eclipse will occur. Hence it is that eclipses of the moon can only occur at or near full moon. Thus there are two important differences between lunar eclipses and those of the sun: the former always occur at full moon, and it is the earth's shadow on the moon's surface that is observed; the latter occur at new moon, and this time it is the moon's shadow which crosses the earth's surface.

To say that the moon is in eclipse, therefore, means that it is passing through the earth's shadow. This circular shadow, as it sweeps across the moon's surface, is seen to consist of a dark central core, or umbra, surrounded by a less dense fringe called the penumbra. The formation of these two zones in the shadow is explained in Fig. 50, which shows a spherical body casting a shadow upon a flat surface; the illuminator is likewise a spherical body and not a point source. Actual experiment will demonstrate that the outer edge of the penumbra and the junction of the umbra and penumbra are not

sharply defined, as in the diagram, but are hazy and indeterminate. A little thought, and the inspection of the figure, will show why the two shadow zones are formed. An observer stationed in the umbra *AB* will discover that the illuminating body is completely blotted from his vision by the eclipsing body; no light from the source can penetrate the cone *ABba*. An observer under the penumbra *CD* will not see a totally eclipsed illuminator, but a partially eclipsed one. The further he is from the umbra—i.e. the nearer to the outer edge of the penumbra—the smaller will be the fraction of the illuminating body that is observed to be eclipsed, and the lighter will be the tone of the shadow upon the screen.

Now suppose that the body *X* is in motion, so that its shadow moves across the screen. The umbra will trace out some such zone

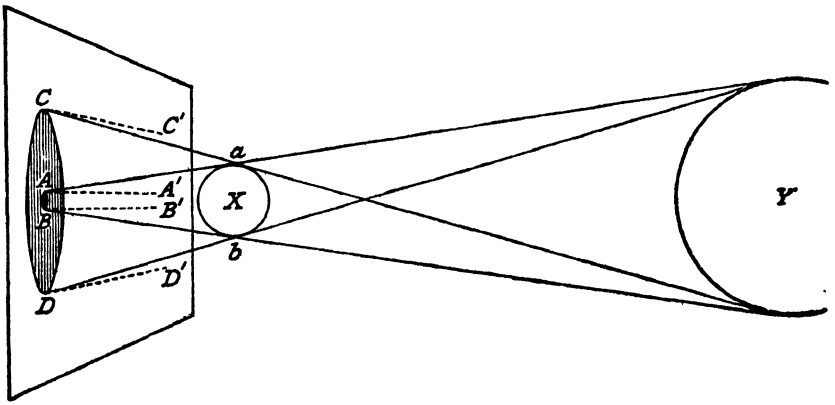


Figure 50.

as *AA'BB'*, and the penumbra the zone *CC'DD'*. We have now reproduced the mechanism of eclipses. In the case of a lunar eclipse, *Y* is the sun, *X* the earth, and the screen the moon's surface (here represented, for simplicity, as a flat instead of a curved surface.)

If the sun and moon reach, not opposite, but the same node simultaneously, then a central solar eclipse will occur. The fact that there is not a total solar eclipse at each new moon proves conclusively, without it being necessary to plot the paths of the sun and moon on a star map, that these two paths do not lie in the same plane. That is, that the planes containing the orbit of the moon and the apparent orbit of the sun (the terrestrial appearance of the earth's circumsolar orbit) are inclined to one another; since they both pass through the centre of the earth, they must nevertheless intersect. The necessary result of this inclination of the ecliptic to the moon's path is that when either the sun or moon are not near a node there cannot be an eclipse.

Suppose that the sun and moon reach a node, not simultaneously, but in close succession. The body of the moon will then not completely cover that of the sun, and the result will be a partial solar eclipse (see Fig. 51). The further the sun is from the node, the smaller will be the fraction of its disc which is obscured. Beyond a certain distance (which is easily calculable) there will, of course, be no eclipse at all.

At the same time, it must be noted that the simultaneous arrival of both bodies at the same node does not inevitably result in a total solar eclipse, though the eclipse must under these conditions always be central. For we have seen that planetary orbits are not circles but ellipses, one of whose foci is occupied by the centre of the sun. Hence the distance of the moon from the earth and of the earth from the sun are subject to small variations which cause a variation in the apparent sizes of any two of the bodies concerned as seen from the third. Thus

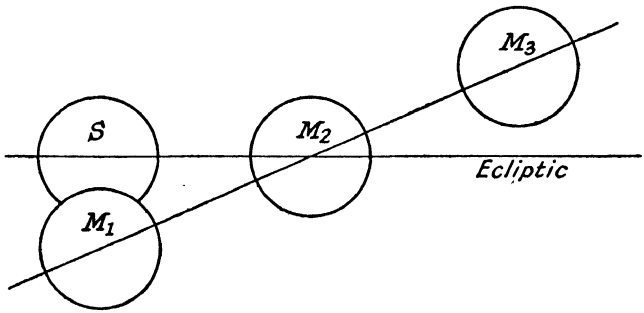


Figure 51. Formation of a partial solar eclipse.

it is possible for the earth and the moon to assume positions in their respective orbits such that the angular diameter of the moon is slightly smaller than that of the sun. Under these circumstances, and when the centres of the sun and the moon are in the same line of sight, the sun will not be totally obscured: a narrow rim of photosphere will show round the moon's limb. Such eclipses are called annular, from the Latin *annulus*, a ring.

The distance and sizes of the moon and sun are related in such a way that the shadow of the former on the earth's surface cannot exceed 167 miles in diameter; usually it is much less than this. Thus as the moon moves across the face of the sun its shadow traces out a comparatively narrow band across the earth's surface. An observer stationed more than about eighty-four miles from the centre of the totality zone (assuming that its size is maximal) will see a narrow crescent of sun round the moon's limb at the moment of the nearest approach to totality; he will, in fact, observe a partial eclipse. The further away he moves, the larger will be the observed solar crescent.

If he is situated more than about 2,000 miles from the centre of the totality zone, he will observe no eclipse at all. These zones of course correspond exactly with the umbra and penumbra of lunar eclipses. The totality zone is the umbra of the moon's shadow; the zone from which the eclipse is observed as partial, is the penumbra. *AA'B'B*, in Fig. 50, is the totality zone, and the wider area covered by the penumbra (still a band-shaped zone, longer than it is wide) is the 4,000-odd mile wide zone of the earth's surface from which the sun appears to be partially eclipsed, and which therefore receives some light direct from the sun.

Lunar eclipse phenomena

When the sun and moon reach their respective nodes at the same time, the moon will be totally eclipsed, i.e. the whole disc will lie under the umbra. If they do not, the earth's shadow will sweep across a section of the lunar disc only. Under these circumstances, nothing but the penumbra may fall on the moon, but at total eclipses the shadow can be clearly seen to consist of a central dark umbra and a peripheral, lighter umbra.

Since, in eclipses of the moon, it is the shadow upon the observed object that is seen, and not the eclipsing body itself as in the case of solar eclipses, a lunar eclipse is visible over the whole terrestrial hemisphere from which the moon is visible. We have seen that the lunar territory under the umbra receives no direct light from the sun. It might therefore be expected that it would be invisible. Actually, however, the totally eclipsed moon is only reduced in brightness, and usually rendered a deep red colour: it is still visible, and—on account of its colour—very strikingly so. The reason for this visibility is that the earth's atmosphere refracts the sun's rays, bending them into the shadow cone, just as a stick held partly under water appears to be bent. Thus all the light that illuminates the totally eclipsed moon has passed through the earth's atmosphere, and for this reason terrestrial meteorological conditions (especially the incidence of clouds) affect the appearance of a lunar eclipse. The redness of the refracted light is due to the preferential action of the atmosphere upon the radiation of mixed wavelengths of which sunlight is composed. The blue light is scattered, while the atmosphere has little effect on the red, which it transmits almost in its entirety. It is for the same reason, as has already been pointed out, that the day sky is blue, and that sunsets (when the sun's light has to pass through a maximum thickness of atmosphere) are red.

The planetary family

The order of the planets from the sun outwards is Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, and (most distant and most recently discovered of all) Pluto. We saw in Chapter III that the distances of the sun and moon can be discovered by the parallactic method, either directly or indirectly. Further, that the distances of the planets may be similarly determined; and, since the distance of the earth from the sun is known, we can arrive at the solar distances of those planets too remote for direct investigation, by the application of Kepler's third law. By a close observation of the motions and successive positions of each planet, it is also possible, with the assistance of the laws enunciated by Kepler and Newton, to reconstruct its complete orbit: to discover its eccentricity, the lengths of the major and minor axes (the planet's greatest and least distances from the sun), the period of revolution, and the velocity of the planet in any part of its orbit.

Mercury: solar distance

In this way it has been established that the mean distance of Mercury from the sun is 36 million miles, or about one-third of the earth's. But the orbit is markedly eccentric—more so than that of any other major planet except Pluto—with the result that the solar distance varies considerably with different positions of the radius vector: it may be as great as 43 million miles or as little as 28 million. Thus Mercury's distance from the sun varies by about 42 per cent. of its mean value, as compared with the earth's 3 per cent. This alone will cause appreciable temperature differences at the Mercurian surface at different times during its year.

Mercury: observation

But even 43 million miles is a relatively small distance, and it is this nearness of Mercury to the sun that causes the apparent proximity of the two bodies in the sky; we have seen that Mercury can never be further than about 28° from the sun. This restriction has important practical effects, for it means that whenever Mercury is visible to the naked eye it is near the horizon—certainly not more than about 20° from it—either the eastern horizon a short time before sunrise or the western horizon just after sunset. Now the least favourable way of observing a celestial object is to view it when it is near the horizon, for it is then being seen through a maximum thickness of atmosphere (see Fig. 52). The effect of the atmosphere upon telescopic 'seeing' is comparable with that of a thick, cloudy plate-glass window of uneven thickness and density, and studded with flaws, interposed between

the observer and the object observed; a window, moreover, which is in constant and erratic motion. Fortunately, Mercury is a bright object, and it is this fact that allows the difficulty to be overcome. For when its position in the sky is known—as, through the application of Kepler's laws, it is known, and recorded in almanacs—it can be

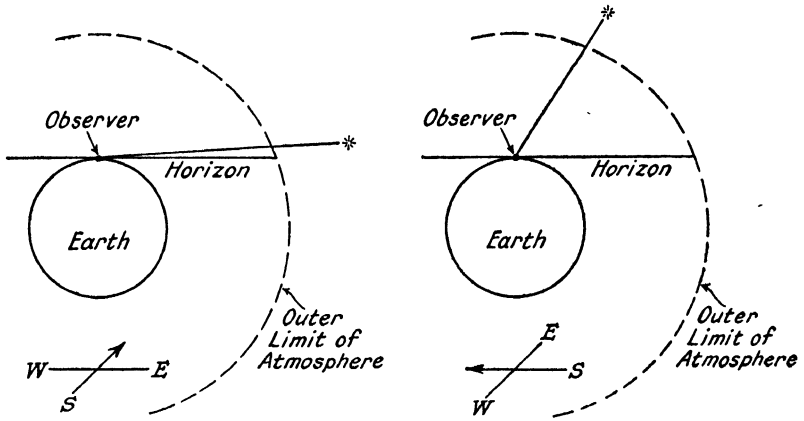


Figure 52. To observe any celestial object at its culmination is to observe it through a minimal thickness of terrestrial atmosphere.

picked up telescopically during the daytime. Its altitude above the horizon is then greater than in the evening or early morning, and the effects of the atmosphere upon the 'seeing' are therefore lessened.

Those who aspire to see Mercury with the naked eye will do well to note the following facts. Elongations, when Mercury and the sun are, as seen from the earth, furthest apart, and when the interval between the rising (and setting) of the two bodies is greatest, are most favourable in the north hemisphere when they occur in spring or autumn, for then the ecliptic makes its largest angle with the horizon. Maximum elongations east in March or April (evening observation), or maximum elongation west when it occurs in September or October (morning observation), consequently provide the best conditions. Telescopic viewing will of course show that it is dichotomized at these times.

Mercury: axial rotation

Mercury's sidereal period is easily determined by inspection, allowance being made for the observer's motion, and is found to be a little less than eighty-eight days: this is the planet's 'year', or period of circumsolar revolution. But its period of axial rotation is less easily determined, and is, indeed, not yet beyond doubt. The obvious directions to an observer wishing to determine a planet's

rotation period are as follows. Observe the transit of some conspicuous surface marking across the planet's central meridian, note the time, observe the next transit, again note the time, and then subtract the first time from the second. But, unfortunately, Mercury exhibits no sharply defined markings that can be used for this purpose. Such markings as it has are so vague in outline that it requires telescopes of considerable aperture, employed when the planet is high above the horizon, to show them at all; even so, their ill-defined nature renders them quite unsuitable for the determination of the rotation period. Schröter believed that Mercury rotates on its axis in about twenty-four hours, and this view was generally accepted until 1889, when Schiaparelli suggested the much longer period of eighty-eight days, that in which Mercury completes one revolution of the sun. If this is so, Mercury always turns the same hemisphere towards the sun, just as the moon always turns the same face inwards to the earth. This longer period is consonant with the negative results of the attempt made at Mt. Wilson to measure with a thermocouple the radiation from the dark side of Mercury, which also suggest that one hemisphere never receives direct heat from the sun.

How difficult it is to estimate the rotation period from observations of surface markings may be appreciated from the fact that the angular diameter of Mercury at elongation is the same as that of a halfpenny seen from a distance of about half a mile. This difficulty is aggravated by the persistent lack of sharply defined markings of any kind, the characteristic features being vague areas of a slightly darker tone than the general background of the planet's disc. Other observations which have been described at various times are a blunting of the cusps when the planet is in the crescent phase, irregularities in the outline of the terminator and certain minute projections over the limb.

Mercury: temperature

Were it possible to isolate the problem of the rotation period from all others connected with the planet, this hiatus in our knowledge would not be of any great importance. But this cannot be done, for the problem is intimately bound up with the further problem of the temperature at the planet's surface. If one hemisphere is always turned towards the sun and the other never receives any solar light and heat whatever, the temperature difference between the two will be considerable. But if, on the other hand, Mercury rotates rapidly upon its axis, the difference of temperature between the sunlit side and the dark will be very much smaller. Radiometric measurements

undertaken at Mt. Wilson Observatory indicate a mean temperature of about 340° C. for the centre of the sunlit side. Owing to the great variations in Mercury's distance from the sun, to which attention has already been drawn, this will vary between a minimum of about 280° to a maximum in the neighbourhood of 410°. As one travels from the sub-solar point[1] towards the edge of the illuminated hemisphere the altitude of the sun will decrease (since Mercury, like the earth, is a spherical body) and the temperature will drop. The temperature of the dark side seems to be very low indeed, and no radiation from it of measurable intensity can be detected with our present instruments. The extremely high temperature of the centre of the sunlit side of Mercury (about equal to that of molten zinc and nearly twice that of molten tin) is of course due to the planet's relatively small solar distance, while the great difference between the temperatures of the sunlit and the dark hemispheres confirms the longer rotation period.

Mercury: atmosphere

Though the lack of definite knowledge regarding the rotation of Mercury must also render our knowledge of its temperature regrettably inexact, one thing is certain. On account of its proximity to the sun and its high temperature (whatever the precise figure may be), Mercury cannot possibly have retained such atmosphere as it may once have possessed. The factor controlling the retention or loss of a planet's atmosphere is the difference between the velocity of its constituent molecules and the velocity at which any body could escape permanently from the planet's gravitational grasp. The first of these velocities is determined by the temperature of the atmosphere (since the hotter a gas is, the greater is the velocity of its component particles), and by the mass of the molecules concerned. The second depends solely upon the planet's mass (since a greater mass involves more powerful gravitational attraction upon the atmospheric particles, and correspondingly higher velocities are required in order to escape from that attraction), and its radius (since the atmosphere is further away from the centre of mass of a large planet than that of a smaller one).

We have already discovered that whatever the exact surface temperature of Mercury may be, it is certainly very high; hence the molecules of such atmosphere as it may possess will be travelling at considerably higher velocities than, for instance, those of the earth's atmosphere. Escape may therefore be expected unless the planet's gravitational pull is correspondingly greater than the earth's.

[1] That is, the point on the surface at which the sun is directly overhead.

But it has been found that this is not so, and the raising of the first threshold together with the lowering of the second make it quite certain that Mercury cannot have retained an atmosphere. As in the case of the moon, this conclusion is verified by the spectroscope, for the spectrum of Mercury bears no trace of absorptions other than telluric.

Mercury: linear dimensions

That Mercury's gravitational pull is less than that of the earth may be suspected as soon as its linear size is determined. This can easily be done—all that is required is the measurement of its angular diameter. This is found to vary from 5″ to 13″ according to the relative positions of Mercury and the earth in their respective orbits. Since the distance separating the two bodies is known, one has only to calculate what linear diameter would be required to give the observed apparent diameter at the distance in question. This linear diameter has been found to be about 3,000 miles. Mercury is thus a much smaller body than the earth, only about half as big again as the moon. The gravitational pull of a body is not, however, dependent upon its size but upon its mass, and though we may truthfully say that the value of g, the gravitational constant, at the surface of Mercury will be less than that of the earth providing the two bodies are composed of the same materials and have the same density, yet we cannot be certain until we have removed that provision by determining its mass.

Mercury: mass

The most accurate method of determining a planet's mass is by a study of the motions and orbits of its satellites and the application of the laws of motion. But Mercury has no satellite, and we are thrown back upon the less accurate and more tedious method of perturbations. As Newton discovered, any two bodies attract one another with a force that is dependent upon certain properties and relations of those bodies. Hence Mercury's gravitational pull on Venus, for instance, will affect or perturb Venus' orbital motion. A careful study of the nature and degree of this perturbation allows the mass of Mercury to be calculated by the application of Newton's laws, though the comparative proximity of both planets to the very much more massive sun largely deadens the perturbations, and the result is correspondingly inaccurate. As would be expected from its small size, however, the mass of Mercury is considerably less than that of the earth—only about 4 per cent.

Mercury: transit phenomena

Several other observations that confirm the conclusion that Mercury has no atmosphere must be noticed in passing. When one of the inner planets approaches the limb of the sun it is being lit from 'behind' by an extremely brilliant light. This light must pass through any atmosphere that the planet may possess, thus rendering it visible as a bright halo. Venus, which is known to possess a dense atmosphere, exhibits such a halo when passing into and out of transit, but nothing of the sort is to be seen surrounding the limb of Mercury under similar circumstances.

Mercury: albedo

The planet's albedo suggests the same conclusion. When a beam of light is directed upon an object a certain proportion only is reflected. Different types of surface reflect the incident light in different degrees, but no terrestrial object or substance is more highly reflective than newly fallen snow, which returns about 75 per cent. of the incident light. This reflected fraction or percentage of the incident light is known as the albedo of the surface; the albedo of snow, for instance, is 75 per cent. or 0·75. The albedo of clouds is also high—that of some terrestrial clouds is about 70 per cent. In the same way, those planets known to have atmospheres have high albedos; those of Venus and Jupiter are about 60 per cent. and 50 per cent. respectively. Not only do those bodies with dense atmospheres have high albedos, but the moon, with no appreciable atmosphere, has a low one—about 7 per cent. only. And measurements of the albedo of Mercury show that it also is 7 per cent.

Thus in many respects Mercury and the moon may be regarded as similar bodies. Both are considerably smaller than the earth; both are without appreciable atmospheres; they are consequently subject to greater extremes of temperature than the earth; and their identical albedos would suggest that the nature of their surfaces is similar or identical.

Venus: solar distance, size and mass

No difficulty is to be encountered in finding Venus, provided that it is not too near conjunction. It is more conspicuous than Mercury because it is larger and nearer to the earth; it is also brighter both because its albedo is higher and because at elongations it is further from the sun and therefore shines from a darker sky.

We have seen how Mercury's distance from the earth and from the sun, its linear size and its mass may be determined; the methods of determination are the same in the case of Venus, and need not be

again described. Suffice it to say that in many respects Venus is more like the earth, and less like the moon, than Mercury. Its diameter of about 7,600 miles is only slightly less than that of the earth, and its mass is about four-fifths of the earth's. Hence the value of g at the surface of Venus is only about 20 per cent. less than that at the terrestrial surface. Provided, then, that the temperatures of the two bodies are not of a very different order, Venus may be expected to have an atmosphere.

Venus: temperature and atmosphere

The questions of Venus's temperature and rotation period have received less certain answers than those of Mercury's. Its orbit lies within the earth's, though further from the sun than that of Mercury, its mean solar distance being 67 million miles. The orbit is almost circular—less eccentric, in fact, than that of any other planet—and this value does not vary by more than 1 million miles throughout the course of the planet's year. It might be expected, therefore, that its temperature is somewhat higher than that of the earth though considerably lower than that of Mercury. Calculation, based upon its solar distance, would suggest a temperature of over 100° C., yet direct measurements record the very much lower figure of about −25° C. for both sunlit and dark sides. It is certain, however, that the relation between its temperature and its mass is such that it could have retained the molecules of all but perhaps the lightest gases.

Every line of independent research substantiates this result. Venus's albedo is high: between 60 per cent. and 65 per cent. Although this is a little lower than the albedo of the brightest terrestrial clouds, it is higher than that of any known surface of rock or soil, and would suggest most strongly that we are not looking down on the solid surface of the planet but upon the upper layers of a dense vaporous atmosphere. That this atmosphere is both dense and deep is shown by two further observations. When describing the appearance of the totally eclipsed moon we found that a gaseous envelope is penetrated more deeply by red light than by blue; that is, by radiations of long wavelengths than radiations of short. Yet photographs of Venus taken in infra-red light, the wavelength of which is even longer than that of visible red light, resemble closely the naked-eye appearance of the planet: they show a perfectly uniform and featureless surface. We must therefore conclude that even the infra-red radiations do not penetrate the fog-blanket to the depth of the solid surface, and that the infra-red photographs merely depict a lower atmosphere level than that visible to the naked eye.

That it is not only dense but deep may be discovered from transit

observations. The atmospheric halo that surrounds Venus when just outside the sun's limb has already been alluded to. Since the distances from the earth to all points on Venus's orbit are known, the measurement of the angular thickness of this halo will give us its linear thickness. In this way it has been established that the atmosphere of Venus cannot be less than fifty to seventy miles deep.

Venus: spectroscopic evidence

All these telescopic indications of the existence of a dense atmosphere are confirmed by the spectroscope. Even if the atmosphere is so dense that no light can penetrate as deep as the planet's surface, some of it will penetrate for a small distance, at least, before being scattered and reflected; absorption bands will accordingly be present in the spectrum.

A faint band in the infra-red region has been identified as that of carbon dioxide. Curiously enough, the absorptions of water are not found in the spectrum of Venus, despite the impenetrability of its atmosphere; neither are those of oxygen. It does not follow, however, that these gases do not occur in the atmosphere of Venus, but only that they do not occur in measurable quantities (about one thousandth of the concentration in the earth's atmosphere, in the case of oxygen) in the upper atmospheric regions which constitute the planet's visible surface. Sunlight only penetrates a comparatively short way into the fog blanket, and what the constitution of the lower levels may be we have no means of knowing. This being so, it is possible that light is thrown on the apparent absence of water in the atmosphere of Venus by the fact that nearly all the water vapour in the terrestrial atmosphere occurs within seven miles of the earth's surface—that is to say, in the lowest atmospheric levels.

Venus: axial rotation

The fact that no surface markings are visible renders the estimation of the rotation period more difficult than that of Mercury. It is true that Venus's disc does bear some markings, but these are not of a type suited to the purpose of recording the period of the planet's rotation. They are vague, greyish areas, particularly noticeable near the terminator when Venus is in crescent phase, and are certainly atmospheric. Nevertheless, the older observers concluded that Venus rotates upon its axis in about the same period as the earth. Schiaparelli, however, favoured a much longer period, and suggested 225 terrestrial days as the correct period. If this is so, the temperature difference between the sunlit and dark hemispheres must be considerable; yet bolometric measurements, although

indicating a lower temperature than might be expected, show only a negligible difference of temperature between the two hemispheres, indicating that the dark side does radiate a measurable amount of heat. This appears to eliminate the very long period favoured by Schiaparelli, and though nothing is yet certain, it is probable that the planet's day is longer than several terrestrial weeks but shorter than one terrestrial year.

The determination of the rotation period by means of the Doppler effect at opposite limbs is made difficult by the fact that when both limbs are visible (at superior conjunction) the angular diameter of the disc is too small for measurements to be made. Conversely, when the disc is large enough for accurate work to be undertaken, only one of the limbs is visible and therefore the size of the shift is reduced by half. Not too much weight can therefore be attached to such results as have been obtained, and the only conclusion that can be drawn from their negative character is that the rotation is not rapid, and is probably not completed in a period shorter than several weeks.

Venus: phases

Venus, moving in an orbit nearer the sun than that of the earth, exhibits a full series of phases. When it is furthest from the earth, at superior conjunction, it will be fully illuminated; as it approaches and then reaches elongation, it will become first gibbous (i.e. of a shape intermediate between full and dichotomy) and then dichotomized, and as it proceeds from elongation to inferior conjunction it becomes an increasingly fine crescent, until at inferior conjunction it is invisible, having its illuminated hemisphere turned directly away from the earth. Now the distance separating Venus from the earth varies from about 26 million to 160 million miles, and this considerable difference is correlated with wide variations in its angular size. At superior conjunction the angular diameter is only 11″, whilst near inferior conjunction as a very fine sickle, the diameter is 64″. Hence the larger the fraction of the whole disc that is illuminated, the smaller is its angular diameter. For this reason Venus is not brightest when fully illuminated, but when a thick crescent, corresponding in phase to the five-day-old moon. At this time Venus is the most brilliant object in the sky (the sun and moon excepted); it is about fourteen times as bright as the brightest star, and is visible to the naked eye in broad daylight.

Mars: observational advantages

Although Mars, the earth's neighbour on the side furthest from the sun, is smaller than Venus and never approaches as near the

earth, yet we have a much more exact knowledge of its size and mass and of the physical conditions prevailing upon its surface. This is due to three factors in particular: the orbit of Mars lies outside that of the earth whereas that of Venus lies within it; Mars has two satellites; and Mars has a clear atmosphere which permits study of the actual surface of the planet. It will be remembered, in connexion with the last point, that the surface of Venus is hidden from us and that of Mercury is seen with difficulty owing to the smallness of the disc and its unfavourable position for observation.

Fig. 19 shows that when Venus is fully illuminated it is furthest from the earth and consequently subtends the smallest angle at the observer's eye; when it is near and comparatively large in angular measurement only a small fraction of the hemisphere that is turned towards the earth is illuminated. These circumstances represent a grave hindrance to telescopic study, a hindrance to which the study of the outer planets is not subject. It will be clear from Fig. 53 that the earth and Mars are nearest one another when the line joining them would, if produced, pass through the sun. When Mars is in such a position in respect of the earth it is said to be in opposition; it is then 180° removed from the sun and culminates at midnight. It is particularly to be noted that, unlike an inner planet, Mars is fully illuminated when nearest the earth and therefore subtending the greatest possible angle at the observer's eye: a concourse of circumstances that greatly aids satisfactory observation. Mars does not, of course, exhibit a complete cycle of phases as the inner planets do, but when in quadrature it is noticeably gibbous.

Mars: geocentric positions and linear size

There is another feature of Mars's orbit which is of the greatest practical importance and which must therefore be mentioned briefly. The orbit is, like that of Mercury and unlike that of Venus, markedly eccentric. The planet's mean solar distance is 142 million miles, with a variation of some 13 million on either side. One result of this eccentricity will be seen from Fig. 53. If a number of radiating lines are drawn from the sun to the Martian orbit it will be found that the sections of these lines that lie between the orbits of Mars and the earth are of different lengths. That is, at some oppositions Mars is nearer the earth than at others. When the two bodies are in the positions *m* and *e* respectively the opposition distance is minimal, and the opposition will be a particularly favourable one for observational purposes. The distance of Mars under these circumstances of favourable opposition is 35 million miles—9 million miles further

than Venus when at inferior conjunction. At least favourable oppositions the distance may be as great as 63 million miles, while at conjunction Mars recedes to an average distance of 235 million miles; it then has an angular diameter of only 3″·5, and is an inconspicuous object in the morning or evening twilight.

But at favourable oppositions its angular diameter is 25″, an angle

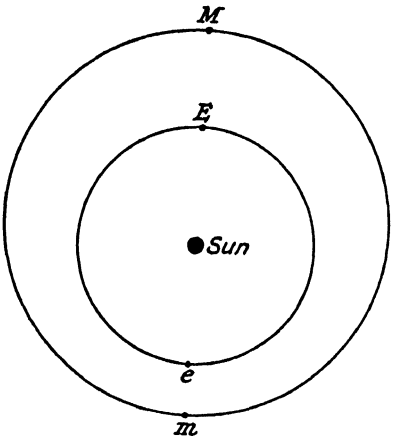


Figure 53. The outer circle represents the orbit of Mars, the inner one that of the earth. When the planets are in positions *m* and *e*, the opposition will be most favourable; when in positions *M* and *E*, least favourable.

subtended, at a distance of 35,000,000 miles, by a linear distance of 4,200 miles. As regards size, Mars is thus intermediate between the earth and the moon.

Mars: mass

Mars has two satellites, and this fact allows its mass and thence its superficial gravity to be calculated with much greater facility and accuracy than is possible in the case of moonless planets such as Mercury or Venus. For it follows from Kepler's third law that

$$\frac{\text{mass of planet}}{\text{mass of the sun}} = \left(\frac{\text{radius of satellite's orbit}}{\text{radius of planet's orbit}}\right)^3 \times \left(\frac{\text{period of planet}}{\text{period of satellite}}\right)^2$$

It will now be seen why an account of the method whereby the earth's mass could be determined was given in Chapter II. For from it the sun's mass can be deduced and this quantity is the only one in the above equation (apart from the unknown) which cannot be derived from simple observation of the planetary system. The period required by Mars to complete one revolution is 687 days; the revolution period of each of the satellites may be determined by inspection; so may the radii of their orbits; and the mean solar

distance of Mars is known (it is, as we have seen, about 142 million miles).

This method is not only simpler but also more accurate than that dependent upon perturbations, and it is found that the mass of Mars is 11 per cent. of the earth's; from this, and its known linear size, it follows that its surface gravity is only 38 per cent. that of the earth's.

