1

Introduction

Basic astronomical phenomena

A great deal of human effort has been expended over the past 4000 years or so in trying to predict and explain the motions of the Sun, Moon, planets and stars. Since Babylonian times, this quest has relied heavily on mathematics, and the developments in man's understanding of the heavens have been inextricably linked to progress in the mathematical sciences. As far as the Sun, Moon and planets are concerned, attempts at an explanation of their motion using mathematical techniques began in ancient Greece and the first mathematical model of the heavens was constructed by Eudoxus in the fourth century BC. The final piece of the celestial jigsaw was supplied by Einstein's theory of general relativity in the twentieth century and, since 1915, all major phenomena associated with planetary motion have possessed theoretical explanations. This should not be taken to imply that we know everything about the future positions of the bodies in our Solar System. Indeed, the researches of Poincaré in the late nineteenth century have led us to a much clearer understanding of the limitations of theoretical predictions. Before embarking on the fascinating story that describes the endeavours of men such as Ptolemy, Copernicus, Kepler, Newton, and Laplace, we will begin by familiarizing ourselves with the various heavenly phenomena that people have for so long sought to explain.

For a variety of reasons, early astronomers thought that the Earth was stationary and that the heavenly bodies moved around it. On the face of it this is an extremely natural assumption to make, the evidence to the contrary is far from obvious. The fact that the natural interpretation of the situation is wrong is one of the reasons why astronomy has such an absorbing history. Progress in man's understanding of the nature of the Universe has not been a gradual refinement of simple and intuitive ideas, but a struggle to replace the seemingly obvious by, what to many, was patently absurd. Nowadays, we accept that the idea of a stationary Earth at the centre of the Universe is wrong and that the Earth, Sun, planets, and stars are all in motion relative to each other, but in order to understand early approaches to astronomy it is often helpful to throw away our modern notions and to try and picture the Universe as the ancients would have done.

The most obvious of the objects visible in the sky is the Sun, and systematic observations of its motion across the sky were made by the Babylonian and Egyptian civilizations using a *gnomon*, which was simply a primitive sundial consisting of a stick placed vertically on a horizontal surface.[1] During the course of a day both the length and direction of the *gnomon*'s shadow vary; when the Sun is high in the sky the shadow is short and although the minimum length of the shadow varies from day to day, the direction of the Sun at the time when the shadow is shortest is always the same. For an observer in the northern hemisphere (and we will assume throughout that the observer is in the northern hemisphere) this direction defines due north. Each day, the Sun rises in the eastern part of the sky, travels across the sky reaching its highest point in the south in the middle of the day, and then sets in the western part of the sky. However, the amount of time for which the Sun is visible and its daily path across the sky are not constant, e.g. the points on the horizon at which the Sun rises and sets undergo variations. Observations over a sufficient period of time would show that these variations repeat, and that the period with which they do so is related to the weather (in most parts of the world); in this way one is led to the concept of a year with its seasonal changes in climate.

In Egypt, where the seasons are not particularly noticeable, this period was recognized to be the same as that associated with the flooding of the Nile and, hence, crucial to people's lives. The Egyptians noticed that for part of the year Sirius, the brightest star in the sky, was invisible as it was too close to the Sun, but that the floods began soon after the time when the star rose in the eastern sky just before dawn, and they measured the period between these so-called heliacal risings at a little over 365 days. We call this period the **tropical year**, and it forms the basis of our calendar. For calendrical purposes, the Egyptians used a year of precisely 365 days and the simplicity of such a system was beneficial in the extreme to astronomers; indeed, it was used by Copernicus as late as the sixteenth century. The division of a day into 24 h is also of ancient origin, though originally these were of unequal length and depended on one's location and the time of year in such a way that there were always 12 h of daylight and 12 h of darkness.[2]

[1] The early use of shadow sticks is described in Fermor (1997).
[2] For the purposes of everyday life, the change to 24 equal hours did not take place until the fourteenth century with the advent of the mechanical clock, but hours of equal length were used

The next most obvious heavenly body is the Moon, and this is seen to undergo not only variations in its track across the sky but also in its form, changing within a period of about $29\frac{1}{2}$ days from thin crescent to circular disc (**full moon**) back to thin crescent again and then disappearing for two or three nights (**new moon**). These changes in form are known as the 'phases of the Moon' and are associated with another characteristic length of time, known as a **lunation** or **synodic month**, that formed the bases of many ancient calendars. When the Moon is full, it is opposed diametrically to the Sun, and we say that the Sun and the Moon are in **opposition**. When the Sun and the Moon are in the same direction – which happens at the new moon – they are said to be in **conjunction**. This same terminology is used for any pair of heavenly bodies. Oppositions and conjunctions are known collectively as **syzygies**.

