1

Observations of Planetary Systems

Planets can be defined informally as large bodies, in orbit around a star, that are not massive enough to have ever derived a substantial fraction of their luminosity from nuclear fusion. This definition fixes the maximum mass of a planet to be at the deuterium burning threshold, which is approximately 13 Jupiter masses for solar composition objects ($1\,M_{\rm J} = 1.899 \times 10^{30}$ g). More massive objects are called brown dwarfs. The lower mass cut-off for what we call a planet is not as easily defined. For a predominantly icy body self-gravity overwhelms material strength when the diameter exceeds a few hundred km, leading to a hydrostatic shape that is near spherical in the absence of rapid rotation (the critical diameter is larger for rocky bodies). Planets (including dwarf planets) are defined as exceeding this threshold size. As planets get larger they typically become more interesting as individual objects; larger bodies retain more internal heat to power geological processes and can hold on to more significant atmospheres. As members of a planetary system the dynamical influence of massive bodies also acts to destabilize and clear out most neighboring orbits. These physical and dynamical characteristics can be used to sub-divide the class of planets, but we will not have cause to make such distinctions in this book. It is likely that some objects of planetary mass exist that are *not* bound to a central star, having formed either in isolation or following ejection from a planetary system. Such objects are normally called "planetary-mass objects" or "free-floating planets."

Complementary constraints on theories of planet formation come from observations of the Solar System and of extrasolar planetary systems. Space missions have yielded exquisitely detailed information on the surfaces (and in some cases interior structures) of the Solar System's planets and satellites, and an increasing number of its minor bodies. Some of the most fundamental facts about the Solar System are reviewed in this chapter, while other relevant observations are discussed subsequently in connection with related theoretical topics. By comparison with the Solar System our knowledge of individual extrasolar planetary systems is meager – in many cases it can be reduced to a handful of imperfectly known numbers characterizing the orbital properties of the planets – but this is compensated

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

in part by the large and rapidly growing number of known systems. It is only by studying extrasolar planetary systems that we can make statistical studies of the range of outcomes of the planet formation process, and avoid bias introduced by the fact that the Solar System must necessarily be one of the subset of planetary systems that admit the existence of a habitable world.

1.1 Solar System Planets

The Solar System has eight planets. Jupiter and Saturn are gas giants composed primarily of hydrogen and helium, although their composition is substantially enhanced in heavier elements when compared to that of the Sun. Uranus and Neptune are ice giants, composed of water, ammonia, methane, silicates, and metals, atop which sit relatively low mass hydrogen and helium atmospheres. There are also four terrestrial planets, two of which (Earth and Venus) have quite similar masses. Mars is almost an order of magnitude less massive and Mercury is smaller still, though its density is anomalously high and similar to that of the Earth. There is more than an order of magnitude gap between the masses of the most massive terrestrial planets and the ice giants, and these two classes of planets have entirely distinct radii and structures. In addition there are a number of dwarf planets, including the trans-Neptunian objects Pluto, Eris, Haumea, and Makemake, and the asteroid Ceres. Many more dwarf planets of comparable size, and possibly even larger objects, remain to be discovered in the outer Solar System.

The orbital elements, masses and equatorial radii of the Solar System's planets are summarized in Table 1.1. With the exception of Mercury, the planets have almost circular, almost coplanar orbits. There is a small but significant misalignment of about 7° between the mean orbital plane of the planets and the solar equator. Architecturally, the most intriguing feature of the Solar System is that the giant and

Table 1.1 *The orbital elements (semi-major axis a, eccentricity e and inclination i), masses and equatorial radii of Solar System planets. The orbital elements are quoted for the J2000 epoch and are with respect to the mean ecliptic. Data from JPL.*

	a / AU	e	i / deg	M_p/g	R_p/cm
Mercury	0.3871	0.2056	7.00	3.302×10^{26}	2.440×10^{8}
Venus	0.7233	0.0068	3.39	4.869×10^{27}	6.052×10^{8}
Earth	1.000	0.0167	0.00	5.974×10^{27}	6.378×10^{8}
Mars	1.524	0.0934	1.85	6.419×10^{26}	3.396×10^{8}
Jupiter	5.203	0.0484	1.30	1.899×10^{30}	7.149×10^{9}
Saturn	9.537	0.0539	2.49	5.685×10^{29}	6.027×10^{9}
Uranus	19.19	0.0473	0.77	8.681×10^{28}	2.556×10^{9}
Neptune	30.07	0.0086	1.77	1.024×10^{29}	2.476×10^{9}