Mars: temperature and atmosphere

Mars is half as far again from the sun as the earth, and radiometric work has shown that its surface is correspondingly cooler. During the day the temperature in equatorial regions probably rises to 10° C., but during the night it falls to at least −85° C., extremes which are consistent with a rarefied atmosphere. Considerations of this fact and of the known mass of the planet show that the velocity of escape is high enough for Mars to have retained an atmosphere, though one of lower density than that of the earth. A number of independent observations confirm this and establish quite definitely that a Martian atmosphere exists. Perhaps the most striking of these proofs is that provided by photographs of the planet taken in infra-red and ultra-violet light by means of colour screens. Mars as depicted in the former is similar in appearance to the planet as photographed with ordinary light or as seen with the naked eye, all features known to be surface markings (they will be described shortly) being shown. But the ultra-violet photographs show a featureless blank. Clearly, the infra-red radiation has penetrated the atmosphere to the surface and thence been reflected to the camera in the terrestrial observatory. But the ultra-violet radiation of short wavelength is scattered in the upper reaches of the atmosphere, never reaching the surface; the ultra-violet photographs, in fact, are photographs of the planet's atmosphere. The particular significance of the photographs, however, lies in the fact that the images of the Martian disc are of different sizes in the two series, the ultra-violet and the infra-red: in the former they are measurably larger than in the latter. Obviously, then, since the former record the outer limits of the atmosphere and the latter the body of the planet itself, the difference between the sizes of the two sets of images represents twice the thickness of the Martian atmosphere. By careful measurement of this difference it has been estimated that the atmosphere of Mars is at least sixty miles in depth.

The albedo of Mars is higher than that of Mercury or the moon, though not so high as that of Venus—about 15 per cent. This is not incompatible with the assumption that Mars has a clear and, compared with the terrestrial, rather rare atmosphere; that it is clear is

shown, too, by the fact that we can see the surface of the planet through it. Three other corroboratory facts should be noted. When the planet is in the gibbous phase a distinct twilight zone is observed; furthermore, surface details situated near the limb are fainter than those at the centre of the disc, and than themselves when in the latter position. This, as has already been pointed out, can only be accounted for on the supposition that there is a rarefied atmosphere overlying them and that it is the greater thickness of this atmosphere through which the objects near the limb must be viewed that causes their partial obscuration. Finally, clouds have been observed floating above the Martian surface, especially near the limb where the sun is rising, and this would appear to put the matter beyond all doubt.

Mars: spectroscopic evidence

The spectroscope, as usual, provides the final and conclusive piece of evidence. The bands of water vapour have been identified in the Martian spectrum, as also have those of oxygen. The atmosphere is nevertheless rare by terrestrial standards, and it has been estimated that at the Martian surface the concentration of these two gases cannot exceed about 5 per cent. and 15 per cent. respectively of that in the terrestrial atmosphere. More recent work has suggested that the latter figure may be too high, and that the incidence of oxygen is nearer 0·1 per cent. of the earth's.

It is worthy of note, in passing, that the spectra of Mercury, Venus and Mars are in complete contrast with one another: that of Mercury lacks all trace of atmospheric absorptions, that of Venus contains bands due to carbon dioxide but none of oxygen or water vapour, while that of Mars has unmistakable absorptions due to these two gases.

Mars: surface features

A glance at Mars with quite a small telescope, when the planet is in favourable opposition, will reveal that the general colour of the surface is orange, and that a number of greenish markings are superimposed upon it. The old astronomers called these blue-green areas maria, or seas, and the name has persisted although it is now understood that they are not areas of water; for one thing, the sun is never reflected in them as it would be if they were liquid surfaces. In 1719 were seen for the first time the features to which the name polar caps was given. Mars rotates upon its axis in about twenty-four and a half hours, and at each pole of this axis is situated an approximately circular white patch. The colour and position of these patches suggested the obvious conclusion that they were composed of snow

and ice, and were in every way analogous with our terrestrial polar icefields.

Mars: seasonal changes

This guess was elevated from the realm of supposition when their changes of appearance were worked out in detail. Mars's axis of rotation is, like the earth's, inclined to the plane of its orbit,[1] and it is the varying position of this axis relative to the sun, as the planet moves round its orbit, that causes the change of seasons: when the northern hemisphere is inclined inward towards the centre of the orbit it is summer in the northern hemisphere and winter in the southern, and vice versa. When the northern half of the axis is just beginning to tilt over towards the sun—when midwinter has just passed in the northern hemisphere—it will be noticed that two changes occur: the northern polar cap begins to shrink, while the southern, which has been small during the past summer, begins to grow larger. As the northern, 'summer' cap continues to shrink it is often seen to be surrounded by a dark band. For some time this band increases in width, its outer edge remaining approximately stationary while its inner edge remains contiguous with the retreating cap. At the same time the tint of the greenish areas of the northern hemisphere begins to deepen, the darkening being first noticed in the vicinity of the cap itself and thence spreading south towards the equator.

The generally accepted explanation of these annual changes is that with the coming of summer and the gradual rising of the temperature in the summer hemisphere, the polar cap begins to melt. The appearance of the dark band surrounding the diminishing cap would seem to indicate the existence of a temporary sea; on account of the prevailing low atmospheric pressure, however, it is unlikely that water in the liquid state could exist—the snow evaporating straight into water vapour without passing through the liquid stage. However this may be, the released water vapour disseminates throughout the atmosphere of the whole hemisphere, travelling southward. The darkening of the green areas is regarded as the germination and proliferation of some form of plant life under the stimulus of increased temperature and the liberated moisture. The polar caps provide another proof of the existence of a Martian atmosphere, for their changes are only explicable on the assumption of the alternate evaporation and precipitation of water into and from an overlying atmosphere. The conjecture, to be found in old textbooks of descriptive astronomy, that the caps may consist of solid carbon dioxide, is

[1] That is, to a Martian (as to a terrestrial) observer, the ecliptic is inclined to the celestial equator.

now known to have been unfounded, since the temperature may be higher than the vaporising point of this gas (−80° C.) while the deposit is still in the ground.

Mars: the 'canals'

At the favourable opposition of Mars which occurred in 1877 Schiaparelli announced the discovery of a new type of Martian surface feature, to which he gave the name *canali*. These appeared to him as faint dusky streaks, more or less straight, which traversed the orange areas of the planet's disc. The Italian word *canale* connotes 'channel' rather than 'canal', and it is therefore unfortunate that the English-speaking world accepted the translation 'canal', which the *Concise Oxford Dictionary* defines as 'an artificial watercourse for inland navigation'; no idea of artificiality is associated with the Italian *canale*. It is probably this circumstance, more than any other, that has convinced the popular mind of the existence of 'Martians', for which there is no direct evidence whatsoever.

Two years later Schiaparelli confirmed his original observations. At neither of these oppositions was any other observer able to see the *canali* and until the first independent observation was made in 1888 controversy raged in astronomical circles regarding their objective existence. The matter was finally clinched in 1905 when Lowell succeeded in photographing some of the more prominent *canali*. Since then they have been drawn many thousands of times by a host of observers, and the objective basis of at least the larger ones is established. The main point at issue to-day is their precise appearance and nature. From the earliest days of their observation they have presented themselves to different observers in two main guises: as fine, straight, well defined lines that might have been ruled across the planet's image with a sharp-pointed pencil, and as much thicker, blurred markings which lack both the definition and the geometrical appearance of the others.

Schiaparelli's original papers, together with the conventionalized method of drawing the Martian markings adopted by some American observers, undoubtedly exercised a potent influence at the time upon Martian observers generally: the danger of seeing what one expects to see (especially when the object is near the limit of vision) is very real. It is only in more recent years that the extreme linearity of the early observations of the *canali* has given way to the 'leopard-skin' conception of their true appearance. Briefly, what was seen and drawn by Schiaparelli, Lowell and others as fine, geometrical lines, are now recognized to be more truly represented as the

demarcations of darker and lighter areas, or as series of more or less discontinuous markings. But to say that the *canali*, as originally described, were largely an optical illusion is far from saying that markings, which may have been misrepresented as linear *canali*, have no objective reality.

Mars: satellites

Mars has two satellites, discovered in 1877 at the same opposition that yielded the *canali* to Schiaparelli's sharp eyes. Their small linear size[1] combined with their nearness to Mars itself renders them extremely difficult to observe, and accounts for their relatively recent discovery; while their *outré* behaviour qualifies them for the title of most interesting and peculiar satellites in the solar system. Phobos is only 3,725 miles from the Martian surface, and since its orbit lies very close to the equatorial plane of the planet it never clears the horizon in latitudes above 70°. It revolves about its primary in 7h. 39m.; that is to say, the Martian month as determined by Phobos is only about one-third as long as the day. Consequently, to an observer on the planet's surface it would appear to rise in the west and set in the east, passing from the new to the full phase in about four hours. Deimos, the second satellite, revolves at a distance of about 12,500 miles from the Martian surface in a period of 30h. 21m. This is so very slightly longer than the Martian day that it remains above the horizon for three days, during which time it passes through all its phases twice over.

Bode's law

It was pointed out by Titius of Wittenberg as long ago as the eighteenth century that the distances of the planets from the sun stand in a curious mathematical relationship to one another. This relationship is usually expressed in the form known as Bode's law. The terms of the series

0 1 2 4 8 16 . . .

are each multiplied by 3, giving

0 3 6 12 24 48 . . .

If 4 is then added to each term a fairly close correspondence is obtained between the terms of the series and the relative solar distances of the planets, taking that of the earth as 10:

[1] Observations of their apparent brightness, combined with the assumption that their albedo is similar to that of Mars, suggest diameters of some ten and five miles.

Bode's Series	4	7	10	16	28
Solar Distances	3·9	7·2	10	15·2	?
Planet	Mercury	Venus	Earth	Mars	?

Bode's Series	52	100	196	388	772
Solar Distances	52	95·4	192	301	395
Planet	Jupiter	Saturn	Uranus	Neptune	Pluto

In the first place it is to be noticed that the correspondence deteriorates progressively in the case of the four outermost planets of the solar system. Secondly, there is no known planet occupying the gap which Bode's series reveals between Mars and Jupiter at a distance from the sun equal to 2·8 times that of the earth.

The asteroids

What can be the meaning of this gap? Its existence had been noted by Kepler who suggested that it might be occupied by the orbit of a planet which was too small to be perceived. In 1800 von Zach calculated the orbit of this hypothetical planet and organized a band of twenty-four astronomers to undertake a systematic search of the zodiac in the hope of detecting it. Early in the following year Piazzi discovered a previously unknown planetary object, subsequently christened Ceres, and Gauss calculated its orbit and showed that it did indeed occupy Bode's gap. In quick succession, however, three more faint planets were discovered and named Pallas, Vesta and Juno. This unexpected result of the work of the 'Celestial Police' suggested that the planet which originally occupied the Mars-Jupiter gap might have disintegrated into a number of fragments, of which Ceres, Pallas, Vesta and Juno were the first four to be discovered. Nowadays, however, it is not conceived that the disruption of a single planet affords a likely explanation of the origin of the asteroids. It has been found that these bodies revolve about the sun in widely dissimilar orbits—both as regards eccentricity and inclination to the plane of the ecliptic: this does not suggest a common point of origin. In the second place there is no known reason why an already formed planet should disrupt in this fashion. What does appear more probable is that the asteroids represent material which has never condensed to form a single planetary body. The whole question is still wide open to conjecture, however.

The discovery of new asteroids proceeded apace from the early

years of last century, and by the time that Wolf introduced the camera as an asteroid-hunting instrument in 1891 some 300 had been discovered. Wolf's technique accelerated the rate of discovery considerably, and to-day over 2,000 are known, a small percentage of which are probably duplicated discoveries.

What the total number of asteroids is cannot possibly be laid down with any precision, although one calculation has suggested 50,000 as a possible rough figure. On the other hand, it is known that their aggregate mass cannot exceed about 1 per cent. of the earth's, for if it did they would cause perceptible irregularities in the motion of Mars, whereas no such perturbations occur.

The asteroids are all small bodies, a fact which accounts for their having escaped discovery until less than 150 years ago. Only one, Vesta, is visible to the naked eye, and even it is very near the limit of naked-eye visibility. Some of the larger present measurable discs in large instruments, and Ceres, the largest, is known to have a diameter of about 480 miles. There is probably a continuous decrease in size from this figure, the smallest asteroids being indistinguishable from the largest meteors. None of the asteroids is sufficiently massive to have retained any trace of atmosphere, and they may be envisaged as completely barren lumps of rock, without air, water or life in any form; probably the smaller ones are not spherical, like the major planets, but irregular in shape. This conjecture is borne out by the fact that the brightness of many of them varies in a regular manner which is suggestive of the axial rotation of asymmetrical and therefore unevenly reflective bodies.

The width of the zone occupied by the asteroids, bounded approximately by the orbits of Mars and Jupiter, is four times as great as the distance from the earth to the sun. Within this zone lie the orbits of the asteroids. But, as has already been mentioned, these are typically eccentric to a degree more reminiscent of cometary than planetary orbits. The eccentricity of some is so great that part of the orbit may lie beyond that of Jupiter or within that of Mars, the earth, or even of Venus. At the same time many of the orbits are steeply inclined to the plane of the ecliptic, the orbital inclination of Pallas being as great as 35°. They are thus not confined to the zodiac, as are all the major planets (except Pluto), the sun and the moon.

In 1898 the asteroid Eros was discovered. A faint object, only about seventeen miles in diameter, it was nevertheless of very great interest because at the time of discovery and for many years afterwards it was the nearest known planetary body to the earth, its minimum distance being 14,000,000 miles. In 1932 Eros was dispossessed of this title by the discovery of an even smaller asteroid,

Apollo, not more than one mile in diameter. Since that date the asteroids Adonis and Hermes have been discovered, both of which approach the earth even more closely than Apollo. Hermes, discovered in 1937, is at times less than one million miles from the earth; with a telescope, its movement against the starry background can actually be seen from moment to moment.

Jupiter: solar distance and linear dimensions

The next planet outside Mars and the asteroid zone is Jupiter, the first of the so-called 'giant' planets. In almost every respect it is strikingly different from the planets we have already described, all of which in greater or smaller degree resemble the earth.

Jupiter's mean solar distance is 483 million miles, but owing to the eccentricity of its orbit it may be 23 million miles nearer or more distant than this. Its distance from the earth consequently varies from 367 million miles at the most favourable oppositions to nearly 600 million miles at conjunction. Yet at favourable oppositions, despite the great gap of nearly 370 million miles that separates Jupiter from the terrestrial observer, its angular diameter is 50″. Its corresponding mean linear diameter must be 86,700 miles; that of the earth, it will be remembered, is less than 8,000. The volume of Jupiter is therefore 1,300 times that of the earth.

Jupiter: mass and density

Its mass can be determined with greater accuracy than that of any other planet, for not only has it nine satellites, each of which allows a different and independent estimation to be made, but it is sufficiently massive and sufficiently far from the sun to cause easily measurable perturbations of Saturn. It has been found that the mass of Jupiter is 318 times that of the earth, or more than that of all the other eight planets together.

Enormous though this value is, it is not as great as one would expect from an inspection of the figure representing its volume. For since its volume is 1,300 times that of the earth and its mass 318 times that of the earth, its density can only be one-quarter of the earth's. The possible explanation of this surprisingly low density will be discussed later.

Jupiter: telescopic appearance, form, and axial rotation

Owing to the large size of its disc, the smallest telescopes will show the more obvious features of the Jovian system. Three in particular

will be noticed at first glance: Jupiter is not a true sphere, but is flattened very considerably at the poles, bulging a corresponding amount in equatorial regions; its disc is crossed, parallel to the equator, by a number of belts, alternately dark and light; four of its satellites move back and forth in the plane of the belts.[1]

The degree of Jupiter's polar flattening is greater than that of any other planet in the solar system with the exception of Saturn, and careful measurements of the angular subtention of its polar and equatorial diameters have shown that there is a difference of nearly 6,000 miles between them. The polar diameter is 82,800 miles in length while that of the equatorial is 88,700 miles. It is known from laboratory experiments on the constitution and behaviour of matter, and from mathematical deductions based upon such experiments, that when a non-rigid body rotates it will bulge equatorially, the polar regions being drawn down towards the equator. Up to a certain point the amount of the distortion from the spherical is proportional to the velocity of the rotation. The telescopic appearance of Jupiter would therefore lead one to expect a somewhat rapid axial rotation. When this is measured it is found that the expectation is justified. The belts and zones that stripe the Jovian disc contain many irregularities and individual features any one of which may be used to determine the rotation period. If the reader were to observe an equatorial spot as centrally placed upon the disc at, say, 6 p.m. one evening, he would find that it is back again on the central meridian at ten minutes to four o'clock the following morning. But if he decided to check this determination by making a number of different observations and then taking the mean of the different values obtained he would find himself in difficulties. Suppose that for his second estimation he chose a spot in high latitudes and, as before, observed it to be on the central meridian at 6 p.m. Then the next morning it would return to the centre of the disc at five minutes to four. This discrepancy of five minutes between the two observations is much too large to be accounted for on the grounds of observational error, and he would quickly discover that the rotation period is actually different in different latitudes. Jupiter, that is (or at any rate the visible portion of Jupiter) does not rotate as a solid body. The equatorial regions rotate once in about 9h. 50m. whereas the period in high latitudes is about five minutes longer. Two points are to be noted: in the first place the rotation is, as we expected, extremely rapid—it imparts to an equatorial object a velocity of nearly 27,000 miles per hour—and, in the second (as we shall see), we have here a remarkable analogy between Jupiter and the sun.

[1] The remainder of the satellites are not visible in small instruments.

Jupiter: surface features

The second conspicuous feature of Jupiter as seen in a small telescope is the system of belts and zones with which its disc is crossed. These lie parallel to one another and to the equator, and since the axis of rotation is very nearly perpendicular to the plane of the orbit, while at the same time the orbits of Jupiter and the earth are only inclined to each other at a small angle, their edges appear as straight and not as curved lines. In the same way a spot or other marking traces out a straight path across the planet's disc, and not the curved one it would pursue if the axis were inclined appreciably towards or away from the earth.

Though the minor detail of the belts is constantly changing, the larger divisions themselves remain tolerably stable, and changes in the relative positions, sizes or general depth of colour of the belts usually require several months at least for their consummation if a large area of the surface is involved. A telescope of quite moderate aperture shows that change of a minor order is going on the whole time. If the appearance of the disc is carefully recorded, preferably by drawing, and then reobserved about nineteen hours later, it will be seen that the finer detail of the belts, invisible in small instruments, is teeming with differences of shape, position, size and, to a lesser degree, colour. The equatorial region is in a particularly unstable state, and it is there that the most widespread and rapid changes in the configuration of the markings occur. Towards the poles the visible surface appears to be in a more stable condition and the belts much less susceptible to disturbance.

After the bright and dark belt-zone system the most important markings are the spots. These may be either dark or light, and are sometimes very short-lived, though they normally persist for several months. New spots are being formed and old ones dying out continuously. Further alterations in the appearance of the markings are provided by the fact that the spots may have rotation periods slightly different from those of the substrata in which they lie. It is a significant fact, and establishes another similarity between the sun and Jupiter, that the most rapid and frequent changes connected with the spots occur in two zones, in low latitudes on either side of the equator.

One other feature of the Jovian surface must be mentioned, for its size and semi-permanence place it in a class by itself. In 1878 a peculiar, oval-shaped marking was noticed in the south temperate latitudes. It was pale pinkish in colour and of abnormal size. The colour darkened and the size increased until the spot was a great, dark red patch, measuring about 30,000 miles by 7,000. After being

the most conspicuous object on the Jovian surface for a period of nearly thirty years it began to fade; in 1919 this fading accelerated, and by 1922 the Great Red Spot, as it had been named when in its prime, was only just visible. It experienced a slight and temporary revival in 1927 but is now rather inconspicuous. The cause of this gigantic and prolonged disturbance of the visible surface has been the subject of endless controversy from which no acceptable conclusions have yet emerged. Like the smaller and more ephemeral spots, its position relative to the substratum was not fixed—i.e. its rotation period was not the same as that of other objects in the same latitude. Furthermore, its velocity of rotation was not uniform, but fluctuated in a regular and cyclic manner. The significance and cause of this fluctuation are not known.

Jupiter: atmosphere

It will be clear from this short account of the appearance and behaviour of the visible surface of Jupiter that we are not looking at the planet's solid surface but upon the upper reaches of a dense atmosphere. In no other way can the continuous and, terrestrially speaking, gigantic and cataclysmic upheavals, together with the dependence of rotational velocity upon latitude, be explained. Again, we can hardly suppose that a surface bearing so many points of similarity with that of the sun can be solid; the state of constant turmoil, the birth and decay of spots, often extremely rapid, the more rapid equatorial than polar rotation, the restriction of greatest activity to two zones situated between the equator and each pole, are all characteristics shared by the two bodies, and their significance cannot be ignored. The high albedo (about 45 per cent.) suggests the same conclusion, while the spectroscope knits all these loose strands together, as it were, to form a rigid demonstration that the visible portions of Jupiter are gaseous.

Jupiter: spectroscopic evidence

The spectra of the outer planets, unlike those of the inner, are all of a type, certain characteristics growing merely more pronounced as we move outward from Jupiter towards Pluto. The spectrum of Jupiter consists of a continuous background due to reflected solar radiations, and in addition a number of strong and broad absorptions; these occur particularly in longer wavelengths, the infra-red region being almost completely absorbed. The strength of the bands indicates that the atmosphere is dense, and their identification as the absorptions of ammonia and methane is now complete; the former is estimated to occur in quantities equivalent to a layer 30 feet thick

at atmospheric pressure. Even if the other physical conditions were favourable to life in any form resembling the terrestrial, such an atmosphere would preclude the possibility of its occurrence on Jupiter: for neat smelling salts and marsh gas would be as lethal as cyanic acid.

Jupiter: internal structure

An independent line of reasoning suggests that Jupiter's atmosphere is not only dense, but extremely deep. For the abnormally low density of the planet entails one of two alternatives: either the material of which it is composed is in some fundamental way different from that of the earth and of the other planets we have described, or else the solid body of Jupiter is small compared with the whole visible sphere; that is to say, the atmosphere accounts for a considerable fraction of the observed diameter. Since the planets are generally believed to be formed of solar material torn from the sun by the gravitational drag of a passing star in some remote epoch, it is believed that the densities of the solid planetary spheroids are not widely divergent (a belief supported by observation wherever this is possible), and consequently the latter alternative is favoured.

It is on such considerations as these that the modern views regarding the physical nature of Jupiter are based. Though it was at one time widely believed that Jupiter might be entirely gaseous, it is now supposed that it consists of a small solid core[1] surrounded by a dense and exceedingly deep atmosphere. Jeffreys, from his mathematical studies of the densities, sizes, ellipticities and temperatures of Jupiter and Saturn, concludes that the former probably consists of a small rocky core surrounded by a thick layer of ice, the whole enclosed by an atmosphere whose depth is about 9 per cent. of the planet's radius. This atmosphere is probably chiefly composed of such gases as hydrogen, helium, nitrogen, oxygen and methane, and carries particulate clouds of some substance, such as solid carbon dioxide, which occurs in our own atmosphere in the gaseous state.

Jupiter: temperature

Jupiter's solar parallels, its probably largely gaseous constitution, and its great size, led astronomers to suppose that it might have retained some heat of its own, that the core might still be molten or at least hot. But recent radiometric work has shown that its surface temperature is very low, certainly between −100°C. and −200°C. and probably not far off −140°C. This is the temperature that a body with no heat reserves would assume if placed at a distance from the sun equal to that of Jupiter. In other words, Jupiter radiates no more

[1] Small, that is, relative to the visible atmospheric sphere.

heat than it receives from the sun. These measurements also dispose of any idea that the cloud belts may consist of water vapour.

Jupiter: satellites

The last readily noticeable feature was the system of four satellites. On account of their high albedo and large linear size these are bright objects, and were it not for the proximity of the much more brilliant Jupiter they would be visible to the naked eye. Indeed, some observers possessed of abnormally acute long sight have claimed so to have seen them when at elongation. In the experience of the author, however, and probably in that of the majority of observers, field glasses will be required to show them. Two are slightly smaller than the moon, and two are larger than Mercury; with the doubtful exception of the satellite of Neptune, no other satellite in the solar system approaches them in size. Their orbits lie close to the equatorial plane of Jupiter, so that they appear to swing back and forth in front of and behind its disc. Their revolution periods vary from about one and three-quarter days to just under seventeen days, and the outermost is rather more than one million miles from the planet. Since they were first seen by Galileo in 1610, seven more have been added with the increasing perfection of the telescope and the addition of the camera to the astronomer's equipment. These are all small bodies, and consequently, on account of their distance, very faint. The outermost moves in an orbit whose radius is 15 million miles and completes one revolution in about two terrestrial years; thus Jupiter has eleven different months, ranging in length from less than one terrestrial day to over two terrestrial years.

JUPITER'S SATELLITE SYSTEM

Number as discovered	*Mean distance (miles)*	*Sidereal period*			*Date of discovery*
		d.	*h.*	*m.*	
V	112,500	0	11	57½	1892
I	262,000	1	18	27½	1610
II	416,000	3	13	13½	1610
III	664,000	7	3	42½	1610
IV	1,168,000	16	16	32	1610
VI	7,115,000	250	14	40	1904
VII	7,290,000	260	1	24	1905
VIII	14,600,000	738	21	36	1908
IX	15,000,000	745	—	—	1914
X	?	?260	—	—	1938
XI	?	?700	—	—	1938

The two outermost satellites whose orbits have been investigated (VIII and IX) exhibit the interesting phenomenon of retrograde motion. All the planets in the solar system revolve about the sun in the same direction; the majority of the satellites likewise revolve about their planets in this same direction. Only a few of the satellites of the outer planets have retrograde motion, Jupiter VIII and IX, Saturn IX, and those of Uranus and Neptune being examples.

Fig. 54 (not to scale) shows the relative positions of the earth, the sun, Jupiter and one of its satellites. It will be seen that a number of

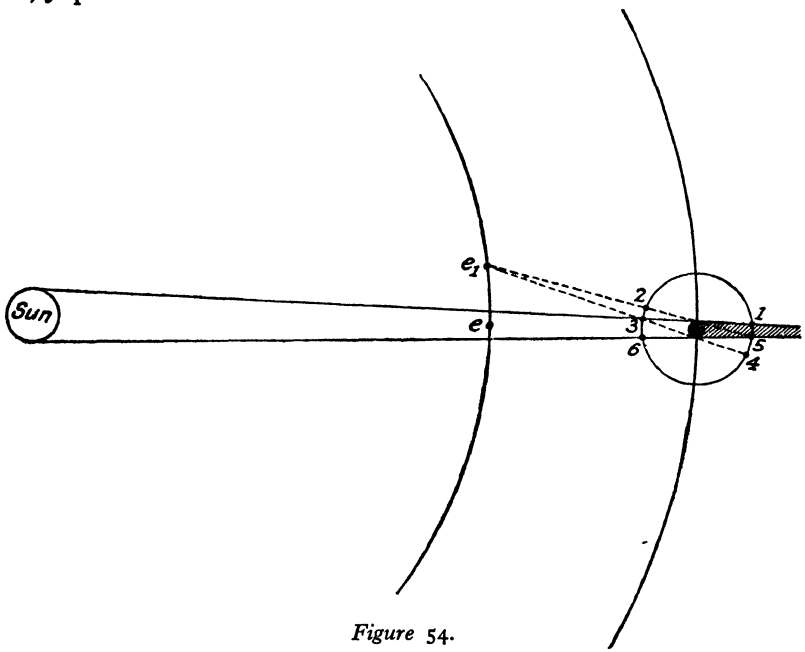


Figure 54.

different things can happen to the satellite, as seen from the earth, during the course of its revolution about Jupiter. We will assume that the satellite is moving in an anti-clockwise direction; when it reaches position 2 it will be directly between the earth e_1 and the limb of Jupiter. Moving to position 3 it will be seen to cross the planet's disc (it will be remembered that the planes of the orbits lie near the equatorial plane of the planet) and pass off it when it reaches 3. Between 2 and 3, that is, it will have been in transit. Just before it passes out of transit at position 3 its shadow will fall upon Jupiter's limb; as the satellite moves towards position 6 the shadow will cross the disc and leave it when the satellite actually reaches 6.[1] Continuing

[1] The relative positions of the satellite and its shadow against the planet's disc are obviously determined by the relative positions of Jupiter and the earth. At opposition, for instance, when the earth is at *e*, the shadow will be invisible, since directly behind the satellite.

round its orbit, it comes to position 4, when it passes behind Jupiter, i.e. is occulted. But before it can pass out of occultation at 5 it has entered the planet's shadow and is therefore eclipsed, not passing out of eclipse and becoming visible once more until it has reached position 1. Like the relative positions of the transiting satellite and its shadow, the order of occurrence of eclipses and occultations is clearly dependent upon the relative positions of Jupiter and the earth. At opposition (to take the simplest case) there can be no observable eclipses of the satellites, since Jupiter's shadow falls directly behind it, as seen from the direction of the earth.

Discovery of the finite velocity of light

It has already been mentioned parenthetically that light is propagated with a velocity of approximately 186,000 miles per second. Though this value may now be derived with greater accuracy experimentally, it was first determined (and the finite velocity of light established) by the Danish astronomer, Ole Römer, as the result of observations of the eclipses of Jupiter's satellites. He found that predictions of the times at which eclipses would occur were marked by a curious rhythmic inaccuracy. Also that the extent of the discrepancy between the predicted and the observed could be correlated with the relative positions of Jupiter and the earth in their respective orbits, and specifically with their distances apart. When the planet was in opposition the eclipses would occur as predicted; as the earth moved away from Jupiter towards the other side of the sun their unpunctuality would increase; the maximum value of this lag—sixteen and a half minutes—would be attained at conjunction. Thenceforward, as the distance between Jupiter and the earth decreased once more, so would the difference between the predicted and the observed. Römer therefore made the brilliantly original suggestion that the extra sixteen and a half minutes was required by the solar light, reflected from the Jovian system to the earth, to travel the extra distance AB in Fig. 17. That is, light travels the diameter of the earth's orbit in sixteen and a half minutes. To accomplish this it must have a velocity in the neighbourhood of 186,000 miles per second.

There is another point of historical interest about the Jovian satellite system. This system is a miniature replica of the solar system: Jupiter taking the place of the sun, and its satellites the place of the planets. Galileo's telescopic observation of the Jovian system was for this reason of particular importance as a factor in the astronomical renaissance and the supersession of the Ptolemaic by the heliocentric hypothesis.