If observers look toward the sky on a clear night, they see a large number of stars that appear to lie on a spherical surface with themselves at the centre; this imaginary surface is called the **celestial sphere**. The stars appear to move over the surface of this sphere but the distance between them remains constant, and it was for this reason that many ancient astronomers regarded the celestial sphere as a real entity with the stars attached physically to it. This physically real celestial sphere had its centre at the centre of the Earth, but it is more convenient to take the imaginary celestial sphere (illustrated in Figure 1.1) as having its centre at the observer O. The point on the celestial sphere directly above the observer is called the **zenith**, Z.

Although we now know that the stars are not all the same distance from the Earth, it is not at all hard to imagine when looking at them that they are equidistant from us, though we have no immediately simple way of determining just how far away they are. Ignoring the question of distance, the celestial sphere gives us an easy way of describing the direction of a star: it is the same as the direction of the line that points from an observer to the point on the celestial sphere where the star appears to lie. We can then talk about the **apparent distance** between two stars as the angle between the lines pointing towards each of them.

Some careful observation will show that the stars move as if they were attached to the celestial sphere and it was rotating about an axis that intersects the sphere at a point in the northern sky. This point, which nowadays is close to the star Polaris, is called the **north pole**, P_{N}, of the celestial sphere and we can easily imagine that there is a **south pole**, P_{S}, in the part of the sky invisible to us.

from ancient times in the context of scientific discussions. A division of the day into equal temporal units was used in China from the second century BC. See Dohrn-van Rossum (1996) for a complete history of the hour.

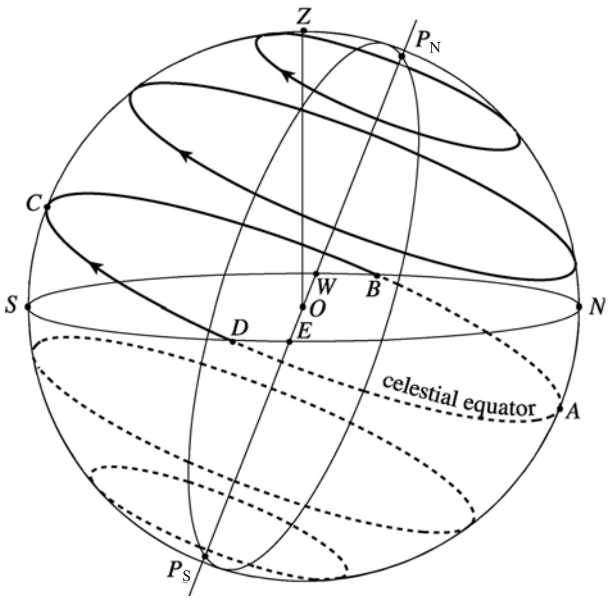

Fig. 1.1. The celestial sphere.

The stars move around the north pole in an anticlockwise direction with some stars always above the **horizon** ($NWSE$), so-called **circumpolar** stars (or, as some ancient observers termed them, ‘those that know no weariness’),[3] and some dipping below the horizon as indicated by the dashed lines in Figure 1.1. We can imagine easily that there are other stars that remain invisible below the horizon. The circle $ABCD$, which lies in a plane through O midway between the poles, is called the **celestial equator**. If we were at the north pole of the Earth, then the north pole of the celestial sphere would correspond to the zenith and all visible stars would be circumpolar, whereas if we were on the equator, the north pole of the celestial sphere would be on the horizon and there would not be any circumpolar stars. For a general point in the northern hemisphere, the angle between OZ and OP_N is 90° minus the latitude of the observer. The celestial sphere completes 1 revolution in about 1 day, and this rotation is called the ‘daily’ or **diurnal motion**. The Moon also participates in this daily rotation

[3] It is these circumpolar stars that are most suggestive of the spherical nature of the heavens. Ptolemy, in AD second century stated explicitly that these stars were instrumental in leading astronomers to the concept of a celestial sphere (PTOLEMY *Almagest*, Book I, 3 (see Toomer (1984)).

as does the Sun, though since the stars and the Sun are not seen at the same time, this is harder to appreciate.