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

terrestrial planets are clearly segregated in orbital radius, with the giants only being found at large radii where the Solar Nebula (the disk of gas and dust from which the planets formed) would have been cool and icy.

The planets make a negligible contribution ($\simeq 0.13\%$) to the mass of the Solar System, which overwhelmingly resides in the Sun. The mass of the Sun, $M_\odot = 1.989 \times 10^{33}$ g, is made up of hydrogen (fraction by mass in the envelope $X = 0.73$), helium ($Y = 0.25$), and heavier elements (described in astronomical parlance as "metals," with $Z = 0.02$). One notes that even most of the condensible elements in the Solar System are in the Sun. This means that if a significant fraction of the current mass of the Sun passed through a disk during the formation epoch the process of planet formation need not be 100% efficient in converting solid material in the disk into planets. In contrast to the mass, most of the angular momentum of the Solar System is locked up in the orbital angular momentum of the planets. Assuming rigid rotation at angular velocity Ω, the solar angular momentum can be written as

$$J_\odot = k^2 M_\odot R_\odot^2 \Omega, \tag{1.1}$$

where $R_\odot = 6.96 \times 10^{10}$ cm is the solar radius. Taking $\Omega = 2.9 \times 10^{-6}$ s^{-1} (the solar rotation period is 25 dy), and adopting $k^2 \approx 0.1$ (roughly appropriate for a star with a radiative core), we obtain as an estimate for the solar angular momentum $J_\odot \sim 3 \times 10^{48}$ g cm^2 s^{-1}. For comparison, the orbital angular momentum associated with Jupiter's orbit at semi-major axis a is

$$J_\mathrm{J} = M_\mathrm{J} \sqrt{G M_\odot a} \simeq 2 \times 10^{50} \text{ g cm}^2 \text{ s}^{-1}. \tag{1.2}$$

Even this value is small compared to the typical angular momentum contained in molecular cloud cores that collapse to form low mass stars. We infer that substantial segregation of angular momentum and mass must have occurred during the star formation process.

The orbital radii of the planets do not exhibit any relationships that yield immediate clues as to their formation or early evolution. (We briefly mention the Titius–Bode law in Section 7.4.3, but this empirical relation is not thought to have any fundamental basis.) From a dynamical standpoint the most relevant fact is that although the planets orbit close enough to perturb each other's orbits, the perturbations are all nonresonant. Resonances occur when characteristic frequencies of two or more bodies display a near-exact commensurability. They adopt disproportionate importance in planetary dynamics because, in systems where the planets do not make close encounters, gravitational forces between the planets are generally much smaller (typically by a factor of 10^3 or more) than the dominant force from the star. These small perturbations are largely negligible unless special circumstances (i.e. a resonance) cause them to add up coherently over time. The simplest type of

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

resonance, known as a *mean-motion resonance* (MMR), occurs when the periods P_1 and P_2 of two planets satisfy

$$\frac{P_1}{P_2} \simeq \frac{i}{j}, \tag{1.3}$$

where i and j are integers and use of the approximate equality sign denotes the fact that such resonances have a finite width. One can, of course, always find a pair of integers such that this equation is satisfied for arbitrary P_1 and P_2, so a more precise statement is that there are no dynamically important resonances among the major planets.[1] Nearest to resonance in the Solar System are Jupiter and Saturn, whose motion is affected by their proximity to a 5:2 mean-motion resonance known as the "great inequality" (the existence of this near resonance, though not its dynamical significance, was known even to Kepler). Among lower mass objects Pluto is one of a large class of Kuiper Belt Objects (KBOs) in 3:2 resonance with Neptune, and there are many examples of important resonances among satellites and in the asteroid belt.