Saturn: solar distance and linear dimensions

Saturn is in many respects similar to Jupiter—and, insofar as this is so, unlike the 'terrestrial' planets—but in one respect, to be discussed later, it is to be sharply differentiated not only from Jupiter but from all the other members of the solar system.

Its mean solar distance is 886 million miles, and its distance from the earth consequently varies from 793 million miles to nearly one thousand million. Its angular diameter of 20″ at the former distance corresponds with a linear diameter of about 75,000 miles. It is thus slightly smaller than Jupiter, although its polar flattening is more pronounced, the polar diameter being only 67,000 miles. We should consequently expect to find that its axial rotation is more rapid than that of Jupiter.

Saturn: surface features and axial rotation

Saturn's disc bears indistinct equatorial bands, very much fainter than those of Jupiter, and also occasional white spots, one of which is shown in Fig. 55. The faintness of Saturn's surface markings as compared with those of Jupiter cannot entirely be accounted for on the grounds of Saturn's greater distance, not only from the observer but also from the sun; they are intrinsically less boldly marked than Jupiter's. The observation of spots has allowed the rotation period to be measured, however, and it has been found to be ten and a quarter hours in equatorial regions. Owing to the rare occurrence of spots, particularly outside the equatorial zone, it has not been easy to determine accurately the rotation period in high latitudes. Such data as are available indicate that Saturn rotates more slowly near the poles than at the equator.

Saturn: internal structure

Saturn thus rotates more slowly than Jupiter[1] despite its greater polar compression. This is somewhat puzzling, but it may indicate that the atmosphere of Saturn is deeper in proportion to the size of the planet than that of Jupiter: Jeffreys gives the depths of the atmospheric shells of Jupiter and Saturn as, respectively, 9 per cent. and 23 per cent. of their radii. This supposition does not conflict with the results of other lines of inquiry which will be mentioned later.

Saturn: mass and density

Light has been thrown on the problem of the size of the solid core of Saturn, as compared with that of the observed oblate spheroid, by

[1] We have seen that the rotation of Jupiter would impart a velocity of some 27,000 m.p.h. to an object on its equator. The corresponding velocity in the case of Saturn is only about 22,000 m.p.h.

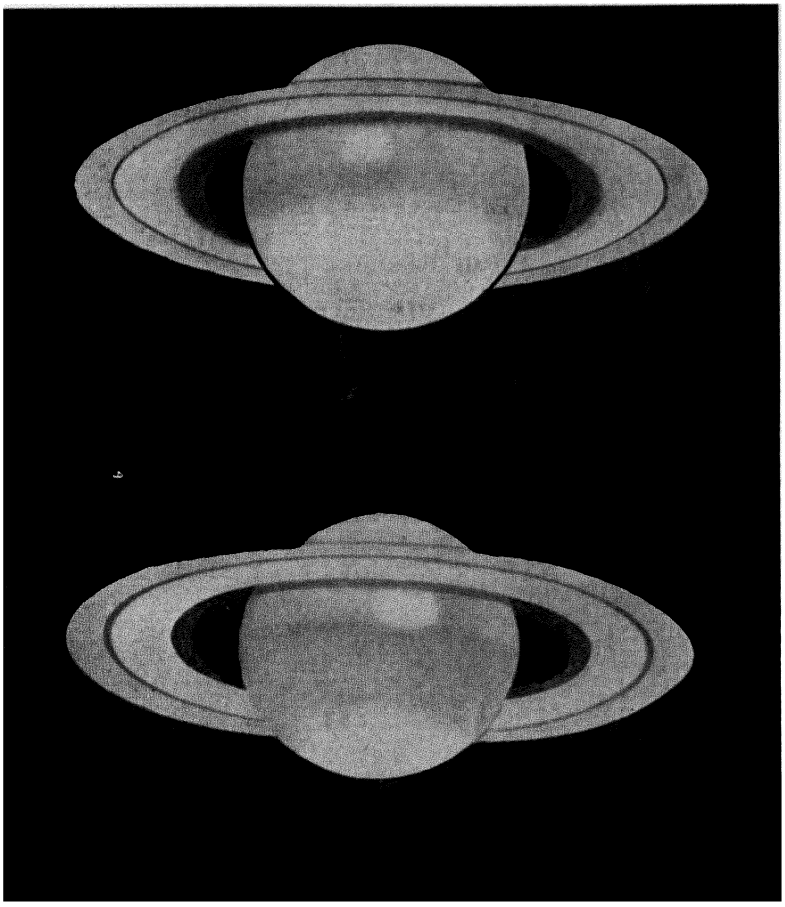


Figure 55. Saturn. Drawings showing the structure of the rings, the characteristic faint belts, and the prominent white spot of August 1933. (*From the* Journal of the British Astronomical Association.)

a consideration of the properties dependent upon its mass. Fortunately, Saturn has a large family of satellites—nine certainly, and perhaps ten—so that its mass can be determined with accuracy. It has been found to be ninety-five times that of the earth. Since the volume of the visible spheroid (an uncertain proportion of which is atmosphere) is 735 times that of the earth, its density is even lower than that of Jupiter: less than one-fifth of the earth's. Such a density has been likened to that of a lightly-packed snowball. This again suggests that the core is smaller, relative to the size of the visible body, than that of Jupiter. Furthermore, mathematical investigation of the motions of the satellites suggests that the visible globe is strongly condensed centrally. It is interesting to note that of all the planets, Saturn is the only one which, were it placed in an ocean of cosmic dimensions, would float.

Saturn: spectroscopic evidence and albedo

Moving progressively further from the sun, we encounter also a progressive change in planetary spectra. The bands occurring in the spectrum of Jupiter occur also in those of the more distant planets, but their intensity increases continuously with increasing solar distance; a few new absorptions are also added. This suggests dense atmospheres and, as is known to be the case, decreasing temperature.

Saturn's high albedo (about 45 per cent.), its Jupiter-like system of parallel belts, and the evidence of the spectroscope, all indicate that the visible surface is certainly gaseous. The extent of this atmosphere is, like the size of the solid core, a still unsolved problem.

Saturn: temperature

In general, then, we may say that so far as is at present known the physical conditions prevailing upon Saturn are similar to those of Jupiter, only more so. This similarity extends to the planet's temperature. Radiometric measurements indicate a temperature certainly as low as $-100°$ C. and probably as low as $-150°$ C., but no lower. That this is somewhat higher than the figure obtained by calculation on the basis that Saturn has no source of heat other than the sun, must be attributed to the uncertainty attaching to all such measurements. It is tolerably certain (and the temperature of Jupiter, which is not only a more massive, and therefore more slowly cooling, body, but is also half as near the sun as Saturn, confirms this) that any heat which Saturn may once have possessed has now been dissipated.

Saturn: ring system

But the feature that differentiates Saturn from all the other planets and sets it in a class by itself is its unique ring system, well illustrated in Fig. 55. It consists of a flat ring, very thin compared with its breadth and diameter, lying in the plane of the planet's equator and separated from it by a wide gulf. This flat expanse is composed of two concentric and closely adjacent rings, separated by a narrow rift which was discovered by Cassini in the second half of the seventeenth century and which bears his name. Under favourable observing conditions it is visible in quite small instruments as a fine black line running round the whole expanse of the ring. That it is a real fissure in the surface of the ring and not merely a surface marking is shown by the fact that it is visible on both sides of the ring; furthermore, when it happens that the ring system occults a star it can be seen shining through Cassini's division with undiminished brightness. In the nineteenth century Encke saw a very faint marking on the surface of the outer ring, *A*, and concluded that it was another and smaller division. Owing to its faintness, however, it is believed that it may be a thinning out of the material of the ring rather than an actual gap; it is nevertheless known as Encke's division. But the nineteenth century did add a third ring to Saturn's system, though not by the subdivision of one of the already known rings. The Crepe Ring, or ring C, is a faint and semi-transparent extension of ring B inward towards the planet. It requires fairly large apertures for its observation on account of the fact that it is of such diaphanous texture, the planet being clearly visible through it.

The total width of the ring system from the outer edge of *A* to the inner edge of *C* is more than 40,000 miles. This can be determined by direct micrometrical measurement when the linear distance of Saturn from the earth is known. In order to make even approximate estimates of the rings' thickness and mass, other methods than direct measurement have to be employed. Unlike the equator of Jupiter, that of Saturn is inclined at a considerable angle (27°) to the plane of the orbit. Hence, as demonstrated in Fig. 56, the plane of the rings will pass through the earth twice in each complete revolution of Saturn about the sun; that is, in a period of about thirty years. At these times the rings will be viewed edge-on, and when this happens they disappear completely for several days; at most they are only just visible for a short distance on either side of the planet. This observation provides us with an upper limit for the thickness, for, if they were more than some ten to fifteen miles thick they would be visible from the earth when in the edge-on position. Compared with their breadth, therefore, their thickness is negligible: ten miles as against

40,000, or less than 0·03 per cent. It may, of course, be very much less than this; all we can say is that it certainly is not more. The determination of the mass of the ring system is entirely mathematical and consists in the calculation of the maximum mass they could have without their gravitational pull affecting to a noticeable extent the motions of the satellites, for no such perturbations are to be observed. Here, too, we can do no more than assign an upper limit, greater than which their mass cannot be. The value obtained is one twenty-seven-thousandth that of the body of Saturn; this is equivalent to about 0·003 times the mass of the earth.

The nature of these gigantic and unique rings engaged the speculations of astronomers from the earliest days of telescopic observation. Nothing definite was established until Laplace, the French astronomer and mathematician, demonstrated mathematically that they could not be solid; that is, continuous planes like concentric rings of cardboard. He showed that such a structure would be

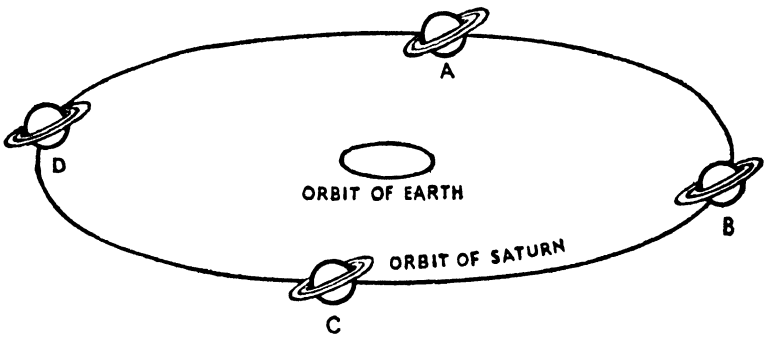


Figure 56. When Saturn is at A or C the rings are presented edge on to the earth. When at D the southern, and when at B the northern, side of the rings is visible from the earth.

unstable and that the gravitational pull of Saturn would disrupt it into a large number of fragments. Though this was only negative information, it was nevertheless a great advance, and paved the way for the further elucidation of the problem by Clerk Maxwell. Maxwell proved that not only could the rings not be continuous, but that they must be composed of innumerable discrete particles, each of which revolves about Saturn in its own orbit and may be regarded as a small, individual satellite.

The Doppler shift has been utilized to prove observationally the truth of Clerk Maxwell's theoretical account of the structure of the rings. The two possible explanations—in so far as Maxwell's allowed of any alternative—were that (i) each ring is a solid structure, or that (ii) it consists of a large number of individual particles, each revolving

about the planet. If (i) represents the true picture of the ring, it is clear that the outer edge will revolve with a greater linear velocity than the inner, since it has further to travel and the same time to do it in. But if the ring is composed of separate particles, each will move in accordance with the laws of Kepler. That is, the nearer a particle is to the planet, the higher will be its linear velocity. Thus the relative velocities of the inner and outer edges of the ring system will be the touchstone by which the rival hypotheses may be put to the test. These velocities can be measured by the amounts of the Doppler shifts of the Fraunhofer lines of the reflected solar spectrum when the spectroscope slit is adjusted to cut the outer and inner edges of the ring. Keeler, who first made this crucial observation, found that the radiations reflected from the inner edge of the principal ring were displaced to a greater extent than those from the outer, the displacements corresponding with velocities of 12½ and 10 m.p.s. respectively. Hence the ring cannot be a solid structure, but must consist of a large number of particles, each reacting individually to the gravitational pull of the planet.

Roche's limit, and the origin of Saturn's rings

After the nature of the rings, the most intriguing problem connected with them is probably that of their origin. Here again mathematics comes to the aid of the observer, and we are indebted to the theoretical work of Roche for some invaluable information concerning the interactions of planets and their satellites. When a satellite is far from the more massive body, the gravitational force exerted by the planet upon different parts of its substance is very nearly uniform. But the nearer it is to the planet, the more disparate will be the stresses within it, until, if the satellite were imagined continuously approaching the planet's surface, a point will be reached at which the conflicting stresses and strains will have become so acute that the satellite will be disrupted. Roche proved that in the case of a large planet and a small satellite, both of the same density, the satellite will be disintegrated when it passes a limit equal to 2·45 times the radius of the planet. Observation affords negative confirmation of this, for no satellite of any planet in the solar system lies within what is known as Roche's limit.

Now the significant fact about the Saturnian ring system is that the radius of the outer edge of ring *A* is 2·3 times the radius of Saturn itself. The obvious interpretation of this fact in the light of Roche's discovery is that at some past epoch one or more of Saturn's satellites have approached too near to the planet—passed Roche's

limit—and have been torn to pieces; and that each individual fragment has continued to revolve about Saturn as one of the many components which together make up the rings. Three observational facts appear to lend support to this hypothesis: all the extant satellites revolve about Saturn in or close to the plane of the rings; both the rings and the inner remaining satellites are of abnormally high albedo, which may indicate similar composition and origin; and thirdly, the innermost satellites both of Saturn and of the other planets are in every case at distances from their primary which are well beyond Roche's limit. On the other hand, it is possible that the rings result from the reverse process. They may consist of matter which at some past epoch in the planet's history was ejected from the body of Saturn itself and which has not yet condensed into satellites. The whole question is still exceedingly obscure, but although no definite answer can be given it does seem probable that Roche's work holds the clue to the enigma.

Saturn: satellites

Saturn's family of moons has nine members; a tenth was claimed by W. H. Pickering in 1905, but the discovery has never been confirmed. The largest, Titan, is larger than Mercury and the largest Jovian satellites. Their periods of revolution range from slightly less than one day to one and a half years; their distances from Saturn from 115,000 to 8 million miles. The main feature of interest about these bodies is that some, if not all, vary in brightness in a regular manner. This light variation is most marked in the case of Iapetus, and was even noticed by Cassini with the poor telescopic equipment of the seventeenth-century observer. The period of variation coincides with the period of the satellite's revolution about Saturn, and it is concluded from this that Iapetus (and others, which exhibit the same phenomenon) always turns the same hemisphere towards Saturn; in this it would resemble our own moon and, probably, the five inner satellites of Jupiter.

The furthermost planets

As regards size, mass and other physical characteristics, Jupiter and Saturn are somewhat similar. The similarity between Uranus and Neptune, the next two planets in order from the sun, is even closer. Owing to their great solar distance (that of Uranus is twice Saturn's) they are faint, and were not known to the ancients, although Uranus is just above the limit of naked-eye visibility and can be found with the help of a star map when its approximate position has been

determined from an almanac. They were discovered in the eighteenth and nineteenth centuries respectively, while the most distant of all, Pluto, was discovered in 1930.

Uranus: solar distance and linear size

Uranus's mean solar distance is 1,783 million miles; this is so great a distance compared with that of the earth from the sun that the planet's brightness and apparent size do not vary to any considerable extent during the course of the terrestrial year. Its mean angular diameter is slightly less than 4″ and corresponds with a linear diameter of about 31,000 miles. It was discovered accidentally by Herschel in 1781. That is to say, he was not engaged in a systematic search for a new planet outside Saturn, but merely noticed a strange object in the field of his telescope; it was at first taken to be a comet.

Uranus: axial rotation

Its period of revolution about the sun is approximately eighty-four years, but its axial rotation is less easily determined. Telescopes of large aperture have shown faint equatorial belts similar to those of Saturn, but no surface marking is sufficiently well defined to be used for an accurate estimate of the rotation period. The polar flattening, however, suggests a fairly rapid rotation, and the period is believed certainly to lie within the limits of eight and twelve hours. The most accurate estimate yet made is that of Campbell. Photometric[1] work has shown that Uranus is subject to a slight and recurrent variation of brightness, the period of which is about 10h. 50m. The most reasonable explanation of such a cyclic variation is that the planet is rotating on its axis in that period and that some areas of its surface reflect less of the incident solar light than others. Campbell's figure agrees well with that of about ten and three-quarter hours derived spectroscopically by Lowell.

Uranus: mass, density and temperature

The mass of Uranus can be derived from the motions of its four satellites and is found to be fifteen times that of the earth; this gives a density about equal to that of Jupiter. This fact, taken in conjunction with the high albedo (similar to that of both Jupiter and Saturn), the evidence of the spectroscope, and the similarity of such surface markings as are visible with those of Saturn, indicates that Uranus has a dense and probably extensive atmosphere. It receives less than 1/360 of the solar light and heat received by the earth, and its

[1] A photometer is an instrument which measures small differences of brightness.

temperature of about −190° C. is slightly higher than would be expected were the sun the only source of heat. But the difference is too small and the practical difficulties attendant upon the radiometric investigation of so distant a body, radiating so little heat, are too great for it to be stated categorically that Uranus has still some reserves of heat unspent. Its physical conditions, therefore, are of the same general type as those of Jupiter and Saturn.

Uranus: satellites

Uranus possesses four satellites, whose periods range from two and a half to thirteen and a half days and whose distances from the planet from about 120,000 to 365,000 miles. The most notable feature of the satellite system is its high inclination to the plane of the planet's orbit; this amounts to more than one right angle, with the consequence that the satellites do not move east and west across the telescopic field (as do those of Jupiter, for instance) but north and south.

Neptune: discovery

Neptune is so similar to Uranus that it may almost be regarded as its twin. In size it is slightly superior to Uranus, but on account of its greater solar distance (nearly 2,800,000,000 miles) it is never visible to the naked eye. The story of its discovery is one of the best known and most interesting in the history of astronomy. Herschel's unexpected discovery of Uranus inevitably led astronomers to wonder whether there might not be other and still more remote members of the solar system. We have seen how the mass of a planet that has no satellites may be calculated by a consideration of its perturbing effects upon the orbits of neighbouring planets. Each planet, were it the only member of the solar system other than the sun, would pursue an orbit which is exactly described by the laws of Kepler; furthermore, its motion in that orbit would be similarly calculable. If, however, there is another planet also revolving about the sun its gravitation will distort the ideal motion of the first, as described by Kepler. Since the mass of the sun is much greater than that of any single planet these perturbations will be small, just as an object held by a strong man can only be moved slightly by a small boy. Accurate observation and mathematical calculation based on the laws of Newton do, however, permit the astronomer to estimate not only the mass of the disturbing planet but also the direction of its action and therefore its approximate position.

About one hundred and fifty years ago it became increasingly clear

that Uranus was not moving to schedule; that is, its observed motion and positions were deviating more and more from the calculated, even after a suitable allowance had been made for all the six planets then known. The only explanation of this disparity between the calculated and the observed was that some factor affecting the latter was not included in the former. That is, that there was a still more distant planet whose perturbing effect was not being taken into consideration. A detailed and extremely laborious mathematical examination of these perturbations led Adams and Leverrier to predict, almost simultaneously though independently, that the unknown planet was situated in such and such a region of the ecliptic. In 1846 Galle, at Berlin Observatory, discovered the new planet close to the predicted position. The discovery was a striking vindication of Newton's theory of universal gravitation.

Neptune: physical characteristics

On account of its great distance very little is known about Neptune, though its physical conditions are probably similar to those of Uranus (which it resembles closely in all ascertainable respects), allowance of course being made for the effects of its greater solar distance. Faint belt-like markings have been thought to have been seen upon it, but since the angular diameter of its disc is only about 2″·5, all such observations must be accepted with caution. Its linear diameter of rather more than 31,000 miles and its mass, approximately seventeen times the earth's, are slightly greater than those of Uranus. Nothing definite is known of its temperature. Assuming that it has no internal heat and is entirely dependent upon the sun for such heat as it has, its temperature must be in the neighbourhood of −220° C. Spectroscopic measures made at the Lick Observatory in 1928 indicate a rotation period of about 15 hours 50 minutes; this is in agreement with the evidence provided by a slight periodic variation of brightness observed many years ago by Maxwell Hall, from which a half-period of 8 hours was subsequently derived.

Neptune: satellite

Only one satellite is known, Triton. It completes its retrograde revolution about Neptune in a period of five and three-quarter days at a mean distance of some 220,000 miles. Its size has not been determined with certainty, but it may be even larger than the giants of Jupiter's satellite system.

Pluto

Most distant of all the known planets is Pluto. Since it was

discovered as recently as 1930, little is yet known of it. Its discovery was made as a result of a planned search based upon calculations of its approximate position upon the star sphere. In broad outline the history of the search and the final discovery was a repetition of that of Neptune. Since, however, Neptune had not completed one revolution of the sun since its discovery in 1846, certain elements of its orbit were not yet determined with great accuracy. Thus any apparent anomalies in its motion were not necessarily due to perturbations on the part of an ultra-Neptunian planet, but possibly to our inexact knowledge of its own motion. For this reason the mathematicians were thrown back upon the perturbations of Uranus for their data. Since the distance from Pluto to Uranus is many times greater than that from Neptune to Uranus, their predictions were less accurate than those of Adams and Leverrier, and Pluto escaped discovery for a correspondingly longer period. When it was identified on a photographic plate exposed in January 1930 it was 5° from the position predicted by Lowell.

Pluto's mean solar distance is approximately 3,673 million miles, or 39·5 times that of the earth. The orbit is abnormally eccentric for a planet, its solar distance varying by something under 2,000 million miles, and at perihelion Pluto is actually nearer the sun than Neptune; the period of revolution is about 248 years.

Its temperature is undetermined, but probably lies in the neighbourhood of −240° C. to −250° C.; an observer situated upon its surface would receive only 1/1,600 as much light and heat as we receive on the earth. These figures follow from Pluto's known solar distance.

The angular diameter of the disc is less than 0″·4, which corresponds with a maximum linear diameter of about 4,000 miles. Our present knowledge of its mass is even less definite, and all that can be said is that it is certainly less than that of the earth, and probably less than half this.

VIII

THE SUN AND THE STARS

The sun: stellar status and size

THE sun is the central and most massive member of the solar system, and is gravitationally the ruler of that system. It is fundamentally unlike the planets; as we shall see, its spectrum alone proves that it is an incandescent body. In fact, it is a star, only differing from the myriad stars of the night sky in its comparative proximity. That it owes its greater conspicuousness to nothing but its nearness is shown by the fact that its absolute magnitude is only 4·85: if viewed from a distance of fifty light years it would only just be visible to the naked eye. At a distance of 32,000 light years (which may be roughly the distance to the centre of the galaxy) its magnitude would be 20, and it would be invisible to the eye even with the aid of the Mt. Wilson 100-inch reflector although it would be within the photographic range of the instrument.

The calculation of the size of the sun is a simple matter, for we know both the size it appears to be (its angular diameter at mean distance is 31′ 59″) and also its linear distance. Thus, for a body to subtend an angle of 31′ 59″ at a point 93,000,000 miles distant from it, its linear diameter must be 864,000 miles. The volume of a spheroid of known diameter is also easily calculated, and it is found that nearly 1½ million earths would be required to build a body the size of the sun.

Telescopic appearance

On turning a telescope (suitably adapted) upon the sun for the first time, the reader will probably be struck immediately by two things: the solar surface bears upon it one or more small black spots; and the centre of the disc is noticeably brighter than the edges. The latter fact indicates that above the shining surface of the sun (known as the photosphere) there must be a cooler atmosphere. Fig. 57 demonstrates that it is possible to see to a greater radial depth at the centre of the disc than at the limb. And since the former is brighter than the latter it follows that brightness increases with increasing radial depth; from which, in turn, it follows that the temperature increases from the outer atmospheric levels towards the solar centre.

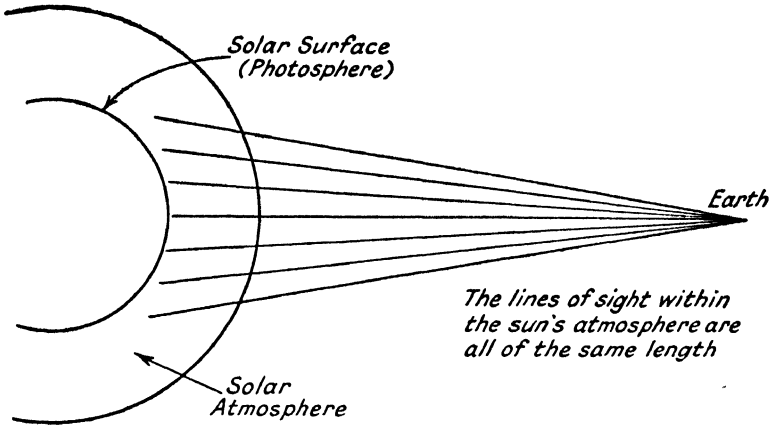


Figure 57. (Not to scale)

Spectrum

The solar spectrum (Fig. 44) consists of a continuous background crossed by many thousands of fine absorption lines. These lines were first studied and mapped by Fraunhofer in 1817, fifteen years after their discovery. The general type of the spectrum thus confirms the inference which we have already drawn from the telescopically observed darkening of the limb: namely, that the photosphere is overlaid by a relatively cool, gaseous atmosphere. It goes beyond this inference, however, in showing that the photospheric surface behaves spectroscopically like an incandescent solid or dense gas under great pressure. Kirchhoff's experiments, and the theoretical work reviewed in Chapter VI, demonstrates that it is the constituents of this atmosphere that impress the Fraunhofer lines upon the continuous spectrum of the photosphere.

Surface temperature and the solar constant

That the photosphere is a radiating surface at a high temperature is obvious without having to undertake any elaborate instrumental investigations. Precisely how hot, can be determined spectroscopically. In Chapter VI the laws of radiation as worked out by Wien, Planck, and Stefan and Boltzmann were stated. Any one of these may be employed to derive the temperature of the photosphere, and it is a satisfactory confirmation of the validity of all of them that the results obtained are in close agreement with one another. Wien's law, based on a maximal intensity at a wavelength of $\lambda 4750$, yields a temperature of 6,070° K.; Planck's equation gives 6,200° K.; and the Stefan-Boltzmann law, 5,960° K.

It may be of interest to follow the application of the latter law as

applied to this specific problem of the sun's surface temperature. It will be recalled that the equation is

$$E=\sigma T^4$$

where E is the radiating body's rate of emission of energy,

T is its absolute temperature,

σ is a constant.

The practical difficulty lies in the accurate determination of E. This is derived from the so-called solar constant, which is, in effect, the amount of solar heat received by a unit area of the earth's surface. It may be expressed in calories per square centimetre per minute, when its value is 1·94; or in ergs per square centimetre per second, when its value is $1{\cdot}35\times10^6$.

But the sun is pouring out radiation in all directions, and not just upon our one square centimetre. In order, therefore, to discover the total amount of its radiation we must calculate the area of the sphere centred upon the sun whose radius equals the mean solar distance of the earth, and then multiply the number of square centimetres in this area by (if we are working in ergs) $1{\cdot}35\times10^6$. If we write R for the mean distance from the sun to the earth in centimetres, the area we require is $4\pi R^2$. Writing k for the solar constant, the total radiation falling upon this sphere in one second=total radiation of the sun in one second=$4\pi R^2k$. Writing r for the radius of the sun in centimetres, we have: radiation emitted by 1 square centimetre of the solar surface per second

$$=\frac{4\pi R^2k}{4\pi r^2}$$

All the terms in this expression are known, and it can therefore be evaluated. The answer is $6{\cdot}25\times10^{10}$ ergs. This, then, is the term, E, which we require.

The value of the constant σ has been found to be $5{\cdot}735\times10^{-5}$ ergs per square centimetre, and substituting the numerical values of σ and E in the Stefan-Boltzmann equation, we get

$$T=5{,}740^\circ\text{K.}$$

A correction for the fact that the sun is not a black body (see p. 150 n.), while the law strictly refers only to a perfect radiator, adjusts this figure to the 5,960° K. already given.

The sun spot cycle

A great deal of information regarding the sun spots can be gained by telescopic observation alone, provided that it is continued over a

long enough period. Moreover, a detailed study of the spots will yield valuable information regarding the sun itself. Early in the nineteenth century, Schwabe, a German amateur astronomer, turned his telescope on the sun for the first time and took note of the spots upon it. He continued his solar observations for nearly twenty years without interruption, except from cloudy days. Each day he made a note of the number of spots visible, and from this simple observation, and as the result of his great patience and perseverance, he discovered their most notable characteristic. Their numbers fluctuate in a fairly regular rhythm. At spot minimum no spots may be visible for days or even weeks on end. Gradually their occurrence becomes more frequent until at maximum there is rarely a day when several spots or groups of spots are not visible. The numbers then begin to fall off again until minimum is once more reached. This cycle from minimum to minimum, or from maximum to maximum, occupies rather more than eleven years; this is only the mean value, however, and periods four or even five years longer and shorter have been recorded. It must be remembered that the fluctuation only involves the number of spots, and not their individual size; large and small spots occur indifferently at all times throughout the cycle.

Solar rotation from sun spots

A few days' observation of the sun will reveal that the spots are moving slowly across the disc from east to west. New spots appear round the eastern limb and, having passed right across the disc (if they persist so long), disappear round the western. The time taken in passing from limb to limb is about a fortnight. It was this observed motion of the spots across the disc *en bloc* that led to the discovery that the sun, like the earth, is rotating on its axis. The earth is revolving round the sun in the same direction as the sun's rotation, and therefore the observed rotation as exemplified by the spots is not the true, or sidereal rotation period, but the synodic period. To obtain the former from the latter we must employ the equation with which we became familiar in Chapter II. If S is the observed, or synodic, rotation period of about twenty-seven days as revealed by the spots; E, the earth's sidereal revolution period of 365 days; and P the sidereal rotation period of the sun; then

$$\frac{1}{P}=\frac{1}{S}+\frac{1}{E}.$$

If the equation is solved for the above values, it will be found that the sidereal rotation period is approximately twenty-five days.