Provided we have an accurate means of measuring time, we can observe that the stars actually complete a revolution about the pole in about 23 h 56 min, so that they return to the same place at the same time in 1 year.[4] For an ancient observer, this observation was more difficult. However, by observing stars rising above or setting below the horizon at about sunrise or sunset, one notices that the stars are moving faster than, and therefore gradually changing their position with respect to the Sun, returning to the same position after 1 year. If we regard the stars as fixed on the celestial sphere, then the Sun must move relative to the stars in the direction opposite to the diurnal motion (i.e. from west to east), completing one circuit of the celestial sphere in a year. There are other features of the motion of the Sun that need to be explained. For example, the points on the horizon at which the Sun rises and sets vary from day to day, as does the midday height of the Sun and, indeed, the length of time the Sun is above the horizon. A great deal of careful observation led ancient astronomers to the conclusion that the complex motion of the Sun was built up from two much simpler motions: the first was the daily rotation of the celestial sphere and the second was a much slower **annual motion** that took place on an oblique great circle ($AQCR$ in Figure 1.2). This circle is called the **ecliptic**, the reason being that for eclipses to occur, the Moon must be on or near it, and the angle at which it cuts the celestial equator is called the **obliquity** of the ecliptic, ε. Nowadays, the value of ε is about $23^\circ\,27'$, but the obliquity actually decreases very slowly with time. In 3000 BC, it was about $24^\circ\,2'$.[5]

In order to describe the position of a body on the celestial sphere, we need some sort of coordinates. One possibility, introduced by Babylonian astronomers, is to measure angles relative to the ecliptic. The **ecliptic longitude** measures the distance around the ecliptic, whereas the **ecliptic latitude** measures the distance north or south of this line. Hence, by definition, the ecliptic latitude of the Sun is 0°. A common alternative is to use a system based on the celestial equator, and here the distance around the equator is known as the **right ascension**, whereas the distance north or south of the equator is the **declination**.

From Figure 1.1, it can be seen that when the Sun lies on the celestial equator the lengths of the day and night are equal (since the arcs DCB and BAD are the same length) and, hence, the points A and C in Figure 1.2 are known as the **equinoctial points**, and the times when the Sun is at these points are the **equinoxes**. The time at which the Sun is at A (i.e. when it crosses the celestial

[4] Twenty-four hours divided by 365.25 (the number of days in 1 year) is just under 4 min.
[5] Thurston (1994).

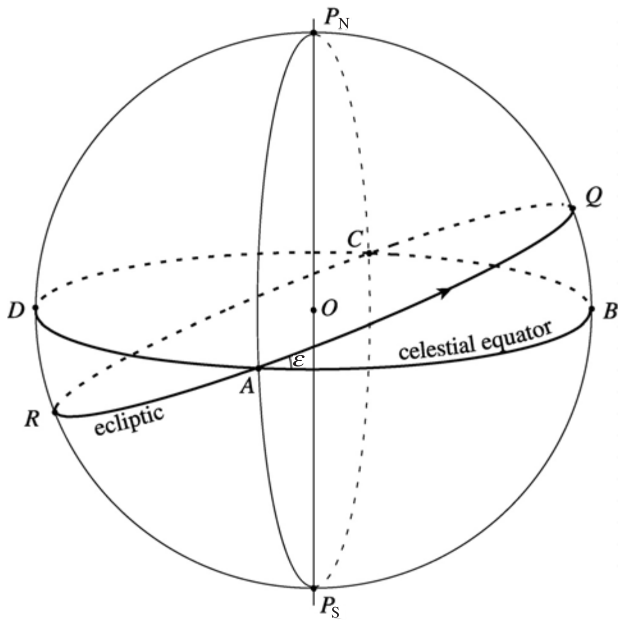

Fig. 1.2. The annual motion of the Sun.

equator from south to north) is the **spring** or **vernal equinox**, whereas the Sun is at C at the **autumnal equinox**. Ancient astronomers were aware of the fact that the Sun does not move around the ecliptic at a uniform rate – the autumnal equinox is about 186 days after the vernal equinox, but only 179 days pass before the autumnal equinox is reached again.

The Sun is at point Q – its most northerly extreme – at the **summer solstice**, and at its most southerly extreme, R, at the **winter solstice**. The points Q and R are known as the **solstitial points**. Observations over a short period of time suggest that the equinoctial and solstitial points remain fixed with respect to the fixed stars, but actually they do not; the equinoctial points rotate around the celestial equator with a period of about 26 000 years. Surprisingly, perhaps, this phenomenon, called the 'precession of the equinoxes', was recognized as early as the second century BC. The fact that the equinoxes precess means that the time taken for the Sun to return to the same position on the ecliptic (the tropical year) is different from the time taken to return to a fixed star. This latter period is known as the **sidereal year**. The modern values for these periods are approximately 365.242 days for the tropical year and 365.256 days for the sidereal year. When we use the term 'year' without qualification we mean the tropical year, though in most situations the difference is unimportant.