1.2 The Minimum Mass Solar Nebula

The mass of the disk of gas and dust that formed the Solar System is unknown. However, it is possible to use the observed masses, orbital radii and compositions of the planets to derive a *lower limit* for the amount of material that must have been present, together with a crude idea as to how that material was distributed with distance from the Sun. This is called the "minimum mass Solar Nebula" (Weidenschilling,1977a). The procedure is simple:

(1) Starting from the observed (or inferred) masses of heavy elements such as iron in the planets, augment the mass of each planet with enough hydrogen and helium to bring the augmented mixture to solar composition.
(2) Divide the Solar System up into annuli, such that each annulus is centered on the current semi-major axis of a planet and extends halfway to the orbit of the neighboring planets.
(3) Imagine spreading the augmented mass for each planet across the area of its annulus. This yields a characteristic gas surface density Σ (units g cm^{-2}) at the location of each planet.

Following this scheme, out to the orbital radius of Neptune the derived surface density scales roughly as $\Sigma(r) \propto r^{-3/2}$. Since the procedure for constructing the

[1] Roughly speaking, a resonance is typically dynamically important if the integers i and j (or their difference) are small. Care is needed, however, since although the 121:118 mean-motion resonance between Saturn's moons Prometheus and Pandora formally satisfies this condition (since the *difference* is small) one would not immediately suspect that such an obscure commensurability would be significant.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

distribution is somewhat arbitrary it is possible to obtain a number of different normalizations, but the most common value used is that quoted by Hayashi (1981):

$$\Sigma(r) = 1.7 \times 10^3 \left(\frac{r}{1\ \mathrm{AU}}\right)^{-3/2}\ \mathrm{g\ cm^{-2}}. \tag{1.4}$$

Integrating this expression out to 30 AU the enclosed mass works out to be 0.01 $M_\odot$, which is comparable to the estimated masses of protoplanetary disks around other stars (though these have a wide spread). Hayashi (1981) also provided an estimate for the surface density of solid material as a function of radius in the disk:

$$\Sigma_s(\mathrm{rock}) = 7.1 \left(\frac{r}{1\ \mathrm{AU}}\right)^{-3/2}\ \mathrm{g\ cm^{-2}}\ \text{for}\ r < 2.7\ \mathrm{AU}, \tag{1.5}$$

$$\Sigma_s(\mathrm{rock/ice}) = 30 \left(\frac{r}{1\ \mathrm{AU}}\right)^{-3/2}\ \mathrm{g\ cm^{-2}}\ \text{for}\ r > 2.7\ \mathrm{AU}. \tag{1.6}$$

These distributions are shown in Fig. 1.1. The discontinuity in the solid surface density at 2.7 AU is due to the presence of icy material in the outer disk that would be destroyed in the hotter inner regions.

Although useful as an order of magnitude guide, the minimum mass Solar Nebula (as its name suggests) provides only an approximate lower limit to the amount of mass that must have been present in the Solar Nebula. As we will discuss later, it is very likely that both the gas and solid disks evolved substantially over time. There is no reason to believe that the minimum mass Solar Nebula reflects either the initial inventory of mass in the Solar Nebula, or the steady-state profile of the protoplanetary disk around the young Sun.

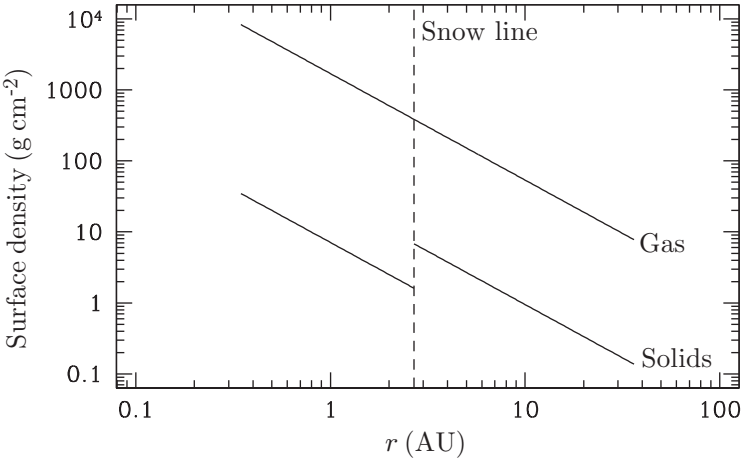

Figure 1.1 The surface density in gas (upper line) and solids (lower broken line) as a function of radius in Hayashi's minimum mass Solar Nebula. The dashed vertical line denotes the location of the snow line.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