But a curious anomaly makes it impossible to state simply that the

sun rotates in such and such a period. For the derived figure depends upon the latitude of the spot upon whose motion the determination is based. A spot in high latitudes takes longer to travel from limb to limb than one nearer the equator. The sun, in fact, rotates in zones and not as a rigid body: the equatorial regions rotate more rapidly than the polar,[1] the two periods being about twenty-four and a half days and thirty-four days respectively. All that we can say briefly is that the sun's *mean* sidereal rotation period is approximately twenty-five days.

Latitude distribution of spots

The regular observer of the sun will sooner or later discover that spots do not occur in all latitudes with equal frequency. They are, in fact, confined to an equatorial band reaching to about 40° N. and S.[2] In slightly higher latitudes than these they are extremely rare and they are completely unknown in the vicinity of the poles themselves. But even the attainment of this conclusion will not have taught the observer the whole story of the distribution of the sun spots, for about twenty years after Schwabe published his discovery of the eleven year cycle, Spoerer showed that the distribution of the spots in latitude varies with this cycle. At minimum there may be a few spots in high latitudes—usually about 35° N. and S. As the cycle proceeds, new spots are born in progressively lower latitudes until, by the next minimum eleven years later, they are confined to the immediate vicinity of the equator. Shortly before the cycle comes to a close and these equatorial spots die out, the first few spots which herald the opening of the new cycle will have appeared in the higher latitudes again. The two zones occupied by these spots and their successors will in turn converge upon the equator, pass their maximum, and dwindle away in the equatorial regions at the next minimum.

Thus we can, by direct telescopic observation carried out over a period of years, establish three facts about sunspots:

i. Their numbers vary throughout a cycle whose duration is approximately eleven years.

ii. They are restricted almost entirely to two comparatively narrow zones, parallel to the equator and situated on either side of it, the latitude of which varies with the eleven-year cycle.

[1] Since no spots occur in the polar regions, another method of determining the rotation period has to be employed, and will be described shortly. In the highest latitudes at which spots normally occur, the period is about twenty-seven and a half days.

[2] More precisely, to two bands, one north and one south of the equator, for few spots occur between 5° N. and S.

Figure 41. Sun spots photographed in integrated light, showing details of umbrae and penumbrae. (*From* General Astronomy, *by Sir Harold Spencer Jones. London: Edward Arnold & Co.*)

iii. The sun's rotation carries them across the visible hemisphere from east to west with velocities which vary with the solar latitude.

Spectroscopic determination of the solar rotation

The Doppler phenomenon allows the rotation of the sun in any latitude to be measured. Since estimates based on spot observations are restricted to the equatorial and temperate zones of the sun, this is of great value; without the spectroscope the determination of the rotation period in the spotless polar regions could never be undertaken. Fig. 73 represents the sun as seen from a point in space directly above one of its poles. It will be seen that, owing to its axial rotation, one limb is approaching the earth, while the other is receding from it. The velocity of this line-of-sight motion can be measured by the extent of the displacement of the Fraunhoferic lines; it amounts to about $1\frac{1}{4}$ m.p.s.—the east limb approaching the earth, and the west receding from it. Hence, since the linear size of the sun is known, the period of rotation in any latitude can be calculated. In this way it has been discovered, and could have been discovered in no other way, that the rotation period near the poles is ten days longer than in equatorial regions.

Appearance, size and structure of spots

Quite a low magnification has sufficed so far, since we have only been observing the motions and positions of the spots. But now we must study a single spot, preferably a large one, with a high enough magnification to reveal its detailed structure. It will then be seen to consist of two regions, an inner called the umbra, and an outer and less dark fringe called the penumbra. These two main divisions of the spot are illustrated in Fig. 58. In addition, one or several minute black spots, known as nuclei, may be visible within the umbra. It will often be noticed that the photosphere in the neighbourhood of spots, and particularly of large spot groups, is brighter than elsewhere. Close scrutiny of individual spots sometimes reveals that filaments or bridges of this brighter material are thrown across the dark umbra from the surrounding photosphere.

That spots are depressions in the photosphere is demonstrated by the so-called Wilson effect: although the umbra of a spot may be centrally placed within the penumbra while the spot is at or near the centre of the disc, as the solar rotation carries it towards the limb that section of the penumbra further from the limb becomes progressively narrower. This, as Fig. 59 shows, is an effect of foreshortening dependent upon the fact that the penumbra slopes in and down to the umbra. The whole question of the formation and structure of

the spots is still vexed, but that they are depressions seems to be well established. Chevallier has estimated that the majority of spots are about 500 miles deep, though they may exceptionally reach down 1,800 miles below the surface.

Spots vary enormously among themselves as regards size, shape, motion (for spots usually have motions of their own, distinct from

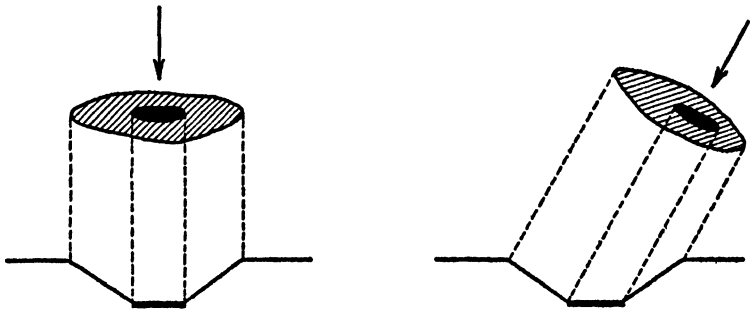


Figure 59. Showing how the appearance of a sunspot varies with the direction of the line of light.

the motion imparted to them by the sun's rotation), and length of life. Some, typically the smaller, may be of only a transitory nature, while the larger ones may last through two, three or, exceptionally, more rotations of the sun. In size, too, the sun spots show great variety. The smallest that can be seen are about 500 miles in diameter,[1] while the largest, and even the more moderate ones, would engulf a body many times the size of the earth. Indeed, a spot must be at least 15,000 miles in diameter before it can be seen with the naked eye, and spots four times this size have been observed.

Though they often occur singly, spots show a marked tendency to cluster into groups of two or more members; such a group is illustrated in Fig. 58. The line joining the separate components of a spot group typically lies parallel to the solar equator; moreover, the larger groups are usually elongated in the same direction.

Lastly, it must be noted that spots appear to be centres of some form of electrical disturbance. Frequently the transit of a large spot or spot group across the sun's central meridian is the herald of magnetic storms, heavy earth currents, and the occurrence of aurorae on this planet. It is significant that these terrestrial magnetic disturbances do not occur simultaneously with the solar disturbances believed to be their cause; the two are separated by a time lag of about twenty-four hours. This is most fortunate in its practical applications, for it permits the forecasting of magnetic storms by the observatories.

[1] Since we have discovered the linear size of the sun, it is a simple matter to discover the linear size of any object upon its surface.

The generally accepted explanation of these terrestrial counterparts of solar activity is that streams of electrified particles are ejected from the disturbed areas of the sun's surface and that these require some twenty-four hours to travel across the 93,000,000 miles separating the earth and the sun. This supposition receives additional confirmation from the fact that very strong magnetic storms are frequently repeated after an interval of about twenty-seven days. Now this is the period of the sun's synodic rotation, and it seems legitimate to conclude that the stream of particles, having encountered the earth, is swept completely round the sun by the latter's rotation and after twenty-seven days encounters the earth once more. Interruption of short-wave radio, on the other hand, is simultaneous with the solar event, and is probably caused by the emission of a flood of ultra-violet light.

In addition to specific correlations of this kind, there is a marked relation between the ebb and flow of the spot cycle and diurnal variations of the earth's magnetic field. At maxima the variations are large, and magnetic storms (which are nothing more than sudden and acute variations) common. At minima the extent of the diurnal variation is reduced by anything up to 50 per cent. while storms occur practically not at all.

Temperatures and spectra of spots

We have seen that the temperature of the general photospheric surface is about 6,000°. At such a temperature no compounds could remain undissociated. But the spectra of spots show the characteristic banded spectra of compounds, and in particular those of titanium oxide, calcium and magnesium hydride, and molecular carbon. Now the dissociation temperature of titanium oxide is in the neighbourhood of 3,000°, and spots are therefore shown to be considerably cooler than the surrounding photosphere—a conclusion which is suggested by their inferior brilliance. Temperature modifications of individual lines in spot spectra tell the same tale. These are of the general type described in Chapter VI: some lines are greatly strengthened while others, especially lines in the infra-red, are weakened; others, again, are unchanged or altogether absent.

Another common feature of spot spectra is the widening, doubling or trebling of many of the lines, especially those situated in the longer wavelengths. This is an example of the Zeeman effect, and indicates, as laboratory experiment has proved, the existence of a strong magnetic field within the source; the spots, then, are centres of electrical disturbance. It is interesting to learn that Hale has detected the presence of spots by means of the Zeeman effect where no spots are visible to the naked eye. It might be suggested that the forces

which normally lead to the formation of a spot are here operative, but are not strong enough to create a visible disturbance of the photosphere.

Faculae and photospheric granulation

Close examination of the brighter material which characterizes the vicinity of spots and spot groups shows that it is not homogeneous or uniform in texture; rather, it consists of numerous irregular patches and worm-like filaments, which are some 15 per cent. brighter than the photospheric background. These are named faculae, and at this point it will suffice to note two facts about them. Congregations of faculae tend to precede the formation of spots, and to survive after they have vanished. Secondly, they appear to be more numerous near the edges of the solar disc than towards its centre. The explanation of this, as we shall see later, is that they are clouds in the atmosphere, situated at considerable heights above the surface of the photosphere. Thus they are not diminished in brightness to the same extent as the photosphere in the peripheral region, while at the same time the reduced luminosity of the background renders them more easily visible.

Quite apart from the faculae, however, the texture of the solar surface is mottled and granular. It appears to be made up of innumerable contiguous grains, only slightly brighter than the photosphere, which are commonly referred to as the rice grains. These granules are from 450 to 1,300 miles in diameter, and Keenan estimates that at any given moment more than 2½ million are visible over the whole disc. The exact cause and nature of this granulation of the solar surface are unknown, but it has been suggested that the granules are the crests of waves of unequally heated photospheric material.

The flash spectrum and the reversing layer

Both visual observation and the nature of the solar spectrum demonstrate that overlying the photosphere there is a cooler atmospheric layer. The question arises, is this atmosphere cool, and therefore non-incandescent, or is it incandescent but nevertheless cooler than the photosphere? In either case its spectrum would consist of dark absorption lines. The question can only be answered if we have an opportunity to observe its spectrum when that of the photosphere (the continuous background) is blotted out. Then, if it were not radiating at all, the Fraunhoferic spectrum would be invisible; whereas if it were incandescent, but less so than the photosphere, the Fraunhoferic spectrum would be reversed, i.e. would consist of bright lines instead of dark. These conditions are provided

at total eclipse. Fig. 60 shows the solar disc, surrounded by the reversing layer, and the body of the moon at a moment just before totality. (To make the diagram clearer it is not drawn to scale, nor are the chromosphere or corona shown.) It will be seen that at this moment the photosphere is hidden by the body of the moon while a narrow crescent of the reversing layer is still visible. At the moment of the final obscuration of the photosphere the Fraunhofer spectrum flashes forth as a set of bright lines; a second or two later, as the moon moves over the reversing layer also, they disappear. It is clear, therefore, that the reversing layer, though cooler than the photosphere, is nevertheless incandescent.

The matching of Fraunhofer lines with lines of spectra produced

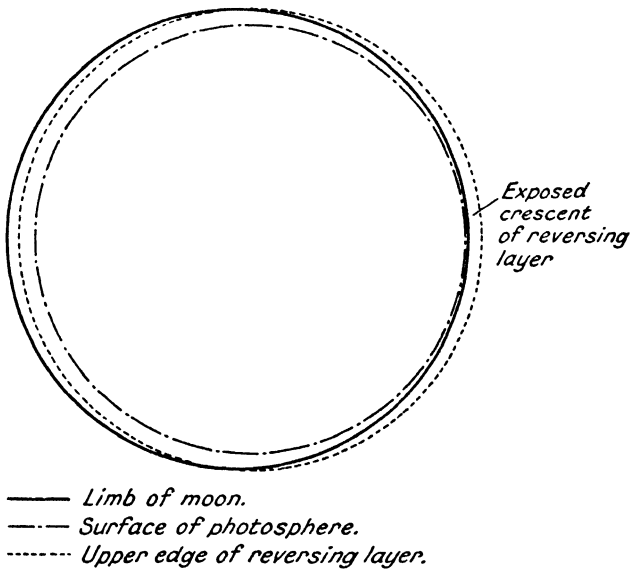


Figure 60. Conditions necessary for the production of the 'flash spectrum'.

in the laboratory, the constitution of whose sources is known, has led to the identification of many terrestrial elements in the reversing layer of the solar atmosphere. This matching of solar and laboratory spectrograms is carried out as follows. The spectroscope slit is divided transversely into an upper and a lower half, either of which may be opened or shut independently of the other. The lower half of the slit is opened, the spectroscope is focused upon the sun, and the plate exposed. Then the upper half of the slit is opened, the lower half closed, and the spectroscope focused upon a source in the laboratory—incandescent iron vapour, let us say—care being taken not to move the plate. The plate is once again exposed, and developed. The spectrogram then consists of two spectra lying side by

side, with equivalent wavelengths directly above and below each other: one the spectrum of the sun, and the other that of vaporized iron. It can be seen at a glance whether any of the iron lines occur in the solar spectrum, and whether, therefore, iron occurs in the reversing layer. By taking a large number of such comparison spectrograms, changing the constitution of the laboratory source each time, it is possible to identify about sixty of the ninety-two terrestrial elements in the reversing layer. Some are represented by only a few lines, others by scores or hundreds; over 3,000 lines of the iron spectrum have been mapped. Altogether, more than 23,000 individual lines figure in the most recent catalogue of the Fraunhoferic spectrum. It is highly probable that the thirty apparently missing elements do in fact occur in the sun. Not only are all wavelengths shorter than about λ3000 absorbed by the ozone in the earth's upper atmosphere, but it must also be expected that the heavier elements in the sun would sink below the reversing layer where alone they are capable of producing absorption spectra.

When the flash spectrum is observed without a slit, each line is curved (being an image in monochromatic light of the still exposed segment of solar atmosphere) and as the moon's limb obscures successively higher and higher levels, so the lines of those elements which occur at different heights above the photosphere will drop out of the spectrum. Finally, only calcium, hydrogen and helium remain before the spectrum disappears altogether. Knowing, as we do, both the sun's distance and the rate of the moon's motion aeross its disc, we can thus determine to what approximate heights above the photospheric surface the various elements extend. The reversing layer, the lowest of the atmospheric levels, stretches for from 100 to 200 miles above the photosphere; some of its constituents are confined to the lower levels while others occur throughout the whole stratum. As already explained, it is possible to deduce the density and pressure of a gaseous or vaporous source from its spectrum, and it has been found that the pressure in the reversing layer is no more than one ten-thousandth that in the earth's atmosphere.

The chromosphere

Without the aid of the spectroscope, the solar atmosphere can be studied only on rare occasions, and even then not in its entirety. By a cosmic fluke, the relative sizes of the sun and moon and their distances from the earth are just such that their apparent sizes as judged by a terrestrial observer are almost identical. It periodically happens that the moon passes directly between the earth and the sun; when this occurs the sun is said to be totally eclipsed, and at the

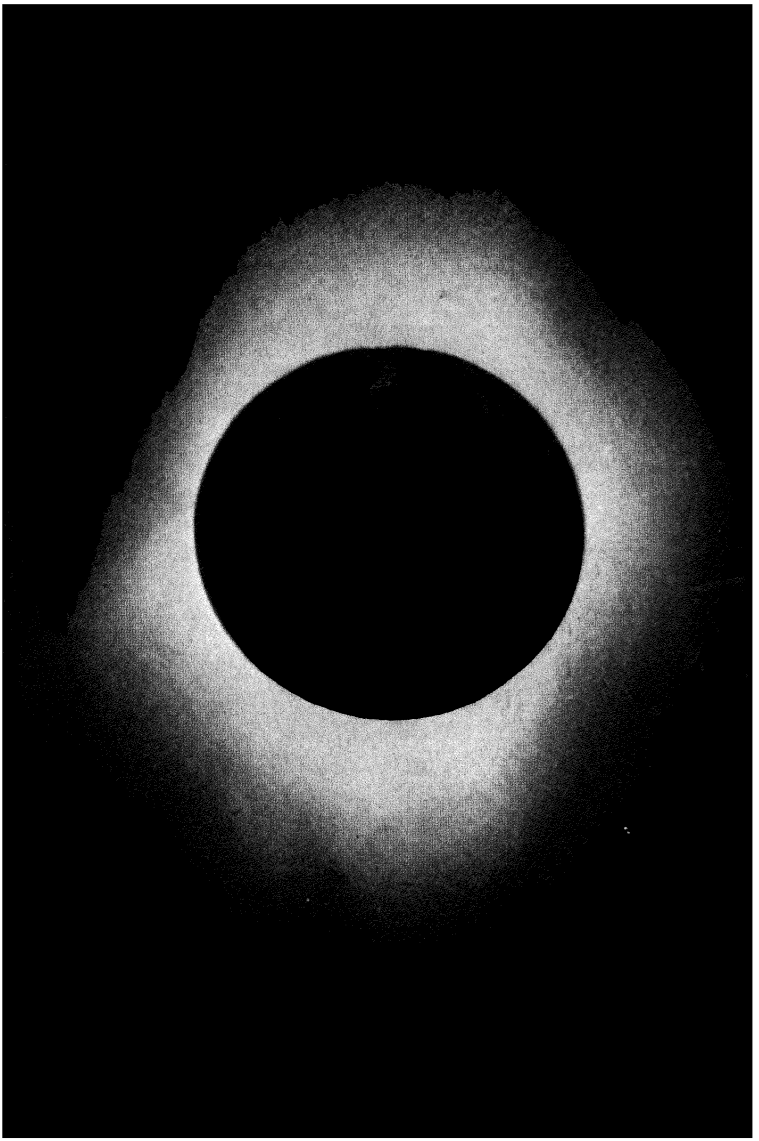


Figure 61. The totally eclipsed sun, showing the corona. (*From the* Journal of the British Astronomical Association.)

instant of totality—which 'instant' may be as long as 7m. 40s., but is usually very much shorter—the solar disc is hidden by the body of the moon. Thus the glaring photosphere is obscured, while the less brilliant atmospheric levels are visible round the moon's limb. Two different zones can then be distinguished. The inner is blood-red, and appears as a narrow rim of flame to the moon's black disc. This atmospheric layer is the chromosphere, and it may be noticed that its outer edge bears small projections, which in fact are gigantic flames and clouds of incandescent gas. These prominences may be attached to the chromosphere at one or more points or may be entirely severed from it.

The spectrum of the chromosphere may also be observed during totality, and it has been established that it does not consist of a complete set of the Fraunhofer lines; furthermore, the lines are bright and not dark. The elements which by their absorption of certain wavelengths of the photospheric light cause the Fraunhofer lines are therefore not situated in the chromosphere. Nor, as we shall see, are they situated in the corona, which is the outermost atmospheric region. This establishes that the reversing layer is situated immediately above the photosphere and below the chromosphere.

One feature of the chromospheric spectrum, which was unexpected at the time of its discovery, was that lines which are weakened in the spectra of spots are here strengthened. Since the solar atmosphere is cooler than the photosphere, just as the spots are, this was somewhat surprising. Later, however, the explanation was revealed by Saha, a physicist. Ionization, the line-modifying agency, is promoted not only by increased temperature but also by decreased pressure, and in the chromosphere the latter overrides the former. Thus line-modifications which were at first thought to indicate certain conditions of temperature are also in fact indicative of reduced pressure.

Until 1868 astronomers could only study the chromosphere during the short and isolated intervals of total eclipse; at all other times the glare of the more brilliant photosphere swamped it utterly. In that year, however, Janssen and Lockyer devised independently a spectroscopic method of observing the chromosphere at any time. Spectroscopic observation of the sun's atmosphere as the moon moves in front of it at total eclipse has proved that when not only the photosphere but also the reversing layer is obscured, there still remain some bright emission lines whose source must be looked for in the chromosphere. Now the effect of high dispersion upon a continuous spectrum is to weaken its intensity, the spectrum being lengthened and the same amount of light spread over a larger area. But the monochromatic lines of the chromosphere are not weakened

by using a spectroscope of high dispersion; they are simply moved further apart. Hence, if a spectroscope of high dispersion is adjusted upon the sun's limb and moved round it till one of the bright chromospheric lines appears, and the slit is then opened, the chromosphere in that region will be visible to the eye; for the intensity of the photospheric light in the sun's vicinity, which usually blots it out of visibility, is greatly reduced by the wide dispersion.

Since this discovery our knowledge of the chromosphere has increased rapidly. Its most remarkable feature is its instability. It is in a constant state of turmoil and upheaval, great flames and clouds of incandescent gas being ejected from its surface. These eruptions are known as prominences, and can be clearly seen by the open slit method with quite small telescopes. They may be roughly divided into two groups, quiescent and eruptive. The former are usually about 50,000 miles in height, and change neither their shape nor their position very rapidly. The eruptive prominences, on the other hand, are subject to very rapid changes: prominences have been observed to be ejected from the chromosphere with velocities approaching one million miles per hour to distances of more than half a million miles from the sun. By regular observation of the numbers and distribution of the prominences Evershed has been able to demonstrate that both are subject to cyclic variation whose period is about eleven years.

The thickness of the chromosphere is about ten times as great as that of the reversing layer, as can be estimated from observations with the slitless spectroscope at total eclipse. Its lower edge tails imperceptibly into the upper reaches of the reversing layer at a height of 100–200 miles above the photosphere, while the highest constituents such as ionised calcium occur exceptionally at heights as great as 8,700 miles. Its chief constituents are calcium, hydrogen and helium, and it is the red light of hydrogen that gives to the chromosphere its characteristic colour. The quiescent prominences, which usually take the form of semi-detached portions of the upper reaches of the chromosphere, have the same spectrum and chemical composition as the regions from which they spring, but the eruptive prominences often contain certain substances, such as iron and tin, which are normally confined to the lower levels of the chromosphere.

The corona

The outermost atmospheric shell is the corona. This appears as a pearly white radiance (the most beautiful of the various striking phenomena witnessed at total eclipses) surrounding the sun to a very much greater depth than the chromosphere (see Fig. 61). Its outer-

edge is not concentric with the sun's limb, as is that of the chromosphere, though it may be approximately so. On the other hand, its figure may be much distorted from the circular by great beams or petal-like rays.

The study of the corona has been rendered excessively difficult by the fact that until 1931 it could only be seen for the short and isolated periods of total solar eclipse: serial or continuous observation has thus been impossible. That this deplorable state of affairs has been remedied is due solely to the historic work of Lyot, whose photographing of the corona without an eclipse has been one of the foremost observational achievements of this century. The difficulty of the observation, which had for so long proved insurmountable, sprang from three causes. In the first place, the enormous brilliance of the photosphere completely swamped the much gentler radiance of the corona: at a distance of 2′ from the solar limb, the brightness of the corona is about one million times less than that of the photosphere, and towards the outer regions it falls off even from this figure. In the second place, there was an instrumental defect to be overcome. A lens does not transmit 100 per cent. of the light directed upon it, a certain proportion being scattered by the two faces and the edge of the lens, as well as by any bubbles in the glass, surface scratches or particles of dust. The resultant halo, when the sun is observed, is at least 200 times as bright as the corona. A third source of diffusion of the photospheric light is dust in the lower levels of the earth's atmosphere: this produces a halo which is at least 100 times brighter than the corona under normal conditions at sea level. Any one of these three sources of diffusion is thus capable of swamping utterly the much fainter glimmer of the solar corona, and all must be eliminated before it can be rendered visible.

Lyot managed to overcome the first two by constructing a type of telescope which he named the coronagraph; this consisted essentially of a system of diaphragms and screens. Also, only lenses of the finest quality were used. The third difficulty was overcome by erecting the coronagraph at the observatory on the summit of the Pic du Midi at a height of nearly 3,000 metres above sea level.

Lyot's first coronagraph was built and put into commission in 1930, in which year also the first coronal spectrogram was obtained without an eclipse. In the following year came the first photograph of the corona itself.

Although the shape and outline of the corona are constantly changing, its general shape varies in a regular manner which is connected with the spot cycle of eleven years' duration. At spot

maximum the depth of the corona is about equal all round the disc; it is at any rate not markedly deeper in one region than another, and certainly bears no particular relation to the solar equator and poles. But at minimum it consists of short tufts at the poles and long, more or less parallel-sided beams or rays at the equator.

Although the whole corona is extremely faint compared with the photosphere, its brightest regions may at times be intolerable to the naked eye. From the immediate vicinity of the chromosphere its brilliance decreases steadily; the inner regions are also yellower in tint than the outer. It is thus possible to make a rather vague distinction between the inner and outer corona, a distinction which is confirmed by the spectroscope. The spectrum of the inner corona consists of less than thirty bright lines, some of which have been known for upwards of seventy years, while others were only detected in 1937. That of the outer corona is quite distinct from this, consisting as it does of a faint replica of the ordinary Fraunhoferic spectrum. This region is therefore understood to consist of non-incandescent particles which merely reflect and scatter the sun's light, superimposing little or no radiation of their own.

But what of the inner corona, with its emission lines? None of these have been matched with laboratory spectra, for the reason that the peculiar physical conditions prevailing in the corona (particularly its great rarefaction, favouring advanced states of ionization) cannot be reproduced experimentally. It was not until as recently as 1941 that Edlén of Upsala fathomed the secret of the inner coronal spectrum, and showed that nine-tenths of the radiation from this structure depended upon unique[1] and 'forbidden' transitions in iron atoms more highly ionized than had previously been envisaged. In addition to iron, calcium and nickel were identified, and at the present time less than 3 per cent. of the total coronal emission is still unaccounted for. It is instructive to reflect that without the theoretical background of ionization potentials and stationary states adumbrated in Chapter VI, and relying solely upon laboratory work, this solution of the coronal spectrum would never have been achieved.

The spectroheliograph

During the last fifty years a vast new field of solar research has been opened up as a result of the invention of the spectroheliograph. This instrument was devised independently by Hale and Deslandres in 1889. If the slit of a spectroscope is brought to bear upon any region

[1] Not only have the coronal lines never been matched in the laboratory, but only for a short time in the spectrum of the nova RS Ophiuchi have they ever been seen elsewhere.

of the sun's disc and the spectrum observed to contain the lines of, for example, hydrogen, we know that a part at least of the region visible through the slit is composed of hydrogen. Now suppose that the eye-piece of the view telescope is replaced by a screen in which is cut a second fine slit. A narrow section of the spectrum will pass through this slit, the rest being caught on the screen. The position of the second slit is adjustable, so that any part of the spectrum can be made to fall upon it. If it is adjusted onto a hydrogen line, only the light from incandescent hydrogen will be able to pass to the photographic plate behind the second slit. The plate, when developed, will record the distribution of hydrogen throughout the area covered by the first slit, and nothing else. Now if the first slit is long enough to cover the whole solar disc, and is moved across it from limb to limb, the second slit being moved with it, a photograph of the sun in hydrogen light will be built up on the plate by a large number of adjoining strips, the images of the second slit. In this way the distribution of hydrogen, calcium, or whatever element is chosen, over the whole visible hemisphere of the sun can be photographed. Fig. 62 is a reproduction of a spectroheliogram in hydrogen light, the $H\alpha$ line of hydrogen being used.

Distribution of solar calcium

Owing to the practical difficulty of occluding all but the required wavelength by the second slit, the exceptionally strong lines of calcium and hydrogen were used first, and it was at once discovered that the distribution of these substances throughout the sun's atmosphere is not uniform. The solar calcium occurs, as may be seen from Fig. 63, in great clouds, to which the name flocculi has been given. These clouds may be either bright or dark, indicating that the calcium vapour is either hotter or cooler than the photosphere. Again the eleven-year cycle is encountered, for the numbers of the calcium flocculi visible increase and decrease in cycles of about eleven years' duration.

Distribution of solar hydrogen

The hydrogen flocculi are distinguishable from the bright calcium flocculi by the fact that they are nearly always dark. Nevertheless, they can never be mistaken for the dark calcium flocculi since the structure of the two types of flocculus is quite different. The calcium flocculi are compact, rounded masses, whereas those of hydrogen are typically wisp-like, and often show a whorled structure suggestive of cyclonic forces. The hydrogen flocculi occur particularly over

disturbed regions of the photosphere, such as spots and spot groups, and these flocculi invariably exhibit the whorled configuration.

It has been found that when the hydrogen flocculi approach the limb they are often associated with prominences. This shows that, as might be expected from their relative darkness, and therefore coolness, they are located in the upper reaches of the solar atmosphere; many of them, in fact, are nothing more than prominences seen in projection against the sun's disc. It is probable, indeed, that the prominences, calcium and hydrogen flocculi and faculae are all objects of a similar nature—i.e. clouds of gas or vapour—only differing from one another in respect of their height above the solar surface, and consequently temperature.

It does not appear unreasonable to seek a connexion between the darkness, and therefore the relative coolness, of the hydrogen flocculi and the fact that hydrogen is lighter than calcium vapour. For we might expect to find hydrogen at greater heights above the photosphere than calcium, and at this level it would certainly be the cooler of the two.

This inference has been verified by an interesting and very important refinement of spectroheliographic technique. Some of the lines of the solar spectrum, particularly those of calcium and hydrogen, are reversed, i.e. in the centre of the absorption line there is a fine, bright emission line. In view of what has already been said of the formation of emission and absorption spectra, it will be recognized as probable that the dark line is caused by absorption in relatively low levels of the atmosphere, while the central bright reversal is caused by the same gas (since the wavelength is the same) at higher levels and in a state of greater incandescence. Thus, in adjusting the second slit first upon the edge of the line and then upon the bright central component, we are photographing different levels of the sun's atmosphere. In this way the vertical distribution of the solar calcium and hydrogen may be studied, and the supposition that the bright flocculi are at a lower level than the dark is substantiated.

The stars

Many facts about the stars can be learnt from direct observation, and it will be convenient to open the following account of the nature of the stars with a description of these observables, considering in turn spectra, temperatures, colours, luminosities, masses, sizes and densities.