Table 1.1. *The periods of the moon.*

Month	Length in days
synodic	29.531
sidereal	27.322
draconitic	27.212
anomalistic	27.555

Observations of the Moon show that its monthly path round the celestial sphere is also a great circle that is very close to the ecliptic but inclined slightly to it. The orbit of the Moon crosses that of the Sun at two points called 'the **nodes** of the orbit' and the straight line joining these points is called the **nodal line**. At one of its nodes, the Moon is moving from south of the ecliptic to north of it, and we call this the **ascending node**, the other point of crossing being the **descending node**. The period of the rotation of the Moon around the celestial sphere is known as the **sidereal month**, which is, of course, different from the synodic month because that measures the motion of the Moon with respect to the Sun, which itself is moving around the ecliptic. There are two other important periods associated with the Moon which were recognized in ancient times. First, there is the **draconitic month**, which is the period between successive ascending (or descending) nodes,[6] or equivalently the time it takes for the Moon to return to the same distance from the ecliptic (i.e. to the same ecliptic latitude) while travelling in the same direction. Second, the speed with which the Moon moves across the sky relative to the stars is variable (varying between about 12 and 15° per day) and the period of time it takes for the Moon to return to the same speed is called the **anomalistic month**. The modern mean values for the various periods of the Moon are shown in Table 1.1, though the actual values for any given month may differ from these by as much as 7 h.

Perhaps the most dramatic of celestial events which are easily observable are **eclipses**, and these are of two types: lunar and solar (see Figure 1.3).[7] In a **lunar eclipse** the Earth passes between the Sun and the Moon and casts its shadow over the Moon's surface, whereas in a **solar eclipse** it is the shadow

[6] In medieval times, the part of the Moon's orbit south of the ecliptic was known as the 'dragon' (which devoured the Moon during eclipses) and from this we get the terminology 'dragon's head' for the ascending node and 'dragon's tail' for the descending node. An example of this usage can be found in Chaucer's *Treatise on the Astrolabe* (1391), one of the oldest surviving scientific works written in English. The periods between successive nodes has, over time, been termed the dracontic, draconic and draconitic month, the words deriving from the Greek for 'dragon'.

[7] Eclipses can be categorized further depending on the precise arrangement of the three bodies (see, for example, Payne-Gaposchkin (1961)).

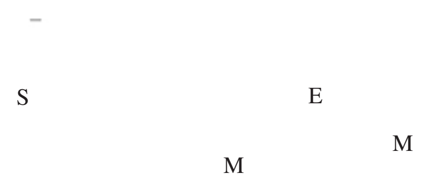

Fig. 1.3. The nature of eclipses.

of Moon that passes over the Earth. Thus, a lunar eclipse takes place when the Moon is in opposition to the Sun (i.e. when it is full) and simultaneously at (or near) one of its nodes. Similarly, a solar eclipse occurs when the passage of the Moon corresponds, through one of its nodes, to a conjunction with the Sun. Now, if the nodal line of the orbit of the Moon had a fixed orientation with respect to the stars, one of its nodes would lie directly between the Earth and the Sun every half year. Actually, the nodal line rotates (a fact known to the ancient Greeks) and as a result a node lies on the Earth–Sun line once every 173.3 days. A solar eclipse will occur if there is a new moon close enough to this time. It turns out that the shadow of the Moon will touch the Earth if the new moon appears within about $18\frac{3}{4}$ days either side of the alignment of a node and, thus, there is a $37\frac{1}{2}$-day eclipse season every 173 days during which a solar eclipse may be visible.

A similar analysis applies to lunar eclipses, but since the shadow cone of the Earth narrows as one moves further from the Sun, the full moon needs to appear nearer to the point of alignment than the $37\frac{1}{2}$-day window that exists for solar eclipses. As a result, lunar eclipses actually are less frequent than the solar variety. In any one calendar year, there are between two and five solar eclipses, but there can be no more than three lunar eclipses per year and there might not be any. The reason solar eclipses are seen so much more rarely than lunar ones is, of course, due to the fact that the area of the Earth covered by the shadow of the Moon in a solar eclipse is very small, and so a particular solar eclipse is visible from only a small part of the surface of the Earth. On the other hand, the shadow of the Earth can block out the light of the Sun from the whole of the surface of the Moon, and such a lunar eclipse will be visible from anywhere on the Earth from which the Moon normally would be visible.