1.3 Minor Bodies in the Solar System

In addition to the planets, the Solar System contains a wealth of minor bodies: asteroids, Trans-Neptunian Objects (TNOs, including those in the Kuiper Belt), comets, and planetary satellites. The total mass in these reservoirs is now small.[2] The main asteroid belt has a mass of about $5 \times 10^{-4}\ M_\oplus$ (Petit *et al.*, 2001), while the more uncertain estimates for the Kuiper Belt are of the order of $0.1\ M_\oplus$ (Chiang *et al.*, 2007). Although dynamically unimportant, the distribution of minor bodies is extremely important for the clues it provides to the early history of the Solar System. As a very rough generalization the Solar System is dynamically full, in the sense that most locations where small bodies could stably orbit for billions of years are, in fact, populated. In the inner Solar System, the main reservoir is the main asteroid belt between Mars and Jupiter, while in the outer Solar System the Kuiper Belt is found beyond the orbit of Neptune.

Figure 1.2 shows the distribution of a sample of numbered asteroids in the inner Solar System, taken from the *Jet Propulsion Laboratory*'s small-body database. Most of the bodies in the main asteroid belt have semi-major axes a in the range between 2.1 and 3.3 AU. The distribution of a is by no means smooth, reflecting the crucial role of resonant dynamics in shaping the asteroid belt. The prominent regions, known as the Kirkwood (1867) gaps, where relatively few asteroids are

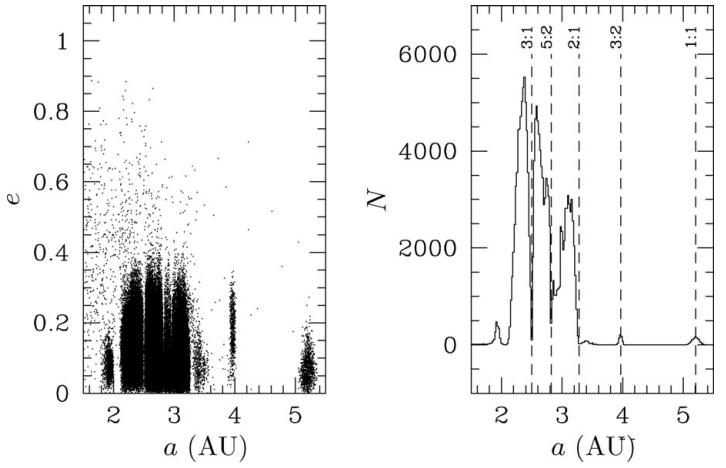

Figure 1.2 The orbital elements of a sample of numbered asteroids in the inner Solar System. The left-hand panel shows the semi-major axes a and eccentricity e of asteroids in the region between the orbits of Mars and Jupiter. The right-hand panel shows a histogram of the distribution of asteroids in semi-major axis. The locations of a handful of mean-motion resonances with Jupiter are marked by the dashed vertical lines.