Secchi's classification of spectra

It has already been emphasized that the sun and the stars are fundamentally similar. This fact is a certain deduction from their

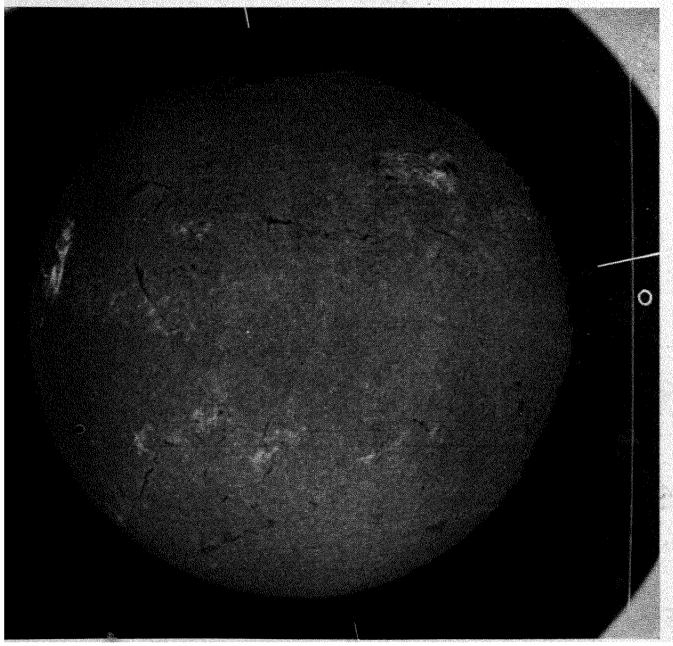


Figure 62. Spectroheliogram taken in $H\alpha$ light at Meudon Observatory on 31 July, 1937, showing the distribution of solar hydrogen.

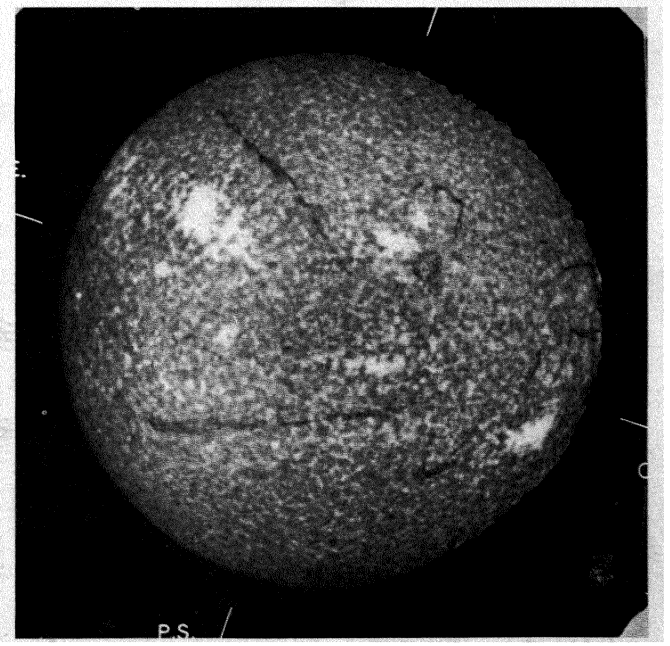


Figure 63. Spectroheliogram taken in the light of ionized calcium at Meudon Observatory on 26 May, 1930.

(*By courtesy of the Director of Meudon Observatory.*)

spectra. Some seventy years ago, Secchi, a pioneer in astrophysics, showed that the vast majority of stellar spectra can be divided into four general groups. A characteristic spectrum of any one of these types is easily distinguishable from that of any other type, while at the same time being closely similar in most of its details to other spectra of its own type. Later work has refined and elaborated Secchi's classification of stars according to spectra, by the division of each of his types into a number of sub-types, and has enlarged its scope by adding several new ones.

The Harvard classification

To-day the accepted stellar classification is the Harvard or Draper Sequence, consisting of a dozen spectroscopic types, decimally subdivided. The advantage of a classification based on spectra is obvious, for the spectrum of a substance or of a star (a conglomeration of substances) is determined by its chemical and physical nature, and thus two stars with similar spectra are also similar in the most important and fundamental respects. In the same way, a classification of the animal kingdom according to skeletal structure is more fundamental, and therefore of greater value, than one depending upon some unsignificant characteristic such as, let us say, colour of skin or hair.

The Draper classification is of the utmost importance in the understanding of modern stellar physics, and without some knowledge of it the general reader will miss much of the significance of the fact and theory of this great branch of astronomy. Over 99 per cent. of all the stars whose spectra have been classified—a total of over a quarter of a million—fall into six of the twelve Harvard types. These are designated B, A, F, G, K, and M, and since stars of the remaining six types are of such infrequent occurrence we may profitably, with one exception, neglect them in this outline account.

Type B: All stars of this type are bluish-white in colour, and their surface temperatures in the region of 20,000° C. The typical spectrum consists of a continuous background upon which are superimposed several prominent absorption lines; although ionized oxygen and nitrogen have been identified, these are primarily due to helium. For this reason, stars of type B are often called Helium stars. This name is, however, liable to lead to confusion, since the primary factor in the production of spectral differences among the stars is not constitution but temperature. For this reason, such names as 'Helium stars', 'Calcium stars' and the like, should be taken as describing spectra and not the stars themselves.

Type A: These stars are of a yellower tint than those of type B, though the majority of them are still whitish; their surface temperatures are of the order of 10,000° C. The strong absorptions of helium have been replaced by those of hydrogen, and faint ionized metallic lines are present in some instances. Sirius is a typical example, and the members of this type are sometimes known as Sirian, or Hydrogen, stars.

Type F: Typically yellowish-white stars, with temperatures of about 7,000° C., of which Procyon is an example. Just as the helium absorptions which characterized type B had faded in type A, so now the hydrogen lines of the latter are relatively inconspicuous. In their place the metallic lines which appeared faint in type A have risen to prominence; the lines of calcium are particularly prominent. F-type stars are also known as Calcium, or Procyon, stars.

Type G: Yellow stars, of which the sun is a typical example. The spectrum is characterized by the very great number of neutral metallic absorption lines that it contains. The hydrogen lines typical of type A are even fainter than they were in the Procyon stars, while the calcium absorptions in the violet are very strong. G-type stars are often referred to as Solar stars. The sun's temperature of rather less than 6,000° C. is typical.

Type K: Deep yellow stars with temperatures of about 4,000° C., of which Arcturus is an example. The hydrogen lines are once again fainter than in the preceding type, the absorptions due to ionized calcium now being at their maximum intensity. The characteristic feature of these spectra is the presence of absorptions due to carbon and certain of its compounds; in this respect they are reminiscent of spot spectra. Type K stars are also known as Arcturian, or Red-Solar, stars.

Type M: Red stars, with temperatures of 3,000° C. or lower; Antares is an example. The fluted bands which first appeared in the last type are here much stronger, while the high temperature lines have disappeared.

The O-type stars must be mentioned at this stage, for although they are of little account numerically, yet their abnormally high luminosities have already drawn them into the discussion at several points. This great intrinsic brightness derives from their high temperatures, which range up to at least 50,000° C. At such a temperature, metallic lines would only occur in wavelengths too short to penetrate the terrestrial atmosphere, and they are in fact absent from O-type spectra. Characteristic, however, are lines of

ionized helium, doubly ionized oxygen and nitrogen, and trebly ionized oxygen.

Preliminary deductions from the Harvard sequence

Perhaps the most remarkable thing about this sequence of types from B to M is the gradual transition both of the colour and of the spectra of the stars throughout it. Blue-white merges into white, white into yellowish-white, into pure yellow, into orange, and finally into red. In the same way we find a progressive fading of certain absorptions and their replacement by new ones. Bands which characterize one type and are strongest in it have faded in the next type and are fainter still, or nonexistent, in the next. The discussion of the significance of this feature of the Harvard Sequence must, however, be left till later.

In the meantime it should be noted that the discovery of the conditioning of the spectrum of a substance by its chemical composition and physical condition allows us to make two presuppositions regarding the stars that fall within these six classes of the sequence; that is to say, regarding 99 per cent. of all stars. (i) We should expect the chemical composition, as well as such physical properties as temperature and mass, of all stars of the same type to be very similar, and (ii) since the spectra change gradually from type to type without any sharp breaks in the continuity of the sequence, we should expect the physical properties of the stars in adjacent types to be more nearly similar to one another than to those of stars belonging to more remote types. Arcturian and Antarian stars, for instance, resemble one another more closely than either resemble Solar stars or, even more, B-type stars. Both these presuppositions have been verified observationally.

Stellar temperatures

We saw in Chapter VI that the temperature of a body which is too distant for direct investigation with a thermometer or thermocouple can be determined spectroscopically in two ways. (i) The brightness of the continuous background is not uniform in all regions of the spectrum; furthermore, the position of the zone of maximum intensity is dependent upon the temperature of the radiating source. The higher the temperature, the nearer to the violet, or short-wave, end of the spectrum does it occur; the lower the temperature, the nearer to the red end. (ii) Not only is polychromatic radiation affected in recognizable ways by temperature, but monochromatic radiation likewise. That is, the absorption lines in a spectrum, each caused by radiation of a single wavelength, are modified by the temperature of

the source of the radiation. Comparison of the stellar spectrum with the flame, arc and spark spectrograms obtained experimentally allow the temperature of the radiating surface layers of the star to be estimated.

By these and similar means it has been possible to verify our two suppositions in regard to temperature. It has been found that there is a steady temperature change along the sequence—the direction of the change being a drop from B towards M—and also that the temperatures of all stars of the same spectral type are of the same order; in other words, the spectral classification is also a temperature classification. The hottest B-type stars have temperatures of about 23,000° C., while that of the coolest M-type stars is about 2,500°. The highest temperature of any star yet precisely investigated is that of Plaskett's star, about 28,000°, although stars of type O normally have temperatures in the neighbourhood of 40,000° or more. At the other end of the scale there are certain stars of types which we have not mentioned, with temperatures as low as 2,000°. These figures represent, with infrequent exceptions, the limits of observed stellar surface temperature. It is to be noted that stars whose temperatures are lower than about 2,000° may well exist, for at such temperatures they would be non-incandescent and consequently we could not expect them to be visible.[1]

Stellar colours

Since star colours are directly dependent upon temperature, it may be as well to interpolate at this point a word on the connexion between colour and spectroscopic type. Everyday experience demonstrates that as a lump of metal—an iron poker, for instance—is heated, it passes through the successive stages of non-incandescence, dull red heat, red heat, yellow heat, and finally white heat. If we had studied its spectrum throughout the changes we should have observed the brightest zone of the continuous background passing steadily from the red towards the violet end of the spectrum. This common knowledge might have led the reader to guess that the temperature is falling steadily along the sequence from B to M, for we have seen that the colour change in this direction is the exact reverse of that just described for the poker. Since the temperature drop along the sequence is gradual, it follows that all tints of the colour range from a blue-white heat to a dull red heat are represented. The following table shows the nature of this correlation between colour and spectral type:

[1] ε Aurigae, the coolest known star, has an estimated surface temperature of 1,700°; the greater part of its radiation lies in the infra-red, i.e. is invisible.

Type	*Colour*	*Percentage occurrence of stars bright enough to be visible to the naked eye*
B	Bluish-white	16·6
A	Pure white	26·8
F	Yellowish-white	9·8
G	Yellow	11·2
K	Deep yellow or orange	26·0
M	Orange-red	9·6
R	Even deeper red	Negligible
N	Even deeper red	Negligible

The reader may be surprised to learn that stars of these conspicuous colours exist, for the majority of those visible to the naked eye appear to be of a nondescript white tint. This is mainly due to the fact that most of the reddish stars are faint, as can be seen from the third column in the above table. Antares is, indeed, the only conspicuous example of a really deeply tinted star, and in our latitudes it is situated so near the southern horizon that it is easily missed. The colours of the majority of stars are various tints of yellow or white, but the difference between, for example, Sirius or Vega (pure white) and Arcturus (yellow) may be seen at a glance. Greenish and bluish stars also exist, but, with one exception, they resemble the deep red stars in being faint. The one exception mentioned is the star known as β Librae, a tolerably bright green star.

Stellar luminosities

The luminosity or real brightness (as distinct from the apparent brightness) of a star may be determined in several ways. Just as, having discovered the distance of the sun, we can calculate its linear diameter from its observed apparent diameter, so in an analagous way, having once discovered a star's distance, we can calculate its real brightness from its apparent brightness. The apparent brightness of a star clearly depends upon two factors: its distance from the earth and its real brightness. Without knowing one we cannot discover the other. The fact that two stars appear to be equally bright is no evidence of their really being so, for one may be twice or a hundred times as distant as the other. But once we know their distances a simple calculation will give us their real brightnesses. To facilitate this, the conception of absolute magnitudes was introduced, as explained in an earlier chapter.

Besides the absolute magnitude scale, luminosity may be expressed by a figure which compares it with that of the sun. Thus a star whose

luminosity is 10 is ten times as luminous as the sun, a star with a luminosity of 0·2, one-fifth as luminous as the sun, and so on. It has been found that the range of luminosity is greater than that of almost any other stellar property. It varies from about 0·000002, the luminosity (discovered in 1944) of the companion of the star whose catalogue number is B. D. +4° 4048, to about 300,000, the maximum luminosity of a variable star named S Doradus.

Visual binaries

The basic observation in the determination of stellar mass is that of the binary star, of which something must now be said. A careful study of the night sky with the naked eye will soon reveal a number of pairs of stars, the two components of which appear to be very close to one another. Telescopic observation shows that they are comparatively common objects. We might at first be led to assume that in each such case the two stars really are near one another in space, as they certainly appear to be. But a moment's thought will show that this need not necessarily be so, for the sun and moon appear to be close together at solar eclipses although the moon is nearly four hundred times nearer the earth than the sun. Thus if two stars happen to lie close to the observer's line of vision—that is, in nearly the same direction from the earth—their angular separation will be small even if one is a hundred times more distant than the other. Angular proximity between two stars on the star sphere is not, therefore, sufficient evidence for their binary nature; the relationship may be purely optical.

There are, however, two ways in which a binary system may be distinguished from an optical double. Firstly, by the detection of orbital motion about their common centre of gravity in one or both of the components, and, secondly, by the detection of common proper motion of the two components. Even if no trace of orbital motion can be detected, we may be tolerably certain that the two stars form a binary system if they have a common proper motion. For the chances against two unrelated stars which happen to appear close together as seen from the earth having an identical proper motion, both as regards direction and velocity, their real velocities being graded to their respective distances from the sun, are too enormous to permit us to appeal to coincidence.

Stellar mass from the observation of binaries

The study of binaries has provided us with the fundamentals of our knowledge concerning stellar mass. If the distance of a binary

system is known, its linear size may quickly be calculated from its angular size; furthermore, mere observation will give the period of the components' mutual revolution, providing this period is not too long for their orbital motions to be perceptible. With these two data —the linear separation of the two stars, and the period of their revolution about the system's centre of gravity—it is possible to deduce the total mass of the system by means of Newton's revision of the harmonic law as enunciated by Kepler.
For

$$\frac{m_1+m_2}{m_s+m_e}=\frac{A^3}{P^2}$$

where m_1, m_2 are the masses of the two stars,

m_s is the mass of the sun,

m_e is the mass of the earth, which, being negligible compared with the other quantities in the equation, may be disregarded,

A is the separation of the two components in terms of the sun-earth distance, i.e. in astronomical units,

P is their period in years.

If, furthermore, the motion of each star relative to the centre of gravity can be determined, their individual masses may be calculated, just as the mass of the moon was calculated.

Such work as this has shown that the limits of stellar mass are more circumscribed than those of size or luminosity. The average total mass of a number of binary systems that have been investigated is 2·2 times that of the sun, the average mass of all these components only differing from that of the sun by 10 per cent. Stars even three times as massive as the sun are rare; so are stars less than one-fifth as massive as the sun. Stars up to ten times as massive are very rare, while still more massive stars are quite exceptional.

Spectroscopic binaries

So far we have only been considering visual binaries—binaries, that is, whose angular separation is great enough for the components to be distinguished individually by the eye, even though a telescope may have to be employed. But there are also large numbers of binaries whose separations are too small to allow of visual resolution. Such systems, of which upwards of one thousand are known, are called spectroscopic binaries, and can be recognized spectroscopically by means of the Doppler shift. Unless the plane of the orbit of the

comes about the primary[1] is at right angles to the line of sight, its distance from the earth will vary: for half of its orbit, its motion will have a positive radial component, and for half a negative. The lines of its spectrum will therefore suffer a continuous, periodic displacement, first towards the red and then towards the blue. If both components are of about the same brightness, both spectra will be visible, and a periodic doubling of the lines will be observed; but if one component is more than two or three times brighter than its companion, then its spectrum will swamp that of the *comes*, and will alone be visible. In this case, the entire visible spectrum of the system will be seen to swing rhythmically back and forth. The first spectroscopic binary to be detected was 3 Ursae Majoris, whose oscillating spectrum attracted the attention of E. C. Pickering in 1889.

Stellar mass from spectroscopic binaries

Although the components of binaries of this type are too close to one another even to be distinguished visually as separate stars, yet their spectroscopic observation and study yield a surprising amount of information regarding them. Only limited data concerning mass can, however—except in special cases, to be described—be derived.

The relative sizes of the shifts of the two sets of lines (where both are visible) give the relative line-of-sight velocities of the two stars, therefore also their relative distances from the centre of gravity, and therefore their relative masses. If, for example, the shift of one set of lines is seven-tenths as large as that of the other, then,

(*a*) line-of-sight velocity of star A: line-of-sight velocity of star B = 7 : 10,
(*b*) linear distance of A from centre of gravity : linear distance of B from centre of gravity = 7 : 10,
(*c*) mass of A : mass of B = 7 : 10.

Only the relative masses of the two components of a spectroscopic binary can be deduced, since the orbit cannot be completely reconstructed owing to our ignorance of its inclination to the line of sight.

Stellar mass from eclipsing binaries

We shall see, however, when we come to discuss eclipsing binaries, that in certain cases the inclination of the orbit to the line of sight is discoverable. In such cases, since we already know the line-of-sight velocities of the two stars, we can calculate their actual orbital velocities, thence the linear sizes of the orbits, and thence their linear separation at any given epoch. The comparison of this linear separation

[1] To simplify the wording, it will be assumed that the *comes* revolves about a stationary primary.

with the angular separation as measured by the interferometer[1] at once yields the distance of the binary. The period of revolution of each star about the centre of gravity is given by the time required for its spectrum to travel from greatest displacement towards the red to greatest displacement towards the violet, and thence back to the red again. From the period and the linear separation of the two stars, their combined mass can be calculated, as we saw a few pages back. And knowing both the combined mass and the relation in which each stands to the other (7 : 10 in the example) we arrive at the individual mass of each star.

Further data from eclipsing binaries

It is interesting at this point to note how much information the ingenuity of the modern astronomer is capable of deriving from the study of these eclipsing binary systems. Knowing the components' distance and their combined apparent magnitude, the astronomer is in a position to calculate their luminosity, either in terms of the absolute magnitude scale or as compared with that of the sun, by means of the inverse square law. The surface temperatures of the stars can be measured spectroscopically, and this, combined with their luminosity, takes us one step further—to the linear surface area which would be required to produce the observed brightness. From the surface area, the diameter is quickly calculated. And finally, from the linear sizes and masses, their densities can (as we shall see shortly) be derived.

Let us summarize the results of this chain of astronomical detective work: the following data have been derived for the spectroscopic binary which we have been considering:

Radial velocity of each component,
Orbital velocity of each component,
Combined and individual masses,
Linear separation at any moment,
Distance,
Absolute magnitude and temperature,
Linear diameters,
Densities.

Bearing in mind Kepler's third law, and the fact that the components of a spectroscopic binary are nearer to one another than the components of a visual binary, we should expect to find that the periods of spectroscopic binaries are shorter than those of the wider pairs which can be resolved visually. This has been found to be so:

[1] See p. 237.

the shortest period of those binaries so far investigated is the two and a quarter hours of γ Ursae Minoris; the upper limit merges into the shortest periods of the visual binaries, there being a smooth transition of period from a few hours to several thousand years. At the lower end of this sequence are the spectroscopic binaries, and at the upper the widest visual pairs.

Stellar size

The luminosity of a star must depend upon two factors: the brightness of its incandescent surface, and the amount there is of it. Increase both, and the luminosity will increase; decrease both, and it will decrease; while an increase in one may just counterbalance a decrease in the other, the luminosity remaining unchanged. This relationship may be crudely expressed in the form

$$\text{Surface brightness} \times \text{Surface area} = \text{Luminosity}.$$

We have already seen (Chapter VI) that the surface brightness of a black body varies as the fourth power of the temperature, for radiation in all wavelengths. It follows, therefore, that if we know the temperature (or the colour index, a function of temperature) as well as the absolute magnitude of a star, then its linear diameter is calculable.

This indirect method at arriving at stellar diameters has proved to be of the greatest value. For both stellar temperature and absolute magnitude are qualities about which a considerable mass of data has been accumulated.

Secondly, as just explained, stellar radii may be derived for a certain type of binary star—the eclipsing binary.

A third method is valuable in that it is especially applicable to the class of star known as white dwarfs, so that the impossibility of applying the interferometer (method 4, below) to these stars is mitigated. It is a deduction (beyond the scope of this book) from the general theory of relativity that the entire spectrum of a source will be displaced by an amount which varies with the value of the expression

$$\frac{m}{r}$$

where m is the mass of the source and r its radius. Before this method can be put into practice, therefore, the star's mass must be discovered. Although the method is theoretical and highly mathematical, it is interesting to note that in the case of the sun it yields a result which agrees very closely with the independently ascertained figure.

The final method of determining the linear diameter of a star requires two data: its angular diameter and its distance from the sun.

If we know the values of both these, it requires only a simple calculation to discover how great its linear diameter must be in order to subtend an angle of x'' at a point y light years distant. Despite the enormous linear sizes of the stars (some have diameters running into hundreds of millions of miles) their distances are of such relative immensity that their angular diameters are minute. So minute, in fact, that not even in the largest telescope now in existence (that at Mt. Wilson, the diameter of whose glass is over eight feet) do they appear as more than points of light.

The telescope, then, cannot unaided show a measurable disc for even the largest and nearest stars. In order to make the determination, the mode of behaviour of light known as interference must be utilized. If light from a source is split up into two beams and then recombined, these two rays will, under certain circumstances, interfere with one another with the production of a characteristic effect. Not to go into too detailed and technical an explanation of the phenomenon of interference, it will be enough to take a specific example. Suppose that the upper end of a telescope is covered with a screen in which two fine slits, some distance apart, are cut. Then light from a source in front of the telescope will be split into two beams, each passing through one of the slits, which are made to converge by the object glass to a focus in the plane of the eyepiece. The result of this combination of the two rays at the focus of the object glass is the formation of alternate dark and bright fringes to the image of the slits. These are known as interference fringes and their formation depends upon this particular property of wave systems.

Now the appearance of the fringes depends, for a given source, upon the distance separating the slits. Adjustment of the positions of the slits will reveal the fact that at a certain distance apart the fringes will disappear. It can be proved, and verified experimentally, that this distance is related in a certain known way to the diameter of the source of the radiation.

This is the principle of the interferometer constructed and used at Mt. Wilson. By its means the angular diameters of a dozen stars of known distance have been measured, and from these data their linear diameters derived. It has been found that stellar size resembles stellar luminosity in the enormous range of its variation, the diameters of these twelve stars alone ranging from 12 to 450 times that of the sun. The diameters of at least two are larger than that of the orbit of Mars. They are Antares (α Scorpii) and α Herculis, and their linear diameters, together with those of three other giants, are given in the table below:

Star	*Diameter in miles*	*Diameter compared with that of the sun*
α Scorpii	$3{\cdot}8\times10^8$	450
α Herculis	$3{\cdot}5\times10^8$	400
β Pegasi	$3{\cdot}5\times10^7$	40
α Tauri	$3{\cdot}3\times10^7$	38
α Bootis	$2{\cdot}3\times10^7$	27

The star ε Aurigae is worth mention at this point. It is a variable, a star whose brightness fluctuates. As recently as 1937 Struve and others have shown that in reality it consists of a close pair of stars, in mutual revolution, and that the brighter is some 3,000 times larger than the sun: if it were placed at the centre of the solar system, where the sun now is, its surface would lie between the orbits of Saturn and Uranus! It is also an interesting object in that it is both the most rarefied and the coolest (1,700° C.) star known.

These stars which we have just been discussing are of course situated at the upper end of the range of stellar size. A more balanced conception of average stellar size is obtained from eclipsing binaries. The mean diameter of twenty-eight stars of this type is a little less than three times that of the sun. This result is substantiated by the indirect methods, already described; these indicate that the average size of stars throughout the galaxy is even lower than that derived from studies of the eclipsing binaries.

The angular subtention of the smallest stars is too minute for measurement by the interferometer, and the first method—that depending upon surface temperature and luminosity—must be employed. The smallest stars of all belong to the class known as the white dwarfs, and are considerably smaller than the sun. The following table illustrates this lower limit of stellar size:

Star	*Diameter in miles*	*Compared with the sun*	*Compared with the planets*
Sirius B	30,000	0·034	$3\frac{1}{2}\times$ earth
Procyon B	7,000	0·008	less than earth
o Ceti B	155,000	0·178	about that of Saturn

Stellar densities

Once the mass and radius of a star are known, its mean density may be derived from the expression

$$\text{density} = \frac{\text{mass}}{\text{volume.}}$$

We have learnt that whereas the masses of different stars are confined within rather narrow limits, their volumes vary gigantically: the smallest stars are comparable with the earth, while the diameters of the giants are in certain cases comparable with that of the solar system. It follows that stellar density must likewise vary within very wide limits. The mean density of such a giant as Antares is only about 0·03 per cent. that of ordinary air; a supergiant like ε Aurigae would be even more rarefied—probably about 3×10^{-9} times water. At the other end of the scale, among the dwarfs, where volume has decreased out of all proportion to mass, densities of the order of 50,000 times that of water are encountered. How matter can be compressed sufficiently to yield such densities will be discussed later.

Having reviewed the methods and results of studying the observable stellar characteristics, reference must be made to two fundamental relationships which connect, respectively, spectral type and mass with luminosity.

The spectral type-luminosity relation (Russell diagram)

In 1913 Russell first produced the diagram which represents one of the most fundamental forms in which astrophysical data may be arranged. He arranged the luminosities of those stars for which data were available, in order of spectroscopic type. As soon as this was done, it became apparent that these two observables—type and luminosity—were intimately connected one with another. The nature of this interdependence is most clearly brought out when it is presented in graphical form. If a blank is prepared, the axes representing spectral type and luminosity respectively (see Fig. 64), and the position of each star marked with a dot—thus one whose type is G and absolute magnitude +5, would be represented by a point at the intersection of the horizontal line from +5 and the vertical line from G—then the disposition of these dots is not random. On the contrary, they cluster about two zones or lines, the first crossing the diagram obliquely from the top left-hand towards the bottom right-hand corner, the second branching off about two-thirds of the way up the first and thence running out towards the upper right-hand corner of the diagram.

Let us inquire more closely into the significance of this two-branched distribution. Stars at the top of the diagram are of high luminosity, irrespective of type, and those at the bottom of low; while from left to right (spectroscopic types B to M) the star becomes progressively redder, irrespective of luminosity. Treating first of all the main zone about which the points cluster—that crossing the diagram diagonally—we see that the redder a star is, the lower is its

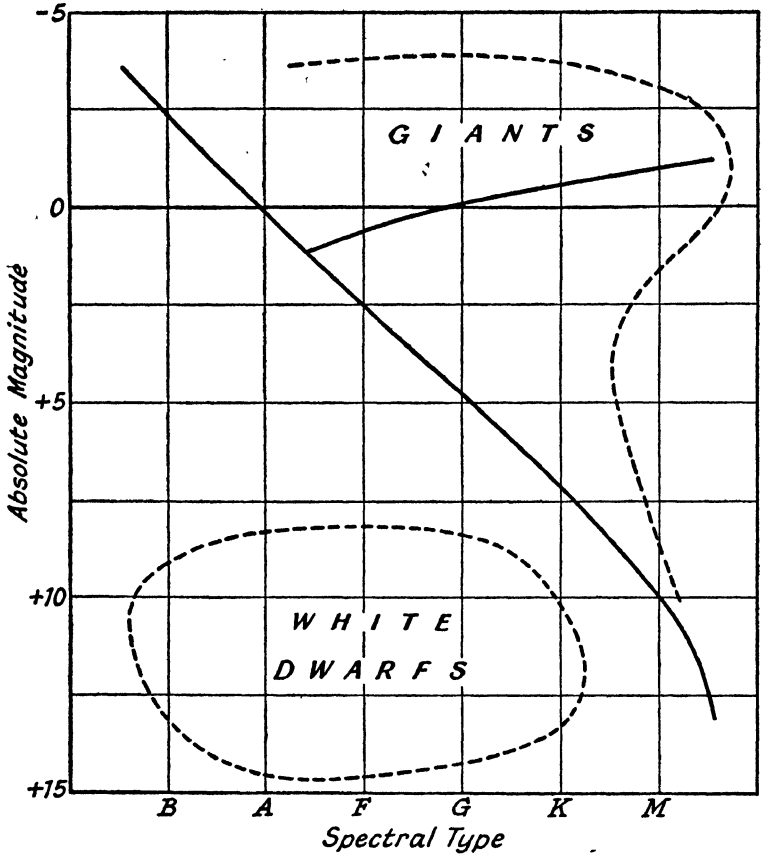


Figure 64. The Russell diagram.

luminosity. Reading off approximate values from the axes, we may derive the following values:

Type	*Absolute magnitude*
B	−2·5
A	+0·5
F	+2·5
G	+4·5
K	+6·0
M	+10·0

This primary zone is known as the main sequence, and stars which lie on or close to it as main sequence stars. About two-thirds of the stars whose luminosities and spectral types have been accurately determined lie within 1 magnitude of the curve which the plotted points define.

The second branch of the diagram is particularly interesting. The

plotted points of the Russell diagram are not in fact confined precisely to the two main branches, any more than the bullets of a good shot are confined exclusively to the bull's-eye, though they will be clearly concentrated upon this point. We have seen that in the case of the main sequence there is in fact an impressive concentration upon the curve defined by them. But in the second branch the crowding upon the central curve (marked, like the main sequence, by a continuous line in Fig. 64) is less clearly marked, and to a certain extent points occur over the whole area which is roughly indicated by the broken line, though progressively less frequently at greater distances from the central zone.