Eclipses of both the Sun and Moon were observed by ancient astronomers and would no doubt have aroused great interest. Over a period of time it would have become clear that eclipses of the Moon occur only when the Moon is full, and that eclipses of the Sun occur only at new moon. The fact that solar eclipses are caused by the Moon passing in front of the Sun would not have been hard to appreciate, but the fact that eclipses of the Moon are caused by the shadow of

Table 1.2. *The zodiacal and synodic periods of the planets.*

	Zodiacal period (years)	Synodic period (days)
Mercury	1	115
Venus	1	584
Mars	1.88	780
Jupiter	11.86	399
Saturn	29.46	378

the Earth passing over it requires a rather deeper understanding of the situation and was probably not realized until much later. The reasons behind both types of eclipse were, however, fully understood by the time of the ancient Greeks.

Ancient astronomers were also aware that five of the star-like objects in the sky changed their position relative to the other stars. These five objects – now named after the Roman gods Mercury, Venus, Mars, Jupiter and Saturn – are the **planets**, from the Greek for 'wanderer'. Careful observations of these objects reveal that, like the Sun, as well as participating in the daily rotation of the heavens, they, too, move around the celestial sphere though with differing periods, and also that while they move predominantly in the same direction as the Sun – from west to east – they sometimes switch back and, for a time, move from east to west in so-called **retrograde** motions. The periods of retrograde motion are linked to the motion of the Sun – the centre of the retrograde motion for Mars, Jupiter and Saturn always occurring when the planet is in opposition to the Sun – whereas for Mercury and Venus this phenomenon occurs at conjunction. The planets also remain close to the ecliptic, the maximum deviation for any of them being 8°, and thus all the wandering heavenly bodies can be found within a strip on the celestial sphere 16° across centred on the ecliptic. This strip, therefore, is very important, and is known as the **zodiac** and was divided by the Babylonians into twelve equal parts: the signs of the zodiac. The average time it takes for a planet to complete 1 revolution around the ecliptic is its **zodiacal period**, and the average period between successive occurrences of retrograde motion is known as the **synodic period** of the planet; for the five planets visible with the naked eye these are given in Table 1.2.

The fact that the Sun, Moon and planets are nearer the Earth than the stars may have been suggested by **occultations**, the temporary disappearance of one heavenly body behind another. The most obvious example is the Moon, which sometimes eclipses the Sun and also is easily observed to pass in front of the planets and stars. Thus, all ancient astronomers agreed that the Moon was the closest of the heavenly bodies. This view was, of course, supported by the Moon's enormous size compared to the other planets and by the fact that various details could be made out on its surface. However, there is no easy way of determining how far away each celestial body is, and so ancient astronomers took the view that their distance probably was related to the speed with which they traversed the celestial sphere. This was consistent with the fact that the Moon is the swiftest of the heavenly bodies.

The Sun's journey around the ecliptic takes, by definition, 1 year, and it was found that Mars returned to the same place among the stars on average after about 2 years, Jupiter after 12 years and Saturn after $29\frac{1}{2}$ years. These planets, being slower than the Sun, were therefore thought to be further away and, hence, higher in the sky, and were termed the **superior** planets. On the other hand, Venus and Mercury both complete 1 revolution about the Earth in the same average time (1 year) and there was considerable disagreement about the correct ordering of these two planets. Eventually, a consensus was reached that they were closer to the Earth than the Sun, and they became known as the **inferior** planets. The observed behaviour of the superior and the inferior planets is different in another important way: Mercury and Venus are never seen far from the Sun, the maximum differences in longitude being about 29 and 47°, respectively, and so they are only ever visible near dawn or sunset, whereas the difference in longitude between the Sun and Mars, Jupiter or Saturn can take any value and they can be visible at any time of night.

The Sun, Moon and five planets, together with the fixed stars, were the only heavenly bodies recognized in antiquity, and this situation did not change until Galileo pointed a telescope at the night sky in 1609–10 and discovered the moons of Jupiter. No further planets were discovered until the eighteenth century, and comets were not considered as celestial bodies until the pioneering observations of Tycho Brahe in the late sixteenth century (prior to this they were thought of as atmospheric phenomena).

The problem for astronomers, then, was to explain the phenomena that have been described above. In the beginning, attempts were limited to qualitative explanation, but the models that were developed could not produce accurate quantitative predictions for astronomical events. The quantitative problem is much, much harder, and it exercised many of the greatest minds over a period of more than 2000 years, from Babylonian times to the twentieth century.

Babylonian astronomy

The heavenly phenomena were of great importance to the Babylonians, as they were perceived as omens and just about every possible astronomical event had