[2] Indirect evidence suggests that the primordial asteroid and Kuiper belts were much more massive. A combination of dynamical ejection, and/or collisional grinding of bodies to dust that is then rapidly lost as a result of radiation pressure forces is likely to be responsible for their depletion.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

found coincide with the locations of mean-motion resonances with Jupiter, most notably the 3:1 and 5:2 resonances. In addition to these locations – at which resonances with Jupiter are evidently depleting the population of minor bodies – there are *concentrations* of asteroids at both the co-orbital 1:1 resonance (the Trojan asteroids), and at the interior 3:2 resonance (the Hilda asteroids). This is a graphic demonstration of the fact that different resonances can either destabilize or protect asteroid orbits (for a thorough analysis of the dynamics involved the reader should consult Murray & Dermott, 1999). Also notable is that the asteroids, unlike the major planets, have a distribution of eccentricity e that extends to moderately large values. Between 2.1 and 3.3 AU the mean eccentricity of the numbered asteroids is $\langle e \rangle \simeq 0.14$. As a result, collisions in the asteroid belt today typically involve relative velocities that are large enough to be disruptive. Indeed, a number of asteroid families (Hirayama, 1918) are known, whose members share similar orbital elements (a, e, i). These asteroids are interpreted as debris from disruptive collisions taking place within the asteroid belt, in some cases relatively recently (within the last few Myr, e.g. Nesvorný *et al.*, 2002).

Figure 1.3 shows the distribution of a sample of outer Solar System bodies, maintained by the IAU's *Minor Planet Center*. Among the known planets outer Solar System bodies interact most strongly with Neptune, and to leading order they are classified based upon the nature of that interaction.

- *Resonant Kuiper Belt Objects (KBOs)* currently occupy mean-motion resonances with Neptune. The most common resonance is the 3:2 that is occupied by Pluto, and such objects are also called Plutinos. The eccentricity of some Plutinos – including Pluto itself – is large enough that their perihelion lies within the orbit of Neptune, and these objects depend upon their resonant configuration to avoid close encounters. The existence of this large population of moderately eccentric resonant bodies provided the original evidence for models in which Neptune migrated outward early in Solar System history.
- *Classical KBOs* orbit in a relatively narrow belt between Neptune's 3:2 and 2:1 MMRs ($39.5\text{AU} < a < 47.8$ AU), and their number drops sharply toward the upper end of this range of semi-major axes (Trujillo & Brown, 2001). These objects are nonresonant and they have low enough eccentricity to avoid scattering encounters with Neptune. They can be divided into two sub-populations. The *cold* classical belt objects have lower inclinations $i < 2^\circ$ (and generally also lower eccentricities) than the *hot* objects, which have $i > 6^\circ$ (Dawson & Murray-Clay, 2012). (Objects with intermediate inclinations cannot be classified reliably using only orbital information.) The dynamical classification matches up to apparent physical differences that are inferred from measurements of the color and size distribution, which suggests that the cold and hot populations derive from distinct source populations.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