Let us consider what this distribution implies in terms of luminosity. All B-type stars are of much the same intrinsic brightness, the limiting absolute magnitudes being about -3 and $+1$. For each succeeding type, however, the range of luminosity increases; the extreme case is provided by the M-type stars, the vast majority of which (those belonging to the main sequence) have absolute magnitudes in the region of $+12$, though very much more luminous stars of the same type are infrequently encountered, up to absolute magnitudes of about -2. All the high luminosity stars are of course represented in the diagram by scattered points lying above the main sequence.

Thus, whereas red stars may be divided into two groups—those with high and those with low absolute magnitudes—this distinction becomes less and less clearly marked as we move back along the main sequence towards type B. Hertzsprung had drawn attention to the dichotomy among red stars as long ago as 1905, when he named the two groups 'giants' and 'dwarfs' respectively. But it was not until Russell investigated the relationship between spectral type and luminosity, eight years later, that this division was also applicable, though in a decreasing degree, to the yellow and blue stars.

The terms 'giant' and 'dwarf' have subsequently come to be used in a rather different sense, Hertzsprung's 'dwarfs' now being known as main sequence stars, the term 'dwarf' being kept for a special class of stars which we shall consider shortly, and the term 'giant' being generally applied to all stars of greater luminosity than main sequence stars of the same type, i.e. to those stars lying well above the main sequence on the Russell diagram, in the area so labelled (Fig. 64). As can be seen from the figure, the main zone of the giants branches away from the main sequence roughly at right angles to the luminosity axis: in other words, the giants, unlike the main sequence stars, are all of the same order of luminosity—several hundred times that of the sun. We shall learn presently that the term 'giant', though

originally based solely upon luminosity, is particularly appropriate in that these stars are also more massive than the stars originally termed 'dwarfs'.

White dwarfs

The third occupied area of the Russell diagram lies well below, and separated from, the main sequence, where a few scattered points record the existence of the peculiar class of star known as white dwarfs.[1] In opposition to the giants, the luminosities of these stars are lower than those of main sequence stars belonging to the same type. Only a handful of white dwarfs have so far been discovered: this, however, does not necessarily indicate that they are of rare occurrence, but rather reflects the fact that owing to their exceedingly low luminosities only those in the immediate vicinity of the sun are likely to be detected.

The subjoined table demonstrates their low luminosities and the apparent faintness which goes with them; all five stars are nearer than about 80 light years:

Star	*Absolute magnitude*	*Equivalent main sequence absolute magnitude*	*Visual magnitude*
Sirius B	+11·3	+2·5	+8·4
Procyon B	16·0	(2·5)	13·5
o Ceti B	7·5	0·0	9·6
40 Eridani B	11·2	0·6	9·7
van Maanen's Star	14·3	2·5	12·3

The investigation of the nature of the white dwarf, Sirius B, illustrates well the Holmesian procedure which the astronomer follows in discovering the mass, luminosity, density, size and other properties of a star. Sirius itself is a comparatively near neighbour, so that the parallactic method gives accurate results: its distance is 8·8 light years. Its spectral type is early A; thus it is a white star with a temperature (determined spectroscopically) of some 10,000°. Since we know its distance we can calculate the linear dimensions of the orbit of its companion from its angular dimensions. Thus we can discover that Sirius B revolves about its primary in an orbit of about the same size as that of Uranus about the sun; its period is forty-nine years. Hence it may be deduced that the combined masses of the two stars is 3·5 times that of the sun. Thus:

$$k.(m_1+m_2)=\frac{a^3}{P^2}$$

[1] Properly so called, in the modern terminology. The stars that used to be known as 'red dwarfs' are those lying at the bottom of the main sequence, i.e. main sequence M-type stars.

where k=a universal constant,
a=the mean separation of the components,
P=the period of revolution.

In the case of Sirius, observation gives $P=49$ years and $a=7''\cdot55$. This angular separation at a distance of 8·8 light years is equivalent to a linear separation 20·4 times as great as that between the earth and the sun. Hence, substituting in the equation,

$$\frac{(20\cdot4)^3}{49^2}=m_1+m_2$$

Or, $$m_1+m_2=3\cdot5.$$

As we have seen, an investigation of the absolute orbits of the two components—i.e. their orbits about their common centre of gravity—is necessary to reveal the individual mass of each component. That of Sirius B turns out to be one-third as great as that of its primary; yet it is very much fainter, only emitting one-ten-thousandth as much visible radiation as Sirius A.[1] In other words, its mass is 85 per cent. that of the sun while its luminosity is less than $\frac{1}{3}$ per cent. The obvious conclusion to be drawn from this is that it is a very low temperature star, radiating much less intensely than either the sun or Sirius A. But it has been found that its spectrum is almost identical with that of Sirius A; that is, it is an A-type star, and must consequently have a considerably higher temperature than the sun. This being so, the only way to account for its faintness is to assume that it is very small. Calculation based on this assumption shows that it cannot have a diameter of more than about 25,000 miles. It is thus a body whose volume is about twenty-seven times that of the earth, and diameter about three times that of the earth. Yet its mass is 250,000 times the earth's. From which it follows that its density is some 40,000 times as great as that of water: 1 cubic inch of the matter of Sirius B would weigh a ton. Even this fantastic figure is surpassed by some other white dwarfs: the type F dwarf known as van Maanen's star has an estimated density of several million times that of water, while the density of another, discovered by Kuiper, is estimated by him to be in the region of 40 million times that of water!

The position of the red dwarfs at the end of the main sequence appears to be secure, but the place of the B, A and F white dwarfs in the stellar evolutionary process is unknown. All that can be said is that they resemble the red, main sequence stars as regards size and mass, while differing from them in the matter of their higher temperatures and therefore whiter colour.

[1] Though it is an F-type star, its absolute magnitude is only 11·3, some nine magnitudes fainter than main sequence stars of the same type.

The mass-luminosity relation

The second important relationship which has been found to exist between the observable stellar characteristics which we have already reviewed, is that between mass and luminosity. Once a sufficient number of binaries had been studied, it was possible to treat their data statistically, and in 1924 Eddington pointed out that if the masses and luminosities of these stars were plotted against one another, the resultant pattern was a smooth curve. Hence we learn that the mass and the luminosity of a star are in some way interdependent. Roughly speaking, the relation between them is such that if the masses of two stars are in the ratio 2 : 1, their luminosities will be as 10 : 1. Mass and luminosity are directly proportional to one another: a star whose absolute magnitude is $-2{\cdot}5$ will be about twelve times as massive as the sun; one whose absolute magnitude is $+4{\cdot}8$ will be equally massive as the sun; and one whose absolute magnitude is $+10$ will have only about one-third the sun's mass. This relation holds good irrespective of spectral type and of temperature: thus, for example, a star of absolute magnitude 0·0 will be about four times as massive as the sun, no matter whether it be a white, type A star of the main sequence or a reddish, K-type giant.

Combining this new knowledge with the already ascertained relation between spectral type and luminosity illustrated in the Russell diagram, it is clear that stellar mass suffers a continuous decrease along the length of the main sequence from B to M, whereas among the giants it maintains, with luminosity, a tolerably steady value.

One notable exception to the mass-luminosity relation is the white dwarfs. These are uniformly too massive for their luminosities. As an instance of this, we have already seen that the mass of Sirius B is more than three-quarters that of the sun, while its luminosity is less than three-thousandths of the sun's.

Had the investigation of stellar mass been confined to binaries, only a comparatively small amount of data could have been accumulated. But it was nevertheless sufficient to open up, by disclosing the mass-luminosity relation, a very much wider field. Once the relationship has been expressed in graphical form, it is only necessary to drop a perpendicular from the appropriate point on the luminosity axis to the curve for the mass to be read off with a probable maximum error of 20 per cent. in the case of any star, binary or otherwise, whose luminosity is already known. This welcome extension of the field of operations substantiates the conclusion drawn from work in the more restricted arena of binary stars: that the masses of the vast majority of stars are not startlingly different from one another, differing at most by a factor of 50.

Spectroscopic parallax

From what we have just learnt of the mass-luminosity relation we should expect that there might be found some distinctive spectroscopic differences between the highly luminous giants and the comparatively dim 'dwarfs' of the main sequence, particularly the lower end of the main sequence. For whereas the masses of the giants have been shown to be of the order of fifty times those of the dwarfs, we have already seen that their radii are some hundreds or even thousands of times as great. Hence the force of gravity at the surfaces of the highly rarefied giants will be, compared with the dwarfs and despite the giants' greater masses, infinitesimal. We have seen in Chapter VI that ionization is favoured not only by an increase of temperature, but also by reduced pressure. It follows, therefore, that a more advanced state of ionization should be expected in the atmosphere of a giant than in that of a dwarf of the same spectroscopic type; and that this should be reflected in the spectra of the two stars.

Such has actually been found to be the case: briefly, as the luminosity of a star increases, certain lines in its spectrum increase in intensity, while others weaken. Adams and Kohlschütter in 1914 constructed curves relating luminosity (absolute magnitudes) with observed intensities of these crucial lines. Thenceforward these curves could be utilized to deduce the distance of any star bright enough to yield a spectrum that could be distinctly photographed: from the observed condition of the crucial lines its absolute magnitude could be read off the curve, and a comparison of this with its apparent magnitude led straight to its distance by means of the formula,

$$M = m + 5 + 5 \log p.$$

This, one of the most surprising and extraordinary developments of recent astrophysics, has proved a powerful weapon in the hands of the astronomer. Thousands of spectroscopic parallaxes have now been determined, many of which would never have been discovered had the only available method been that of trigonometrical parallax. Nevertheless, the resources of spectroscopic parallax have to-day been rather exhaustively explored, and the spotlight has shifted to other of the methods, of wider application, described in Chapter IV. Within its limits—and it cannot be profitably applied to the upper half of the main sequence—it is reliable; indeed, beyond a certain distance (some 65 light years) it is more accurate than the photographic method. For whereas the margin of inaccuracy in the latter method widens with increasing distance, the error in the spectroscopic is constant and independent of the smallness of the parallax, being a slight indefiniteness in the calibration of the curves themselves,

which results in a uniform uncertainty of about 0·5 absolute magnitudes. This is equivalent to an error in the derived parallax of about 20 per cent. The percentage error to be expected has been arrived at by determining the spectroscopic parallaxes of stars, other than those employed in the calibration, whose trigonometrical parallaxes were already known; the correspondence between the two was found to be reasonably close.

The sun as a star

Once we have arrived at some idea of the range of stellar sizes, luminosities, temperatures and other characteristics, the conclusion is forced upon us that the sun is a very ordinary member of the stellar hierarchy: it cannot even boast the inverted distinction of being abnormally small, dim or in any other respect undistinguished. The essence of the sun's stellar status is contained in the simple statement that it is a G-type star of the main sequence, i.e. it is situated about midway between the blue giants on the one hand and the red dwarfs on the other: all other features follow from that.

The sun's mediocrity can best be demonstrated by tabulating the maximum and minimum normal values for the different stellar characteristics, adjusted in each case to bring the solar value to unity (the figures are necessarily approximate):

Characteristic	*Maximum*	*Minimum*
	(*Sun*=1)	
Temperature . .	5	0·4
Luminosity . . .	50,000	0·000002
Mass	10	0·1
Linear diameter . .	500	0·01
Density . . .	400,000	0·0000003

Variable stars

One further stellar characteristic must be mentioned—variability. A very large number of stars vary in apparent brightness: this may be intrinsic in the star—i.e. its luminosity is variable—or may be extrinsic. In the latter case the star cannot truly be considered as variable; such stars have already been mentioned in connexion with the determination of stellar mass: they are the eclipsing binaries.

Extrinsic variables: eclipsing binaries

The types of variation exhibited by different variable stars are diverse, but several general types may be distinguished. The

simplest way to study the behaviour of a variable is to note its magnitude at a series of intervals, spaced according to the speed of the variation, and then to construct a curve whose axes designate Time and Magnitude. This curve will be a direct representation of the star's changing brightness, and its magnitude at any moment during the period covered may be read off at will; Fig. 65 is an example, and from even a cursory glance at it a considerable amount of information can be derived about β Persei, the star in question. The variation being continuous, and the duration of minimum short, the eclipse must be partial, and not central as shown in the figure; were the eclipse central (that is, total or annular) the magnitude at minimum

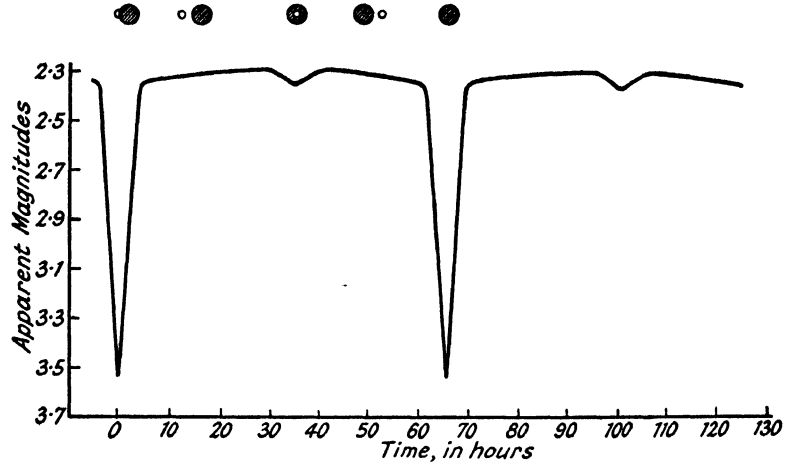


Figure 65. Light curve of β Persei.

would remain fairly constant for an appreciable duration before mounting towards maximum again. Hence it is possible to deduce the approximate inclination of the orbit to the line of sight. A further point to be noticed about the light curve of β Persei is that the magnitude at maximum is not steady. This means that the brightness of the system (i.e. of one or both stars) must vary in some way outside eclipse. This is most probably effected by the distortion of the spherical figures of one or both stars as a result of their mutual proximity: the effect of the rotation of non-spherical bodies is superimposed upon that of their mutual eclipses.

Intrinsic variables: Cepheids

Eclipsing variables—other than those which are distorted by tidal forces—cannot be regarded as true variables, for the actual light-output of each star is constant. Such variations as are observed from the earth are due solely to the relative positions of the sun and the

two stars in space. But there are certain types of variable whose light curves do not permit of interpretation along these lines. These stars are not binaries, but isolated stars, and the observed variation is intrinsic, involving not only apparent but also absolute magnitude. We must confine our attention to one only of the several groups into which such variables may be divided—the Cepheids.

Fig. 32 shows the light curve of δ Cephei, the type star after which the Cepheids are named. It exemplifies three of the most characteristic features of this type of star: the variation is continuous, the rise to maximum is quicker than the fall, and the decline is not represented by a smooth curve. Why Cepheids should behave in this manner is not known, but the theory that they are pulsating stars is regarded as the most likely explanation. A star in such a condition would undoubtedly be variable, while other considerations indicate that its variation might well be of the Cepheid type. For example, the spectroscope shows that Cepheids have fluctuating radial velocities. These, however, are not of a type that could be accounted for on the grounds of the mutual revolution of two stars, for the maximum brightness is attained while the line-of-sight motion is towards the observer, and is minimal when the radial velocity is positive.

The period of a Cepheid may lie anywhere between a few hours and about 50 days; those whose periods are less than twelve hours are, as we have already learnt, called cluster variables, although they are not entirely confined to the globular clusters. Their great intrinsic brightness is, from the practical viewpoint of discovering stellar distances, almost as important as the period-luminosity relation: a Cepheid of short period (say twenty-four hours) has at maximum a luminosity about 100 times greater than the sun: if the period is increased only so far as ten days, the luminosity leaps to about 1,000 times the sun's.

Novae

One further class of variable deserves mention. Occasionally it happens that, for no ascertained reason, a faint star suddenly blazes up, to shine with a brilliance that may outrival any other object in the night sky, only to die down again to its former insignificance. Such stars are called novae. The name dates from pre-telescopic days when novae really did appear to be new stars, for in almost every case they are too faint to be seen with the naked eye before the outburst. It is now known that we are concerned rather with the sudden brightening of an already existent star than with the birth of a new one, since the examination of photographs (where available, and of

sufficiently low limiting magnitude) taken before the outburst invariably establishes the fact that a faint star existed in the exact position of the nova.

Little is known of the early history of novae, for the reason that they are not usually noticed until well on the way to maximum brightness. This rise to maximum is extremely rapid, the peak usually being reached within a day or two (at most, several weeks) of its discovery; during this interval its brightness may have increased a hundred thousandfold. Maximum having been reached, the brightness at once begins to decline, rapidly at first and then more and more slowly; the original magnitude is reached, in typical cases, several years later. This decline, many times more gradual than the meteoric leap to maximum, is usually varied by considerable oscillations of brightness.

It is therefore possible to summarize the history of a typical nova as follows. A sudden and spectacular rise to a prominence which may outshine all its neighbours and be the wonder of the night sky; Lundmark calculates that at maximum the average nova is 25,000 times more luminous than the sun. The glory is short-lived, however, and gradually the star slips back towards its former inconspicuous level; a few years later it is not even visible to the naked eye. Such a story as this might well be told as a cautionary tale to astrologically-minded dictators.

Novae do not occur with equal frequency in all regions of the star sphere, but tend to congregate about the plane of the galaxy. Of thirty prominent novae whose positions and histories are accurately known,

16 lie within 10° of the galactic plane,
23 .. .. 20° .. ..
28 .. .. 30° .. ..
29 .. .. 40° .. ..
30 .. .. 50° .. ..

These figures show a sharp falling off of numbers in zones progressively further from the plane of the Milky Way. Furthermore, novae tend to occur near the edges of the Milky Way rather than in its more central regions.

What can be the cause of the cataclysmic conflagration that the typical nova must represent? Four main lines of approach have been employed in an endeavour to answer this question. The first assumes the existence of binaries with very long periods, and also highly eccentric orbits and close periastron passage; that is, the two components pass extremely near to one another once in each period of

revolution. It is argued that at periastron we may suppose great tides to be raised in each star, possibly with the formation of a third body; at any rate, it is assumed that tidal filaments would be wrenched from the bodies of the two stars. The weakness of this suggested explanation is that in order to produce anything like the appearance of a nova, the mutual approach of the two components at periastron would have to be very close indeed, and it is doubtful whether binary systems possessing at the same time highly eccentric orbits and a sufficiently close periastron passage do in fact exist.

A more general hypothesis on the same lines supposes that when any two stars—not necessarily the components of a binary system—approach closely or possibly collide, the resultant cataclysm would have all the appearances of a nova. While this contention may very well be true, it is difficult to believe that all novae are caused in this way. For, as we have seen, it is possible to arrive at some sort of idea of the density of stars in space; and this, combined with our knowledge of stellar motions, indicates that such stellar conjunctions must be too infrequent to cause all novae, bearing in mind their known rate of occurrence. (Though there may not be more than half a dozen bright novae in the course of a century, faint novae, discovered by the examination and comparison of photographic plates, are of comparatively frequent occurrence.) Furthermore, it is difficult to see how, after such an interaction between two stars, stable conditions could be restored in the space of a few years.

The third theory supposes that a star—dark or luminous—is carried into a region occupied by dark nebulous material, and that it causes a conflagration in much the same way that a meteor is heated to incandescence during its passage through the terrestrial atmosphere. This theory can claim to furnish an explanation of the preference shown by novae not only for the Milky Way (which is the zone both of the diffuse 'galactic' nebulae and also of the obscuring medium) but more specifically for the edges of the Milky Way. It also gives an explanation of some of the features of the complex spectroscopic history of the typical nova. It does not, however, explain all the observed spectroscopic changes and it is furthermore extremely unlikely that there exist any nebulae of sufficient density to cause the outburst.

It is the fourth hypothesis that is regarded with the greatest favour by astronomers to-day. This suggests that the nova is a—perhaps abnormal, perhaps common, perhaps inevitable—stage in the life-history of every star, and results from changes in its internal constitution and structure. A rapid expansion of the internal material of the star, amounting to an explosive outburst, is hypothecated. This

may or may not be the result of a sudden temperature rise, but in any case the outer and visible regions of the star would be 'blown out' very rapidly without at first rising appreciably in temperature. This would explain certain of the spectral changes associated with the rise to maximum which indicate high negative line-of-sight velocities combined with little or no alteration of temperature as indicated by change of type. At maximum the star would begin to shrink again, while its temperature would rise steadily, perhaps as far as the 50,000° mark. This shrinkage associated with increasing temperature would continue until, some years after the initiation of the outburst, the star would be a white dwarf.

Stated briefly and baldly in this manner, the hypothesis may sound like pure romancing, an *ad hoc* fitting of the facts, without regard to probability or even possibility. Nevertheless, it does appear to follow from the mathematical investigation of the internal structure of stars that such a course of events may well be a stage—even a necessary stage—in the evolution of each individual star. The whole matter is still in an uncertain state, however, and few astronomers would be bold enough to state categorically that they know what causes a star to burst forth as a nova.

Excursion into speculation

It was stated in the Preface that in this book would be found a systematic presentation of demonstrable and established facts, involving a minimum of controversial topics and such as require the reader to accept statements on trust. To the present point this claim has been faithfully substantiated.[1] The subsequent pleasure derived from breaking them is, however, the main justification for making good resolutions.

The allied problems of stellar energy production and stellar evolution are still subjects for controversy and speculation. Something should nevertheless be said of the latest developments in this field, though their treatment will necessarily be summary; it must also be recognized that the results mentioned in these last sections are very far from final.[2]

Sources of solar heat

We learnt earlier in the present chapter that the sun pours $1{\cdot}35 \times 10^6$ ergs of radiant energy on to every square centimetre of the earth's surface each second; this involves the total radiation from the sun of

[1] With one or two lapses: as, for instance, the supposed origin of the lunar craters and ring plains.

[2] Those who wish to pursue the subject further are recommended to refer to *The Birth and Death of the Sun* by George Gamow (Macmillan, 1941).

$1{\cdot}2 \times 10^{41}$ ergs per year. The problem of accounting for this very high rate of energy emission becomes more pressing when it is learnt[1] that the sun has already been keeping it up for something like $1\frac{1}{2}$ to 2 thousand million years. For it follows that its total radiation to the present day is of the order of $2{\cdot}4 \times 10^{50}$ ergs, or $1{\cdot}2 \times 10^{17}$ ergs per gram of its mass.

Three theories have in the past been proposed to explain whence the sun gets this great reserve of energy:

(*a*) By simple burning. This is the obvious explanation: the sun is burning up its substance in the same way that, for example, a coal fire transforms its fuel into ash. But a coal sun would burn out in some 5,000 years, and we know that it has in fact been 'burning' for at least 1,500,000,000 years. Clearly, some other explanation must be sought.

(*b*) By contraction. When a gas is compressed, as every user of a bicycle pump knows, its temperature rises. If the sun were once a much larger, less dense, and cooler body than it now is—a giant, in fact—its own gravitation would have caused it to contract. Helmholtz, about the middle of last century, suggested that the heat of the sun had derived from gravitational contraction in this way. Yet this explanation, though an improvement on the combustion hypothesis, can only account for one-thousandth of the $2{\cdot}4 \times 10^{50}$ ergs which represent the sun's actual radiation.

(*c*) The third suggested source of the sun's remarkable store of energy is certainly the correct one, though advances in this field are of recent enough date to have been impeded by the war, and the details are certain to be considerably modified in the future. In essence, it is suggested that the sun's energy is subatomic; that is to say, it is derived from changes within the very nuclei of the atoms of which it is composed.

Nuclear disintegration[2]

By 1919 the phenomenon of radioactivity, or the spontaneous transformation of the atomic nuclei of one element into those of another, was a commonplace, and had been the subject of much investigation since its discovery by Becquerel twenty-three years previously. But in that year Rutherford changed the outlook of physics

[1] By means of the 'radium clock'. Radioactive elements in the earth's crust disintegrate at a known rate. By measuring the proportionate amounts of the end-product of the process (lead) and of the remaining active element, it is possible to gain a rough idea of the period during which the process has been going on: the earth's crust solidified some $1{\cdot}6 \times 10^9$ years ago.

[2] Written before the advent of the 'atom bomb', which throws some of these remarks into rather a lurid perspective.

by realizing the mediaeval alchemists' ambition of *artificially* transmuting one element into another. Into a chamber containing ordinary air (which consists largely of nitrogen) he projected a stream of high-velocity α-particles—the nuclei, bereft of their orbital electrons, of the element helium. One of Rutherford's α-particles achieved a head-on collision with the nucleus of a nitrogen atom, and smashed it, with the resultant formation of two nuclei, one of oxygen and one of hydrogen. Since that epoch-making experiment, the new science of nuclear physics has progressed rapidly, and to-day many dozens of nuclear reactions have been promoted in the laboratory.

Two characteristics of such reactions are particularly to be noted:

(*a*) They are characterized by enormous liberations of energy, a typical nuclear disintegration releasing many thousand times more energy than even so violent a molecular reaction as the combustion of T.N.T. Whereas 1 gram of coal, on complete combustion, liberates only 3×10^{11} ergs, 1 gram of a mixture of lithium and hydrogen, entering into the nuclear reaction which produces helium, would liberate $2{\cdot}2 \times 10^{18}$ ergs of subatomic energy—10 million times as much!

(*b*) Nuclear reactions are extremely difficult to promote, since only rarely does a direct collision between a nucleus and a projectile—α-particle or proton—occur. And the stability of nuclei is such that nothing but head-on collisions with extremely powerful projectiles has any disruptive effect upon them. Hence the *total* energy liberated during an experimental bombardment is of microscopic proportions. Were it otherwise, the physicist, his laboratory and half the county as well would be blown sky high with a force never dreamed of by Bomber Command.

Thermonuclear reactions

Whereas a sun composed of coal would go out in about 5,000 years, a sun deriving its energy from nuclear reactions would have ample reserves to continue radiating for the thousands of millions of years which we require. But how is it that elements like nitrogen and carbon will enter into nuclear reactions in the sun, while they will not, except on the most niggardly scale, in the laboratory?

In 1929 this question was answered by Atkinson and Houtermans with their theory of thermonuclear reactions. Eddington has shown that the sun's temperature reaches a figure of about 2×10^7 degrees at the centre. Under such conditions, nuclear transformations would be greatly facilitated, for the individual nuclei (and free electrons) would themselves be travelling with velocities comparable with the projectiles of the physicist's laboratory—the component particles

of a gas becoming more and more violently agitated as its temperature rises. Thus instead of there being, as in the laboratory, comparatively few projectiles with high enough kinetic energy to shatter nuclei, all the nuclei themselves become 'bullets' capable of disrupting one another. Nuclear reactions would then be possible which in the laboratory would require millions of years to complete. Such reactions are termed thermonuclear reactions.

Once started, thermonuclear reactions will generate enough heat to keep themselves going indefinitely, without the application of any further energy from an external source.

The solar reaction

Having established the fact that conditions in the solar interior are such as would permit the nuclear transformations of the common, stable elements, and the further fact that these reactions would constitute a rich enough source of energy to provide the $2{\cdot}4 \times 10^{50}$ ergs which we require, it remains to identify the precise reaction or reactions that are responsible for the sun shining at the present moment. And here we are treading on to very unsure ground.

The energy liberation of each individual reaction can easily be calculated for any given temperature, and this figure compared with the observed energy production rate of the sun. In 1938 Bethe and Weizsäcker independently hit upon the reaction—or, more correctly, chain of reactions—that would produce the sun's $1{\cdot}2 \times 10^{41}$ ergs per year. Bethe concluded that the thermonuclear reactions proceeding within the sun are six in number, and form a closed chain whose first and last stages are the same. The cycle is thus repetitive, and will continue until the raw materials used up in the reactions are exhausted. Its net result is the building up of one helium nucleus from four hydrogen nuclei, energy being liberated in the process.

This energy does not, of course, appear from nowhere. According to relativity theory, mass and energy are equivalent terms, so that the 'destruction' of mass involves the liberation of the equivalent amount of energy. Now the mass of four hydrogen atoms is slightly greater than that of one helium atom. This lost mass reappears as energy, according to Einstein's equation linking mass and energy ($E=mc^2$).

Bethe concludes, therefore, that the sun is deriving its energy from the synthesis of helium nuclei out of hydrogen nuclei. Hydrogen is the 'fuel' and helium the 'ash' of the solar thermonuclear production of energy.

Energy production in main sequence stars

The question is at once suggested, do the other main sequence stars derive their energy from the same reaction as is responsible for the sun's energy production?

Just as the physical conditions at the centre of the sun can be calculated, so it is possible to discover the central pressure and temperature of any star whose mass, total radiation, and radius or surface temperature are known. It is also possible to calculate the energy production of the solar thermonuclear cycle at different temperatures, and it is found that the figures derived agree well with the observed luminosities of stars in the centre and upper regions of the main sequence.

But among the low temperature red dwarfs the solar reactions would be retarded or inhibited by the comparative thermal slowness of the protons. Here it is possible that the thermonuclear interaction occurs between protons themselves: three hydrogen nuclei combining in two stages to form a helium nucleus (α-particle) with the liberation of energy.

Energy production in the red giants

In the still cooler red giants, thermonuclear reactions would be virtually at a standstill. Gamow and Teller suggest that during the early stages of a star's existence it is a giant with large diameter and low density, and that gravitational contraction is the sole source of heat. Contraction would continue until the internal temperature had risen to a level at which the easiest and most rapid thermonuclear reactions are possible (about one million degrees). The generation of subatomic energy within the star would thereupon put an end to contraction, and a self-regulating, thermostatic mechanism would come into operation.

As soon as the supply of each element is exhausted—first those most easily induced to enter into thermonuclear reactions, and then those progressively more recalcitrant—gravitational contraction would set in again, and would continue until the central temperature were high enough to permit the reaction involving the next heavier element.

The temperature of a giant will therefore rise continuously from the onset of the first reaction. In other words, it will pass backwards across the Russell diagram, traversing the Harvard sequence in reverse order; eventually it will reach the main sequence where, as we have seen, the temperature is high enough for the solar cycle to begin operation.

Passage along the main sequence

On reaching the main sequence, the star's short youth is over, and it is envisaged as entering upon its long period of maturity.

It would be natural to suppose that as its hydrogen supply is depleted, it will grow cooler and pass *down* the sequence towards the red dwarf region. This, however, is not so, for the cloud of helium nuclei generated by the process is sufficiently opaque to radiation to prevent the internally produced radiation from escaping to the stellar surface whence it may be radiated away from the star. Thus the star's temperature and luminosity will increase with increasing rapidity, the star meanwhile passing *up* the main sequence.

The last hydrogen nuclei will ultimately be converted into helium, and the final contraction will begin. The star will shrink very rapidly, with a parallel falling off of luminosity, and since it has no further effective energy-producing reserves to offset this, it must leave the main sequence.

Calculation indicates that the sun's luminosity will increase until it surpasses its present value by a factor of about 100. Luminosity and size will then decrease rapidly, as the sun changes over from thermonuclear to gravitational energy production. Thus it would appear that the fate of the earth is to be burnt to a cinder, and not, as once thought, to be slowly frozen.

The end of stellar evolution

Once the possibilities of thermonuclear energy production are exhausted and the final contraction supervenes, the star ceases to belong to the main sequence: its decreasing luminosity means that its position on the Russell diagram is lapsing towards the lower edge: it enters the region occupied by the white dwarfs.

White dwarfs, then, with their small diameters and disproportionately great masses, are suggested as representing the last few tottering steps to stellar extinction.

IX

THE NEBULAE

Absorbing material within the galaxy

IT will be recalled that the problem of plotting out the extent of the star system by assigning distances to such galactic objects as the open clusters was complicated by the hitherto unsuspected existence of a rarefied absorbing medium. This material pervades the galaxy and has the effect of stretching distance determinations which are based on apparent brightness. Two lines of approach led to this conclusion: first, Trumpler's work on the sizes and distances of the open clusters; and secondly, the marked avoidance of the galactic plane by the globular clusters and extragalactic nebulae. To remove the discrepancies following from the neglect of this factor, Trumpler was forced to assume an optical density of about 0·8 magnitudes per 3,250 light years for the absorbing medium. He was further led to conclude that, although extending for considerable distances in the galactic plane, the medium's thickness measured at right angles to this plane was no more than a few hundred light years.

Not only did interstellar absorption cause the dimming of the light from distant objects, but it also implanted colour-excesses upon the most remote sources of all, provided that these were within about 10° of the galactic plane. This not only indicated clearly that the absorbing layer was thin, but also that it was partly gaseous in composition. Different investigators of the characteristics of the interstellar absorption have reached curiously various conclusions. Seares, for example, deduces from his star counts that absorption of the type outlined by Trumpler cannot occur; a similar conclusion was reached by Elvey from different evidence; while Hubble and Humason have shown that the few extragalactic nebulae situated near the galactic plane exhibit little or no trace of general galactic absorption.

These divergent results may probably be reconciled on the assumption that although diffuse material concentrated in the galactic plane does effect absorption, yet it is far from homogeneous, probably being more accurately thought of as a patchy, flocculent structure, with areas of higher and lower density, than as a uniform medium.

Galactic nebulae

So far we have only considered indirect evidence for the existence of galactic absorbing material. There is, however, a great deal of

direct, visual evidence also. The stars—isolated, in binary and multiple systems, and in clusters—and, perhaps, a few planets here and there, are not the sole constituents of the stellar system. There are, in addition, nebulae of various kinds. Nebula (Lat. cloud) was the name given to these objects by the early astronomers as being descriptive of their appearance, and the first class that we shall consider are in fact vast, structureless clouds of apparently glowing gas. Figs. 39 and 40 show two of these nebulae, the first being situated in the constellation Cygnus, the second being that which involves the stars of the Pleiades. These two photographs illustrate clearly some of the more characteristic features of the diffuse, irregular or galactic nebulae. They have no definite structure or shape; they are of vast proportions (long exposure photographs have shown that the nebula in the Sword of Orion, visible to the naked eye, extends over a great part of the constellation); and they are intimately associated with stars, which, as can be seen from Fig. 40, are not merely superimposed upon the nebula or seen through it, but are actually involved in it. Thus the distances of many of the galactic nebulae may be determined from the distance of the stars embedded in them.

Radiation from galactic nebulae

At one time it was thought that these bright galactic nebulae shone by virtue of their own incandescence, and even that the involved stars might have been born from them by some process of condensation. Both beliefs have now been discarded, our ideas concerning the nature of the diffuse nebulae having undergone a complete revolution within the last fifty years. In 1912 it was discovered spectroscopically that the Pleiades nebula is not self-luminous, as had previously been believed, but shines by the reflected light of the involved stars; without these stars it would not be incandescent and would therefore be invisible. In one circumstance only it might be indirectly visible: were it projected against a rich part of the galaxy its presence would perhaps be deduced from the absence of stars in that region.

Later it was shown by Hubble that both the spectroscopic type and the visible size of any galactic nebula are determined by the temperature and luminosity of the involved stars; the higher the temperature and the greater the luminosity of the stars, the greater will be the illuminated expanse of nebulosity. Furthermore, no bright nebulae exist which do not contain stars of sufficient brightness to account for their visibility. Another observation which points to the same conclusion, that the more luminous stars are capable of rendering this type of nebula visible, is that faint stars are noticeably absent

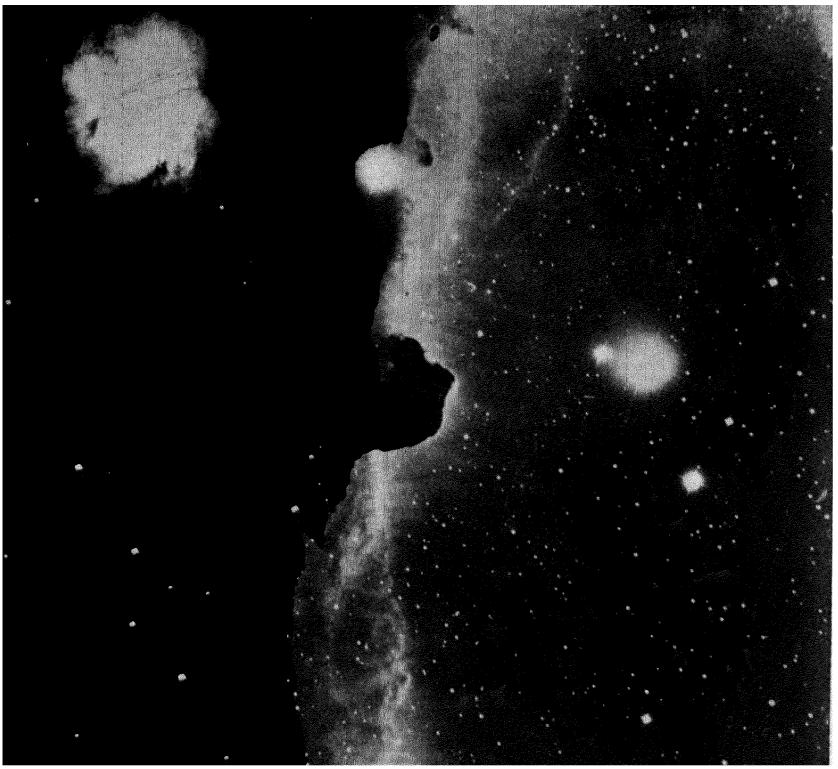


Figure 66. Dark galactic nebulosity: the Horse-head Nebula in Orion. (*By courtesy of the Director of the Science Museum, South Kensington, London.*)

from those regions occupied by high temperature stars and diffuse nebulae. This is believed to indicate that while the brighter stars are capable of illuminating the nebula, the fainter stars are either at too low a temperature to do so, or are situated at very much greater distances from the sun than the nebula; in either case the nebulous material would obscure them from the sight of the terrestrial observer.

Galactic nebulae and involved stars

The theory that the involved stars are born from galactic nebulae has suffered a complete reversal, and instead of believing that an inward movement of nebulous material towards a number of points resulted in the formation of stars at these points, modern astronomy favours the view that the radiation pressure of the high temperature stars is scattering the nebulous material away from them in all directions. After what has been learnt of the nature of electromagnetic radiation in Chapter VI, it should not come as a surprise to learn that radiation possesses momentum and therefore exerts a pressure upon any body, surface or particle that either reflects or absorbs it. The force of expulsion depends upon the temperature of the source of the radiation; the higher the source temperature, the more violent the repellent force. It is the radiation pressure of the sun, acting upon the particles in the tail of a comet, which forces this tail always away from the head, with the result that when the comet is receding from the sun it proceeds tail foremost.

Two conflicting forces, therefore, are centred in the massive 'early type' stars which are found to be associated with diffuse nebulae: gravity, acting inward towards the stars' centres, and the repellent radiation pressure. The mechanism of radiation pressure being known, it is possible to calculate the relation between these two forces for stars of the types concerned, and it is found that in their immediate vicinity the radiation pressure outstrips the gravity many times.[1] Hence any nebulous material in the neighbourhood of the involved stars would be swept away towards regions devoid of the hottest B- and A-type stars. Internal motion has been detected in some of the diffuse nebulae spectroscopically, and so far as it goes (for these motions are not large) the evidence does not conflict with the conclusion arrived at through the theoretical consideration of radiation pressure. The spectroscope also shows that the diffuse nebulae are for the most part almost stationary; when allowance has been made for the component of the velocity shift due to the sun's

[1] At the surface of the sun, which is a comparatively cool star, radiation pressure amounts to 65,000 tons per square mile; at the distance of the earth this is reduced to 2·6 pounds per square mile.

motion, their spectra, considered statistically, are not appreciably displaced.

Distances and distribution of the galactic nebulae

As already mentioned, their distances can be deduced in those cases where the distances of the involved stars are known. Some at least of the diffuse nebulae are near neighbours of the sun, the Pleiades nebulosity being only some 325 light years distant, and that in Orion about 600 light years; some of the fainter, on the other hand, are certainly very much more remote, and it must be supposed that they exist far out beyond the limits to which our telescopes can reach.

Their alternative name, galactic nebulae, derives from the fact that, like the open clusters, they occur most frequently in and near the galactic plane. The majority are to be found within 10° of the galactic plane itself. It will be remembered that this is the zone of avoidance of the globular clusters.

Spectra and composition of the galactic nebulae

Characteristic lines in the spectra of the galactic nebulae were for many years not matched by any produced in the laboratory. It was, therefore, supposed that an unknown gas, which was named 'nebulium', was present in the nebulae, and that it was through the excitation of the atoms of this gas that the unknown radiations were produced. But more recent knowledge of the structure of matter and of the table of elements has rendered this explanation highly improbable; it is now known that the lines of 'nebulium' do not indicate the presence of an unknown element, but rather that of familiar elements existing under unfamiliar physical conditions. Extremely low pressure—representing densities far below the minimum attainable in the laboratory—would be such a condition, and it is even possible to arrive at an idea of the mechanism whereby the 'forbidden' lines are produced.

Under normal conditions of pressure, the atoms of a gas suffer many thousands of collisions every second; even at the most extreme rarefaction that the physicist can achieve, the average interval between inter-atomic collisions is only 1/1,000 of a second. In the gaseous nebulae, however, it is estimated that this interval is increased to from 10^4 to 10^7 seconds. Any transition that requires an appreciable time for its consummation, therefore, would be impossible at normal pressures and densities, since interference from another atom would always occur first; but under the conditions prevailing in the gaseous nebulae this would not necessarily be the case, and the

atom would be left undisturbed for a long enough interval to complete the transition. These atomic states, from which transitions can only occur at long intervals, are called the metastable states, and Bowen has shown that they are at the back of the unknown radiations from galactic nebulae once thought to be due to 'nebulium'. In fact, they are caused by doubly ionized nitrogen and oxygen, and trebly ionized oxygen. Other, familiar lines in the spectra of the diffuse nebulae were long ago identified as belonging to the spectra of hydrogen, helium, carbon, and neutral nitrogen and oxygen.

Spectroscopically, therefore, the diffuse nebulae fall into two classes: those with gaseous emission spectra consisting of bright lines only; and those, like the Pleiades nebulosity already mentioned, which have continuous spectra with superimposed dark lines, similar to those of the involved stars. This difference is attributed to the involved stars rather than to the nebulous material itself: for it is found that where the stars are of the hot early types (O and B), the gaseous spectrum is produced; where the stars are cooler than type B, continuous stellar-type spectra are produced. In the former case, the gaseous material is being excited to incandescence in a way perhaps similar to that productive of aurorae; whereas in the latter, the nebula is simply shining by the reflected light of the involved stars. This discovery in turn throws light upon the probable constitution of the nebulous material. For the emission spectra must depend upon the presence of isolated atoms, while the capacity of the cloud to reflect starlight argues the existence of larger particles than atoms. In other words, the diffuse nebulae are partly gaseous and partly dusty or meteoric—a conclusion in complete agreement with those of Trumpler which, it will be remembered, involved the existence of large particles (total absorption) and small particles (causing the observed colour-excesses).

Dark nebulae

If the diffuse nebulae do not shine by virtue of their own unassisted incandescence, we might expect to find certain nebulae which, being situated near no high-temperature stars, are dark. Such nebulae, did they exist, would necessarily be confined (so far as human observation is concerned) to the galactic zone, since nowhere else is there provided a bright background against which they might be silhouetted. In 1919 Barnard published the first catalogue of 182 dark nebulae, one of which is illustrated in Fig. 66. Their existence is an integral part of the theory that diffuse nebulae do not shine by their own light. The occurrence of dark nebulae had been known for over one hundred years, but until comparatively recent times they were

regarded as starless holes running through the stellar system and aligned upon the earth. Once the frequency of their occurrence had been demonstrated by Barnard, the highly improbable nature of this explanation became clear, and the theory of obscuring clouds was substituted. Like the bright diffuse nebulae, they are not very distant, some being only a few hundred light years from the sun and none that are visible being more distant than, probably, about 1,000 light years. (The distance can be gauged when, as often happens, they are associated with bright nebulosity and high-temperature stars.) This does not mean that they are necessarily restricted to the comparative vicinity of the sun—on the contrary, they are without a doubt distributed through the whole stellar system: but a dark nebula situated at a much greater distance than 1,000 light years would be difficult to detect since the obscuration of still more distant stars would be negatived by the superposition upon it of nearer stars.

These diffuse nebulae, both bright and dark, are certainly of exceedingly low density, as shown by their limited powers of absorption despite enormous extent. Though more rarefied than even the most perfect vacuum that can be produced in terrestrial laboratories, they must nevertheless be regarded as localised areas of exceptional density in the general absorbing stratum which pervades the whole galactic plane.

Interstellar calcium

Mention must be made of two other types of nebula encountered in the stellar system. The first of these is not a nebula in the usual sense of the term, for it is not visible (even by superposition upon a bright background) and appears to pervade the whole of galactic space: it is, indeed, something very like the absorbing medium that Trumpler was forced to hypothecate. Its existence went unsuspected until 1904, in which year Hartmann noticed that prominent calcium lines in the spectrum of the binary δ Orionis (two stars revolving about one another at too small a distance for them to be distinguished as separate stars) did not share the oscillations of the rest of the spectrum, caused by the orbital motion of the brighter component about the centre of gravity of the system. Similar stationary lines have since been detected in the spectra of hundreds of high temperature white stars, and in some cases the lines of sodium, titanium and potassium are also stationary. Slight displacements of the stationary lines towards the red or violet are accountable for by the sun's motion, and when this effect is eliminated it is found that the residual displacement is in the majority of cases negligible. The only possible explanation is that between the star and the observer there is a small

amount of calcium, sodium, potassium, titanium, and possibly other elements as well, in the form of an extremely tenuous interstellar cloud, and that it is absorption by this matter which implants the stationary lines upon the spectra of certain stars.

The spectra of all binaries do not contain Hartmann's lines. In general, two conditions must be fulfilled; the star must be distant, and it must be of an early spectral type, usually O or B. The first condition is readily understandable, for a great thickness of such rarefied matter (it is estimated that the calcium cloud cannot consist of more than one atom per cubic inch) would be necessary for an absorption of detectable strength to be caused. Their apparent preference for the spectra of the early-type stars is presumably due to the fact that even were they present in the spectra of stars of later type than about B5, they would be obscured by the heavy lines that normally occur in these.

Planetary nebulae

The third type of intragalactic nebula is the planetary. Fig. 67 shows that these objects are quite unlike the diffuse nebulae. They are small, round, well-defined, and might almost be mistaken for planets when seen with small telescopes. They are comparatively rare objects, fewer than 200 being known, and are all invisible to the naked eye; this is due both to their faintness and to the smallness of their discs, the majority of which are not more than a few seconds of arc in diameter. Their apparent smallness is, however, the result of distance rather than of linear insignificance, for van Maanen has shown that the diameters of over twenty planetaries are not less than several thousand times that of Pluto's orbit.

In small instruments even those planetaries with the largest angular diameters show no detail; they appear as faint, round discs of pale greenish or bluish light. But instruments of larger aperture reveal the existence of a faint central star in many of them. The spectroscope shows that they resemble the galactic nebulae in two important respects, however dissimilar the two classes of object may appear at first glance. The planetaries are gaseous, and they shine by reflection (or at least by excitation) of the central star; this is typically of the high temperature O-type. That they are not the flat discs they appear to be, but are spherical, is indicated by two observational facts; in no instance is any foreshortening of the disc observed, while it is proved by the Doppler effect that at least some of them are rotating axially.

The spectroscope has been instrumental in providing almost all our knowledge of the planetary nebulae. The spectrum is of the type

which we saw in Chapter VI to be typical of glowing gases; the bright lines of hydrogen are invariably present, those of helium usually, and those of nitrogen frequently. The temperature modifications of those lines show without a doubt that the temperature of the central star cannot in the majority of cases be much below 50,000°. Yet the luminosity of these stars (which can be deduced as soon as the distance is known) proves to be low. Considering their temperature therefore they must be very small.

The spectroscope has also revealed that the planetaries differ from the diffuse galactic nebulae in the matter of velocity. It will be remembered that the spectra of the latter showed, at most, small velocity shifts. The radial velocities of the planetary nebulae, on the other hand, are abnormally high; in extreme cases they may reach 125 m.p.s., the average being in the neighbourhood of 20 m.p.s.

The planetaries follow the same distributional pattern as novae, open clusters and galactic nebulae. That is, they crowd about the galactic plane and are only rarely encountered at any considerable distance from it. This is at any rate true of the fainter,[1] and in general more distant, planetaries; the brighter and angularly larger are found outside the galactic region, but if we conclude that their relative size and brightness indicate nearness to the earth, then their real divergence from the galactic plane will naturally appear to be greater than it is. That the faint planetaries are in general more distant than the brighter is also indicated by the fact that they occur most commonly in that region of the galaxy with which we are by this time becoming somewhat familiar—the Sagittarius region where lies the galactic centre.

Extragalactic nebulae: summary of distribution

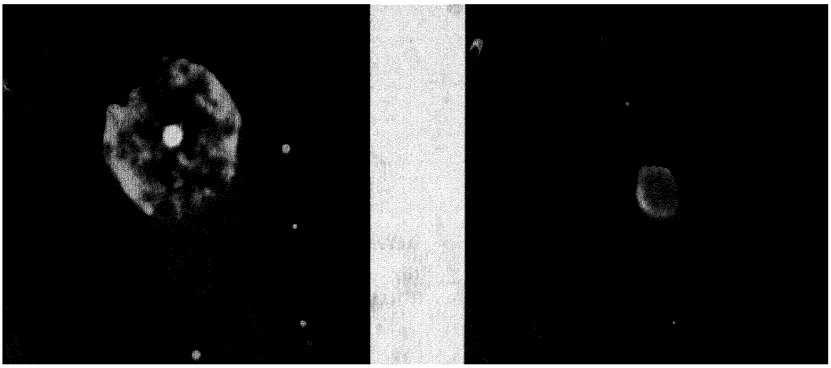
Whereas in Chapter V we treated the extragalactic nebulae rather as anonymous counters in the game of probing the universe to the very limit of human inquisitiveness, we are now better qualified to inquire more closely into their nature. First, however, let us briefly summarize what we have already learnt of their spatial distribution:

i. The extragalactic nebulae visible with present equipment occupy a volume of space stretching from less than 100,000 light years to 5×10^8 light years distant.

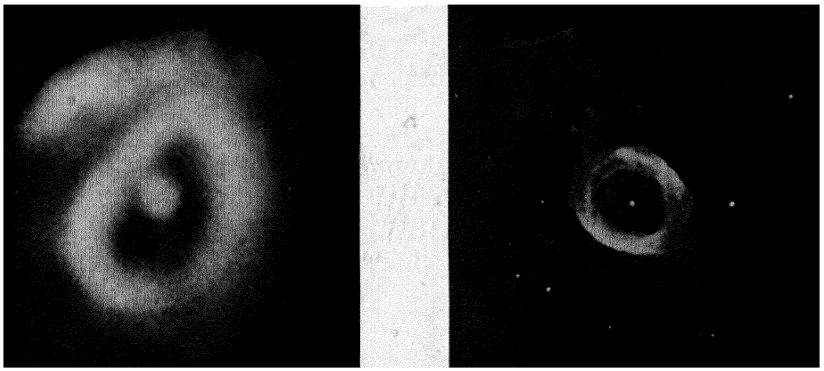
ii. It is estimated that within this region there are some 10^8 nebulae.

iii. Throughout the whole of the observable region, the large-scale distribution of the nebulae is uniform.

[1] I.e. apparently, not intrinsically. Some of the angularly smallest are in fact more luminous than larger (and hence, probably, nearer) examples.



N.G.C. 1501 N.G.C. 2022



N.G.C. 7662 N.G.C. 6720

Figure 67. Four planetary nebulae: N.G.C. 1501, 2022, 7662 and 6720. (*By courtesy of the Director of Mt. Wilson Observatory.*)

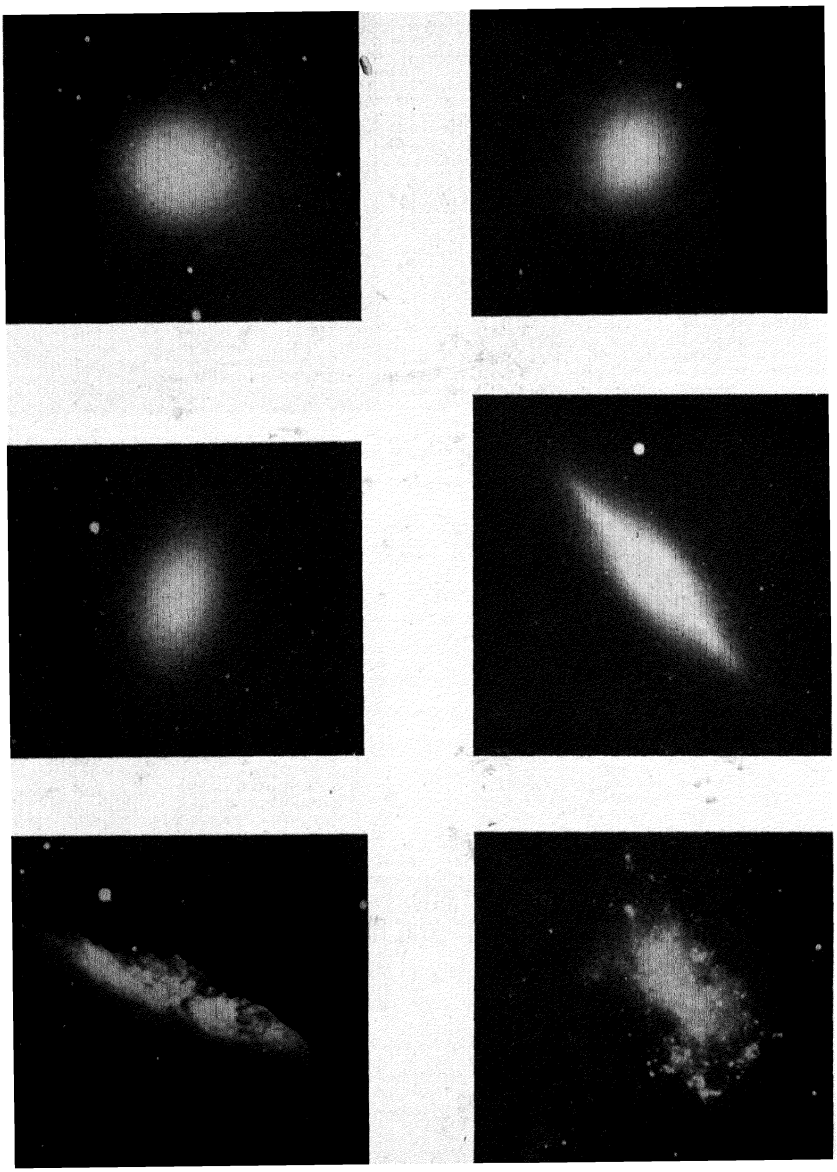


Figure 68. Elliptical and irregular nebulae. (*From* The Realm of the Nebulae, *by Edwin Hubble. Yale University Press.*)

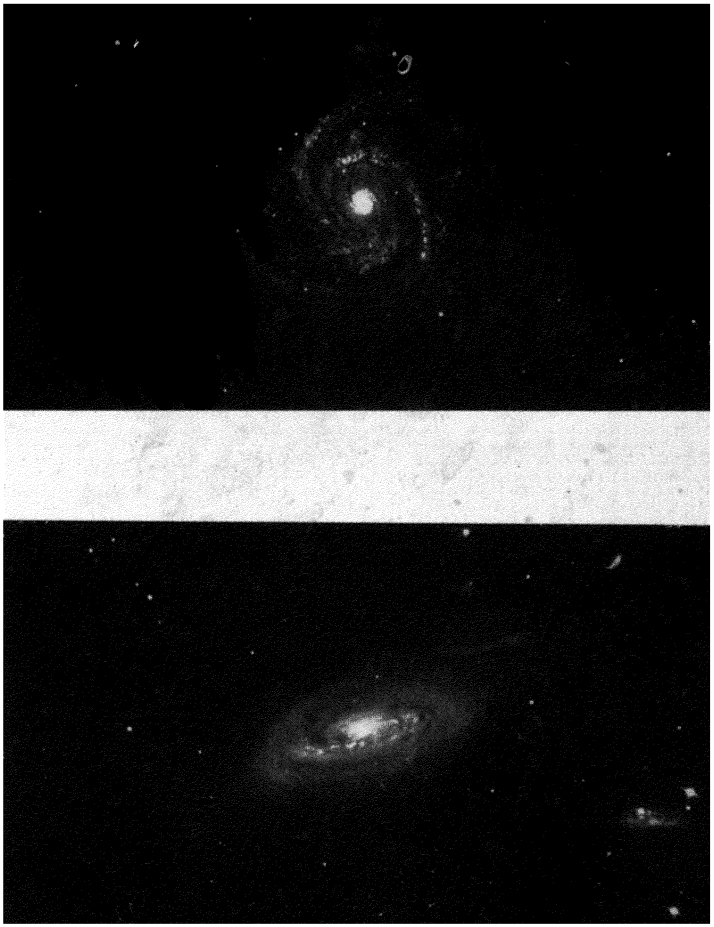


Figure 69. Normal spiral nebulae. (*From* The Realm of the Nebulae, *by Edwin Hubble. Yale University Press.*)

Figure 70. Spiral nebula viewed from a point near the projection of its central plane. (*By courtesy of the Director of the Science Museum, South Kensington, London.*)

iv. Their small-scale distribution is, however, irregular, consisting of single nebulae, small groups and great clusters. The nearest of these clusters, that situated in Virgo, is about 7×10^6 light years distant.

v. No falling off of numbers towards the outer edge of the observable region can be detected, indicating that we have not yet fathomed the system of extragalactic nebulae.

vi. The red-shifts in nebular spectra, if interpreted as velocity shifts, indicate that the more distant a nebula, the greater its velocity of recession.

Classification of the extragalactic nebulae

The extragalactic nebulae are for the most part extremely faint objects, and only one, that in Andromeda, can be detected easily with the naked eye; on clear moonless nights it appears as a faint misty spot resembling a nebulous star. When studied with telescopic cameras the extragalactic nebulae are found to present a number of closely allied forms, and a classification based upon the systematic differences displayed by the several hundred brightest nebulae has been evolved; for only these few are angularly large enough for anything of their structure to be made out, the vast majority of the white nebulae showing on photographic plates as formless specks, hardly distinguishable from faint and ill-defined stars. Among the more conspicuous extragalactic objects, however, two main classes are distinguishable, the second of which may be further sub-divided into two groups:

1. Irregular,
2. Regular: (*a*) globular, or to a greater or lesser extent elliptical;
   (*b*) spiral: (i) normal,
   (ii) barred.

Irregular nebulae

Irregular nebulae need hardly detain us, since they comprise only about 2 per cent. of the total. The Magellanic Clouds (see Fig. 41) are prominent because near examples of this type of system, and about one half of all known irregular extragalactic nebulae are generally similar to them: they show no traces of rotational or other symmetry, and are largely resolvable into stars, star clusters and clouds, and irregular nebulosity of the galactic type. Without doubt they are stellar systems.

Elliptical nebulae

Members of Class 2 (*a*) are to be observed in a graded variety of forms, but all differ from the irregular nebulae in three important

respects: their rotational symmetry, their lack of resolution, and the existence of a bright nucleus. This variety of form is due to a combination of two factors; different inclinations to the line of sight, and real differences in shape. Visually they vary from round discs to elongated, cigar, or spindle-shaped forms (Fig. 68) whose axes are related to one another as about 3 : 1. Nebulae exhibiting every gradation of flattening between these limiting forms are known.

Now if these nebulae were truly discoidal (as two plates, rim to rim) these varying appearances would result merely from different inclinations of their major axes to the line of sight: if viewed edge-on they would appear cigar-shaped; if directly from the prolongation of the minor axis, round; and intermediately, elliptical with a greater or lesser degree of flattening. But a sufficiently large number of these nebulae are known to treat them statistically, and by this means it has been found that more of the circular nebulae occur than the law of averages would warrant. Some of the round nebulae must, therefore, not be discoidal and viewed from a point on the projection of the minor axis, but truly spherical; thus they would appear spherical from whatever direction they might be viewed. Once this was established, it was realized that many of the intermediate types might be ellipsoidal (viewed edge-on) rather than discoidal (viewed from a point between the projections of their median plane and minor axis). Thus we have a transition from a spherical nebula, a globe of gas, through increased polar flattening to a discoidal nebula which may appear either cigar-shaped or circular, according to the angle of vision.

Spiral nebulae

The spiral nebulae resemble the elliptical in being 'flat'; that is, their diameters are very much greater than their thickness. Normal spirals consist of a central, discoidal nucleus, from diametrically opposed points on whose periphery project two arms. These arms leave the body of the nebula approximately tangentially, and the nebula, when seen from a point on the projection of its minor axis, is reminiscent of a Catherine Wheel (Fig. 69). That it is a flat structure may be seen in those cases where the nebula is viewed edge-on (Fig. 70). In some spirals the two arms appear to have bifurcated with the formation of a four-arm spiral.

About one spiral in three is not of this normal type, but falls into the category of 'barred' spirals. Fig. 71 shows an example, and it can be seen that the arms instead of issuing direct from the nucleus as in normal spirals, issue from the outer ends of a bar of matter in which the nucleus is centrally placed.

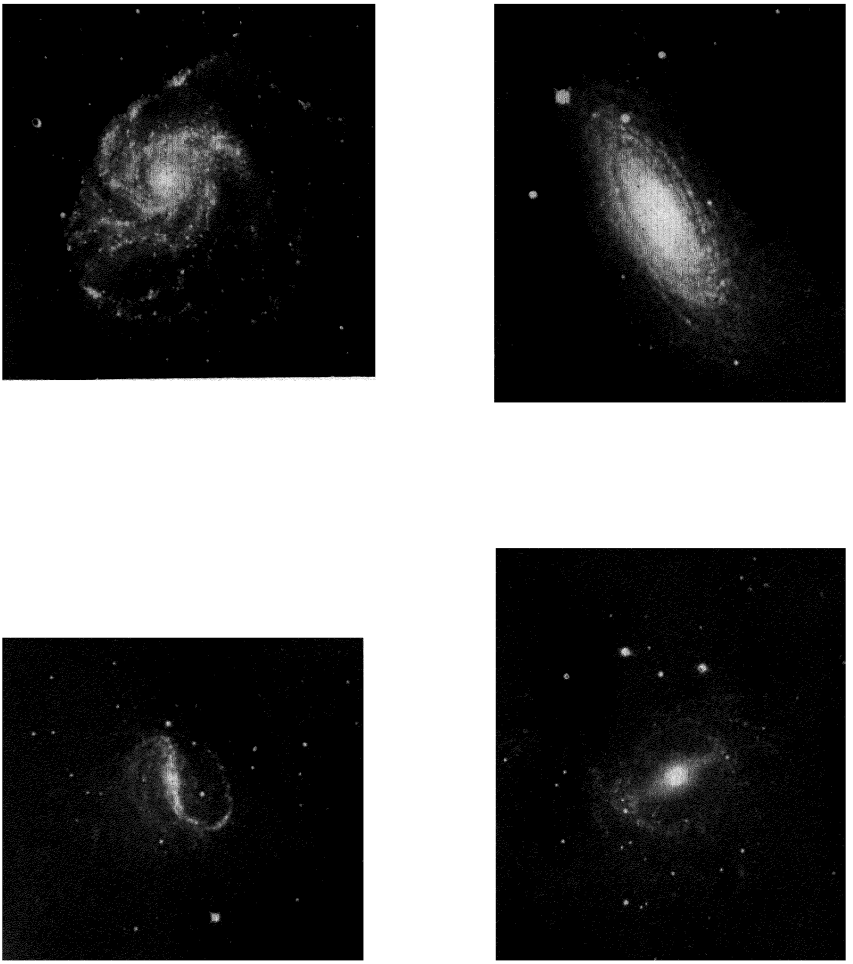


Figure 71. Normal and barred spiral nebulae. (*From* The Realm of the Nebulae, *by Edwin Hubble. Yale University Press.*)

Just as a sequence of forms—from spherical, through increased polar flattening to the critical 3 : 1 figure—is observed among the elliptical nebulae, so a formal sequence, which in all probability also represents a temporal sequence, is exhibited by the spiral nebulae, both normal and barred. Treating the sequence as temporal—i.e. as one through which each nebula passes in the course of its existence—we may describe the successive changes as follows. At the beginning of the sequence the nebula is hardly distinguishable from a 'late' elliptical nebula. The arms are inconspicuous and small compared with the main body or nucleus of the nebula. A steady outward transference of matter from the nucleus to the arms results in the latter becoming increasingly more massive at the expense of the former. Simultaneously the arms uncoil. About half way along the sequence the homogeneous texture of the nebulous matter begins to break down: condensations start to form, at first sparsely, in the outermost regions of the arms; these increase in number and spread inward to the nucleus, which by this stage will have shrunk to a mere shadow of its former state. In the final stage the nebula consists entirely of widely opened arms (the nucleus having disappeared), which in turn consist of clouds and clusterings of faint star-like points, interspersed with unresolved material which may either be nebulous or may consist of stars too faint and too numerous to be individually visible. Another feature of late spirals which is reminiscent of our own galaxy is the common occurrence of patches and blotches of dark obscuring material, presumably analogous with our dark nebulae. Obscuring matter is also to be observed among early spirals of the normal type; here it takes the form of a peripheral band encircling the nebula in its median plane; this formation is consequently only visible in the case of nebulae which are oriented edge-on to the observer (Fig. 70).

The extragalactic sequence as a process

During the course of the modifications of the early globular type whose end point is the late spiral, the brightness of the nebula does not alter to any material extent. Its size, on the other hand, does. How this interesting fact has been established will be explained at a later stage. As the nebula passes along the sequence its diameter grows steadily from (taking average and only approximate figures) about 6,000 light years to 30,000 light years; this extension in the median plane is of course the necessary concomitant of the reduction of the axis perpendicular to it as flattening proceeds.

The most striking feature of the extragalactic objects, when classified in this way, is the uniform nature of the changes to which they

are subjected throughout the sequence: starting as undifferentiated globular masses, they are progressively flattened; when a certain critical stage in this process is reached a new factor emerges, and the formation of arms begins; more and more matter is transferred from the body of the nebula to its arms, which at a second crucial point begin to break up into star-like condensations. The largely resolved, wide-open spirals are then only differentiated from the irregular nebulae by the traces of rotational symmetry that still remain. It is impossible not to conclude that we are dealing with different members of a single family, all of which conform to a basic pattern which is modified systematically from the beginning to the end of a limited sequence.

The rotation of gaseous masses

Added significance is attached to this conclusion by Jeans' theoretical investigation of the behaviour of a rotating gaseous mass, and the bearing of this work upon the question of the size and nature of the extragalactic nebulae. It is supposed that the mass is far enough from other bodies to be gravitationally insulated. It will then assume a spherical form which rotation will flatten at the poles. Under stress of its own gravitation it will shrink, and with increasing shrinkage its rotation will increase in accordance with the law of the conservation of angular momentum. This in turn will cause an increasing flattening of the poles and bulging of the equatorial regions. When a certain stage in the accelerating rotation is reached, the mass will become unstable; in the same way a flywheel that is allowed to rotate too rapidly will shatter into fragments. Jeans has shown in another connexion that a liquid mass on reaching this critical stage of instability will split in two. The gaseous mass, however, behaves differently. It will now be discoidal with a comparatively sharp periphery. As the rotational velocity increases, matter will be ejected from two diametrically opposed points on the circumference. This matter will issue as jets and will be of low density, since the heaviest constituents would have gravitated to the centre of the mass, leaving the less dense in the upper levels. With always increasing shrinkage and rotation, more and more of the substance of the nucleus will be forced into the arms; these will gradually assume a spiral configuration as a result of rotation. Thus, starting with a perfectly spherical body of gas, we reach a stage when it is transformed into a discoidal nucleus of decreasing mass and spiral arms of increasing mass. The more rarefied matter in these arms will tend to condense into separate agglomerations until they have been transformed into discontinuous systems of knots, clusters, and single condensations. Thus the more

condensed the arms the smaller will be the nucleus, since both processes start and run concurrently. Eventually the nucleus will have been completely sapped, the arms will have become entirely condensed, and the original gaseous mass will have been transmogrified into a heterogeneous spiral system of isolated masses.

The observed and theoretical sequences compared

The various stages in this process are strikingly reminiscent of the different types of extragalactic nebula that are to be observed. In order of time, on the analogy of the theoretical sequence, we have the spherical nebulae, slightly flattened nebulae, and discoidal nebulae; all these are, so far as telescopic resolution is concerned, uncondensed. Next we have nebulae with large discoidal nuclei and homogeneous arms, followed by all stages of the transference of matter from the nuclei to the arms; along this sequence it is noticed that progressive condensation of the arms, with the formation of star-like points and clusterings, is occurring. Even in details the correspondence between the theoretical and observed sequences is exact; for instance, the rotation of the nebulae, the outward movement of matter from the nuclei along the arms, the inversely related size of the nucleus and degree of condensation in the arms have all been established observationally.

Whether the problem is approached from the observational or theoretical direction, therefore, it is tolerably certain that the fundamental difference between the various types of extragalactic nebula is one of age.

Spectroscopic evidence

The spectroscopic observation of the extragalactic nebulae is rendered difficult by their faintness, which forbids the use of the high dispersion prisms required to produce a spectrum long enough for detailed study. Not only are the spectra necessarily short, but spectroscopic examination is confined to the brighter central regions of the brighter nebulae. The majority of these yield spectra of a stellar type—a continuous background upon which are superimposed numerous fine absorption lines—and most commonly of type G, that of the sun. A minority of spiral nuclei have spectra of quite a different type: the bright-line emission spectra, similar to those of the planetaries and some diffuse nebulae, which are indicative of incandescent gases under low pressures. Owing to the practical difficulties just mentioned, qualitative analysis of extragalactic spectra has not progressed far, but calcium, iron, and hydrogen have been identified in a number of nebulae. So far as the evidence goes,

therefore, the spectroscope indicates that even the apparently undifferentiated nuclei of the majority of spirals are not gaseous masses but star clouds. The first step towards the direct photographic resolution of a spiral nucleus (M. 31) was made by Baade in 1944, an observational achievement of first importance.

The nature of the red-shifts

We have seen how the discovery of the red-shifts provided astronomers with an additional distance criterion—or, rather, with a welcome confirmation of results obtained by the use of other criteria. The red-shifts are an objective phenomenon, and the linearity of the relation between distance and spectral displacement is unimpeachable. So long, in fact, as the displacements are merely accepted as spectral shifts to the red, and no questions asked, all is plain sailing. Inevitably, however, the nature and cause of the shifts became the subject of speculation. When first detected and measured, they were unquestioningly accepted as velocity shifts of the normal, familiar type. But as Humason pressed on his investigations among more and more remote nebulae the immense velocities[1] encountered threw the gravest suspicion on the validity of the theory seeking to explain the red-shifts in terms of the Doppler phenomenon. It is verging on the incredible that such velocities as have already been measured are real, yet what alternative explanation is there of the observed shifts? The question is of the most fundamental importance, for upon it depends the nature of our whole world picture and a great deal of theoretical cosmology. The question is still unanswered, although some preliminary clearing of the ground has been achieved.

At present no agency other than recession is known which could produce the red-shifts. If, therefore, they are not velocity shifts, they are a reflection of some physical principle of which we are ignorant. Although this alternative is of altogether too problematical a nature for unquestioning acceptance, the phenomena force us to explore its possibilities. We saw in an earlier chapter that the energy and corresponding wavelength of a quantum were related in the manner

$$E\lambda = C$$

where E is the energy, λ the wavelength, and C a constant. If the wavelength is increased—i.e. the spectrum is shifted towards the red—the energy of the radiation is correspondingly decreased. This loss of energy attendant upon a red-shift may be envisaged as occurring primarily in one of two ways. Either the wavelength is

[1] It must be borne in mind that these velocities are purely geocentric, i.e. relative to the terrestrial observer.

increased, as it would be were the source in motion away from the observer, which would result in the familiar Doppler or velocity shift; or else the energy itself might be reduced in transit from the source to the observer. If such a decrease of energy could be effected, the spectrum would be shifted towards the red, and the natural deduction would be that the source had a positive radial motion. It is at this point that we have to fall back upon the mysterious 'unknown' as an explanation, for we are ignorant of any mechanism whatever that could effect the leakage of energy during transit without causing parallel effects which would easily be observable, but which are in fact absent from nebular spectra.

Nor is it possible to differentiate between a red-shift due to a lengthening of the wavelength by movement of the source and a shift due to leakage of the energy carried by the quanta as they traverse inter-nebular space. For while it is true that a source in rapid recession would appear fainter than an equally luminous stationary source at the same distance, we utilize apparent luminosities as the only distance criterion capable of reaching to the most distant nebulae where these effects are of considerable proportions. Until, therefore, a distance criterion independent of apparent luminosity can be developed, no certain answer can be given to the crucial question: Do the red-shifts really represent recession?

For all these reasons the shifts are commonly referred to, outside the popular Press where sensation is rated more highly than truth, either as red-shifts or as apparent velocity shifts; and where they are called velocity shifts, without qualification, it is with the unspoken proviso that this is a convenient and obvious term until more specific contrary evidence is forthcoming.

Linear sizes of the extragalactic nebulae

It has already been stressed that although luminosity remains sufficiently constant all along the sequence from globular to irregular nebulae to be used as a criterion of distance, this is not true of linear diameter; for if a spherical object is subjected to a process of flattening, it must expand in a direction at right angles to the axis of flattening: thus a rubber ball, compressed between two plates, will expand laterally between them. In the same way it might be expected that the extragalactic nebulae will exhibit linear diameters which are a function of their degree of flattening. This is in fact the case. As soon as distances, and therefore linear diameters, of a representative collection of each type of nebula had accumulated, it became apparent that the globular nebulae had the smallest diameters and late spirals the largest, intermediate types having diameters appropriate to

their positions between these two extreme types. The ascertained facts regarding the linear diameters of the extragalactic nebulae may be summarized as follows:

	Globular	*Late elliptical*	*Early barred spiral*	*Late normal spiral*	(*Irregular*)
Diam. (L.Y.)	6,000	16,000	18,000	31,000	(21,000)

These figures are only approximate and, if anything, are on the low side: the true figures may well be twice, or even three times, as great. The reason for this is that the longer the photographic plate is exposed, the larger grows the image of the nebula, showing that the outer regions are not only faint but extensive. These figures refer to what is rather vaguely known as the 'main body'—that part of the nebula which is visible on any well-exposed plate.

Masses of the extragalactic nebulae

Similarly approximate figures may be derived for the masses of the nebulae. One method was developed by Öpik and depends upon a single datum: the shift in a nebula's spectrum at a given distance from the nucleus, due to rotation. If a nebula is viewed edge-on, and is in rotation, one side will be approaching the observer, and the other receding from him (see Fig. 73). Hence one edge of the nebula will show a Doppler shift indicating recession, the other a shift indicating approach. The nebulae exhibit such unmistakable signs of rotational symmetry that even were such shifts not measurable, we should still be perfectly justified in postulating their rotation about their minor axes. But the spectroscope proves this to be a fact by revealing the existence of these rotational shifts. Clearly the size of the shift, varying with the line of sight velocity, will depend upon how far out from the nucleus the spectroscope slit is adjusted. Knowing, in the case of a nebula whose distance has been determined, the linear diameter of the mass between the two positions of the slit, and also the velocity of rotation at a preselected distance from the axis, the laws of motion allow the calculation of the total mass within this limit. Since the outer regions of any extragalactic nebula are too faint for investigation, it follows that this method will yield results short of the true figures for the whole nebular mass. The masses of four of the brighter nebulae derived in this way are respectively (in terms of the sun) 1,000,000,000; 9,000,000,000; 30,000,000,000; and 35,000,000,000.

Mean density of matter in the universe

Knowing the volume of the observable region, the approximate number of nebulae in the region, and the average nebular mass, it is a simple matter to calculate the average density of matter in that region of space which we know. The result emphasizes the extreme emptiness of the universe: for if all the matter in the observable region were distributed uniformly through it, the density of the resultant medium would be between 10^{-28} and 10^{-30} grams per cubic centimetre, equivalent to one grain of sand distributed throughout a volume equal to that of the earth.

The galaxy and the extragalactic nebulae compared

The idea has for many years been entertained by speculative thinkers that the galaxy might itself be a late-type spiral. Our knowledge both of the extragalactic nebulae and of the galaxy has now reached a stage at which it is possible to give serious consideration to this hypothesis. In view of the facts that have already been set out in the preceding pages, let us compare the extragalactic nebulae, more especially the late spirals, with the galaxy, paying particular attention to the following characteristics of each: (i) rotation, (ii) mass, (iii) size, and (iv) content.

At the outset it might be said that perhaps the most attractive general feature of the hypothesis is that it introduces uniformity into the large-scale organization of the observable region of the universe. On the one hand we have some hundred million extragalactic nebulae, all closely related, and one other object of a different type, an unique system which also happens to be our home—a conception uncommonly like an aftertaste of Aristotelian geocentricity. And on the other, we have a universe stocked exclusively with nebular-stellar systems, all built to the same fundamental pattern.

i. Rotation

The discovery and subsequent measurement of the axial rotation of extragalactic nebulae naturally raised the question, Is the galaxy likewise rotating about its minor axis? If the reader goes out of doors on a starlit night, and, gazing up into the firmament, considers this question for himself, he will probably be forced to admit that, if the matter rested in his hands alone, we must for ever be ignorant of the answer. Yet the problem has been tackled and solved. We not only know that the galaxy *is* rotating, but we know approximately how long the sun requires to complete one galactic revolution, the direction of the galactic centre, and the direction of the sun's orbital motion at the present time.

The fact of stellar motion was already a commonplace in the eighteenth century, when the elder Herschel set himself to discover the direction and velocity of the sun's motion relative to its stellar neighbours. We have already learnt in Chapter IV that later studies of the proper motions and radial velocities of large numbers of stars situated in all regions of the star sphere have permitted the location of the solar apex and antapex, as well as its velocity towards the former and away from the latter relative to the stars chosen for the investigation. But what information does this discovery, itself a masterpiece of observation and deduction, provide about the supposed rotation of the galaxy? Unfortunately, none. For all the stars chosen as 'street lamps' were comparatively bright and near, and therefore members of the Local Cluster. This was the result neither of chance nor mistake: bright and near stars were deliberately chosen so that their spectra should be clearly visible and accurately measureable for velocity shifts, and also because a near star may be expected to have a larger proper motion than a remote one. Hence, although the derived velocity of 12 m.p.s. towards a point near Vega is likely to be accurate, it refers only to the motion of the sun within the framework of the Local Cluster, and not to the motion of the sun within the framework of the galaxy. The problem of what the Local Cluster itself is doing—whether or not it is rotating about the galactic centre—is not touched. The only way in which information concerning this rotation can be obtained is to use stars outside the Local Cluster, i.e. faint and distant stars.

If the Local Cluster, and with it the sun, is revolving about the galactic centre, non-cluster stars between the sun and the centre will be revolving more rapidly than stars situated between the sun and the edge of the galaxy.[1] Let us consider the case of the eight stars represented in Fig. 72. It is assumed that they are all galactic stars lying well outside the Local Cluster; also, for the sake of simplicity, that they have no random motions, their only motion being that imparted to them by virtue of the galactic rotation. In the case of each star the length of the arrow is proportional to the distance that it will be carried by galactic rotation in a given period of time. If a terrestrial observer now studies the motions of these stars, making the necessary allowance for the motion of the earth about the sun, he will get the following results. Stars 1 and 2 will possess no measurable proper or radial motions—they are stationary, relative to the sun. Star 3 also will possess no radial motion, but will, owing to its greater space

[1] This follows from Kepler's harmonic law, and assumes that there is a concentration of mass at the galactic centre. Subsequent work on stellar motions has indicated the correctness of this assumption: the galaxy does not rotate 'solid', like a flywheel.

motion, appear to drift forward. Similarly star 4 will drift backward, while showing no line-of-sight motion. Stars 5 and 6, on the other hand, will not only show a forward drift but will also have, respectively, negative and positive radial motions. In the same way, stars 7 and 8 will combine a backward drift with radial motion.

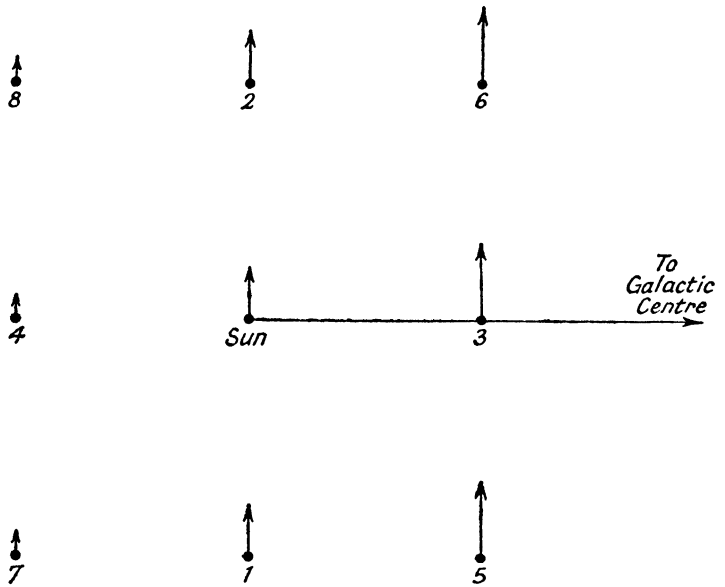


Figure 72.

The clearly defined differences in the observed motions of stars lying in different directions, assuming the galaxy to be rotating, are sufficient to provide the basis of a crucial investigation. If stars do in fact exhibit motions of this character, notably of drifting in two preferential directions, then we not only know that they are due to galactic rotation, but may also distinguish the direction of the galactic centre. Independent analyses of the proper and radial motions of different groups of remote stars by a number of observers have shown conclusively that such is the case; furthermore, they mutually agree in placing the galactic centre in the direction already indicated by Shapley as the result of his work on the spatial distribution of the globular clusters—towards Sagittarius. That this result is materially correct cannot be doubted: independent analyses of the motions of the bright O- and B-type stars, of Cepheids, and of planetary nebulae up to distances of 60,000 light years all yield the same result.

The galactic centre is located among the star clouds of Sagittarius at an approximate distance of 30,000 light years.

The sun, together with its near neighbours, is at present moving

towards the constellation of Cygnus. The velocity imparted to them by the galactic rotation is in the neighbourhood of 170 m.p.s. From these two facts it follows that the sun completes one revolution of its galactic orbit in about 220,000,000 years.

ii. Mass

Once these facts were established, it was further possible to gain a rough idea of the mass of the galaxy by means of the Kepler-Newton laws of motion. The average mass of a star being known, simple division gives the number of stars in the galaxy. Approximations being involved in each of the steps to this result, it is not to be wondered at that the results obtained should be discordant within fairly wide limits. They range from about 3×10^{10}, the more probable estimate of Seares, to $2{\cdot}7 \times 10^{11}$, that of Eddington; the estimates of other workers mostly fall between these limits. Considering the nature of the data, it is a matter for surprise that the agreement is so close. It seems probable, therefore, that the stars in our system are to be numbered in ten-thousands of millions, and perhaps in hundreds of thousands of millions.

The first two stages of our comparison of the galaxy with the late-type spirals are complete. Like the extragalactic nebulae, the galaxy is in rotation. Though only established in recent times, this conclusion has been suspected ever since the characteristic ellipsoidal form of the galaxy was demonstrated by Herschel at the end of the eighteenth century. And secondly, the galaxy and the only spirals concerning which we have data are comparably massive. The figures in each case are of necessity only approximate, but the correspondence is suggestive and probably significant.

iii. Size

The Cepheid determinations of twenty years ago established beyond cavil that the apparently largest and also the brightest—and therefore, as a safe bet, the nearest—spiral was extragalactic. But this conclusion was not without its attendant difficulties; and though all these have now been resolved, one of them is worth a brief glance. Not only had speculation long toyed with the idea that these objects might be extragalactic, but had constantly added the rider 'and stellar systems like our own', or, as they were often termed, 'comparable galaxies'. The first spirals to be plotted on a linear scale of distance all turned out to be considerably smaller than the galaxy: astronomers have never forgotten the salutary lesson that Kepler and Copernicus taught the believers in the unique status of the home of *homo sapiens*,

and any theory that sets our earth or our galaxy on a plane above the normal scheme of things is viewed with suspicion, and will only gain provisional acceptance as a last resort to save the phenomena. Yet if the spirals were as near as they were thought to be, and the galaxy as large as was thought, then certainly the galaxy was a giant among them. The distance determinations were accordingly suspect. This slur on the early distance determinations was, however, removed by two subsequent developments. First, it transpired that the estimations of galactic diameter were too large: its supposed diameter of

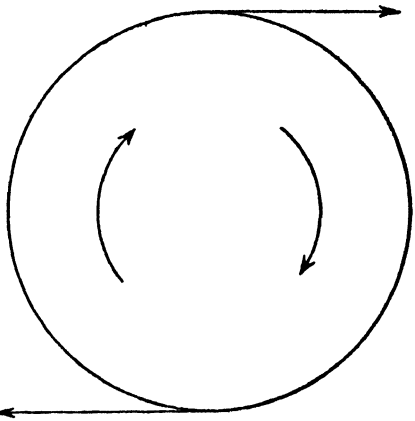


Figure 73. The rotation of an edge-on spiral (or of the sun) causes one limb to approach the earth and the opposite limb to recede from it. If the figure represents the sun (see Chapter VIII), the earth may be imagined as lying off the edge of the page at a distance of about 6 yards.

300,000 light years has been reduced, as a result of new knowledge about galactic absorption, to something nearer 100,000 light years. And at the same time the spirals have been shown to be larger than originally suspected. That the brightness of a spiral such as M.31 falls off gradually towards its perimeter has been demonstrated by the fact that the longer a plate is exposed, the larger the area of nebula that is recorded. But the degree to which this extension could be carried was not realized until extremely delicate photometric investigation with a photoelectric cell proved that the nebula is at least twice as large as is visually apparent. The combination of these revised estimates of the size of the galaxy on the one hand, and of the spirals on the other, has gone far towards removing the original discrepancy and has reduced the star system to a scale comparable with that of the late-type extragalactic nebulae. Though there is still considerable uncertainty regarding the precise size of the galaxy, the investigation has proceeded far enough to justify the confident statement that the spiral nebulae are indeed 'comparable galaxies'.

iv. Content

Having established these preliminary agreements between galaxy and spirals as regards rotation, size and mass, we can now proceed to a comparison of their respective contents, so far as these are known. And here some of the most impressive evidence for the identity of the two is to be found.

Stellar resolution in a considerable number of the nearer nebulae has been confined to the late-type spirals. A number of interesting facts concerning the stellar and other identifications are worth collecting together.

i. The luminosities (absolute magnitudes) of the brightest galactic stars and the brightest stars in the spirals are of the same order: so far as the evidence goes, galactic and nebular stars are built on the same pattern.

ii. Variables of precisely the same types as galactic variables are found in great profusion in the resolvable nebulae.

iii. Novae, which also occur in great numbers in the spirals, are similar to galactic novae in appearance and behaviour.

iv. Patches of bright diffuse nebulosity are of common occurrence in the spirals, as in the galaxy.

v. Dark nebulae likewise occur abundantly in both. Among spirals, it is most noticeable in the edge-on nebulae, where it appears as a peripheral band.

vi. Vast star clouds and even open clusters of the galactic type occur in the late spirals: in addition, the spectroscope indicates that in many cases the visually unresolvable nebular nuclei consist of star clouds, which, seen from a smaller distance, would no doubt resemble the clouds of our Milky Way.

vii. The peculiar distribution of the globular clusters has already been described: though certainly to be classed as galactic objects, they are for the most part situated outside the galaxy as defined by the stars, galactic nebulae and open clusters (see Fig. 38). In 1932 Hubble announced the discovery of 140 objects similarly situated with regard to the great Andromeda spiral. These are conjectured to be globular clusters, though the identification has not yet been definitely established. If they are, the discovery will provide still another, and very striking, detailed similarity between the galaxy and a representative extragalactic object.

viii. Finally, there is the close correspondence between the observed shapes of the spirals and the conjectured shape of the galaxy. Both the galaxy and a spiral are circular in plan and highly flattened in the direction of the axis of rotation. In each case the flattening is the result of this rotation.

The galaxy as a late-type spiral

All this constitutes an impressive mass of evidence that the spirals and the galaxy are identical types of object: that, in other words, our own stellar system is a late-type spiral. Reviewing this evidence, it becomes abundantly clear that the galaxy and the extragalactic nebulae are truly 'comparable galaxies'; that they differ from one another primarily as regards age, or degree of development along the sequence of observed forms; and it is highly probable that the galaxy is a late spiral. That it is not an early spiral is indicated by the prevalence of star clouds and, generally, the high degree of resolution reached. That it is not an irregular nebula, despite this advanced stellar resolution, is established by the evidence for rapid rotation, which must involve rotational symmetry—a characteristic which the irregular nebulae notably lack.

We may, then, envisage the galaxy as a lenticular, heterogeneous congeries of stars and other matter, probably between 50,000 and 100,000 light years in diameter. Its major axis is from seven to ten times as great as its minor axis, this highly flattened form deriving from its rapid rotation. The galactic centre lies among or beyond the star clouds in the Sagittarius region of the Milky Way, and at a distance of about one-quarter of the galactic diameter.

In addition to stars and star clouds, the system contains a vast amount of undifferentiated matter, partly gaseous and partly meteoric. Not only does irregular nebulosity, capable of the complete obscuration of radiation passing through it, abound in the median plane, but the whole system is permeated by an even more tenuous absorbing medium which is probably of a gaseous nature. An analogy may possibly be drawn between the peripheral absorbing matter seen in edge-on spirals, and that which is responsible for the great rift which splits the Milky Way into two streams between the constellations of Cygnus and Centaurus. The obscuring matter is concentrated towards the galactic centre and actually masks this centre from the terrestrial observer.

It is probable that the remains of spiral arms unwinding from the central regions would still be visible to an extragalactic observer. Not only do the studies of the extragalactic nebulae show the existence of such arms to be most probable, but Oort's complex and painstaking analyses of stellar distribution and comparative star density may be interpreted as revealing their actual existence. It is within them that the star clouds (including that containing the Local Cluster) would be chiefly located, the regions between the arms being characterized by considerably lower star densities.

Trumpler's three-dimensional model of the open cluster system

also reveals traces of spiral structure. In particular, the clusters in Auriga, Cassiopeia and Perseus appear to delineate a spiral arm, and Trumpler believes that the galaxy is, as viewed from the north galactic pole, a right-hand spiral.

This is as far as we can go with existing equipment. All the more fundamental problems connected with the large-scale structure and organization of the universe must probably await the advent of the 200-inch giant of Palomar for solution. The future will always hold in store greater and more magnificent conceptions than the past; and, despite the awe-inspiring achievements already realized, it is to the future that we must turn our eyes.

INDEX OF NAMES

Printed in Great Britain by
The Camelot Press Ltd., London and Southampton