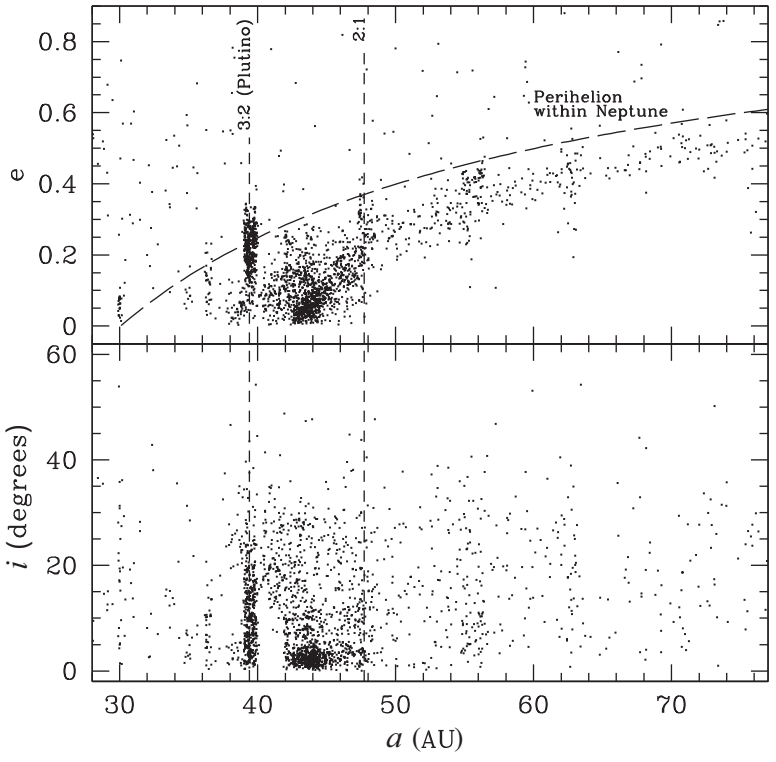

Figure 1.3 The distribution of eccentricity and inclination for a sample of minor bodies in the outer Solar System beyond the orbit of Neptune. The dashed vertical lines indicate the locations of mean-motion resonances with Neptune. Objects with eccentricity above the long-dashed line have perihelia that lie within the orbit of Neptune.

- The *scattering* population have perihelia $a(1 - e) \approx a_{\rm Nep}$, where $a_{\rm Nep}$ is the semi-major axis of Neptune. These objects are in close dynamical contact with Neptune, and their orbits evolve as a result of the planet's perturbations.
- The *detached* population makes up the rest – objects on typically quite eccentric orbits that are not currently in dynamical contact with Neptune. Some of these objects are so detached that their orbits must have been established under different dynamical conditions earlier in Solar System history. A notable example is the large object Sedna, whose perihelion distance of 76 AU lies way beyond the orbit of Neptune.

Comets that approach the Sun are more easily accessible messengers from the outer Solar System. The Jupiter Family comets have low inclinations and are thought to originate from within the TNO reservoirs discussed above. Other comets, however, have a clearly distinct origin. In particular, among comets that are identified for the first time there is a population that has a broad inclination distribution and semi-major axes that cluster at $a \sim 2 \times 10^4$ AU, far beyond the Kuiper Belt. It was this evidence that led Jan Oort to postulate that the Sun is surrounded by a

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

quasi-spherical reservoir of comets, now called the Oort cloud (Oort, 1950). The Oort cloud was established at an early epoch and delivers comets toward the inner Solar System over time as a consequence of Galactic tidal forces and perturbations from passing stars.

Planetary satellites in the Solar System also fall into several classes. The regular satellites of Jupiter, Saturn, Uranus, and Neptune have relatively tight prograde orbits that lie close to the equatorial plane of their respective planets. This suggests that these satellites formed from disks, analogous to the Solar Nebula itself, that surrounded the planets shortly after their formation. The total masses of the regular satellite systems are a relatively constant fraction (about 10^{-4}) of the mass of the host planet, with the largest satellite, Jupiter's moon Ganymede, having a mass of $0.025\ M_{\oplus}$. The presence of resonances between different satellite orbits – most notably the *Laplace resonance* that involves Io, Europa, and Ganymede (Io lies in 2:1 resonance with Europa, which in turn is in 2:1 resonance with Ganymede) – is striking. As in the case of Pluto's resonance with Neptune, the existence of these nontrivial configurations among the satellites provides evidence for past orbital evolution that was followed by resonant capture. Orbital migration within a primordial disk, or tidal interaction with the planet, are candidates for explaining these resonances.

The giant planets also possess extensive systems of irregular satellites, which are typically more distant and which do not share the common disk plane of the regular satellites. These satellites were probably captured by the giant planets from heliocentric orbits.

The sole example of a natural satellite of a terrestrial planet – the Earth's Moon – is distinctly different from any giant planet satellite. Relative to its planet it is much more massive (the Moon is more than 1% of the mass of the Earth), and its orbital angular momentum makes up most of the angular momentum of the Earth–Moon system. The Moon's composition is not the same as that of the Earth; there is less iron (resulting in a lower density than the uncompressed density of the Earth) and evidence for depletion of some volatile elements. Some aspects of the composition, in particular the ratios of stable isotopes of oxygen, are however essentially indistinguishable from those measured from terrestrial mantle samples. Qualitatively these properties are interpreted within models in which the Moon formed from the cooling of a heavy-element rich disk generated following a giant impact early in the Earth's history (Hartmann & Davis, 1975; Cameron & Ward, 1976), though some of the quantitative constraints remain challenging to reproduce. Pluto's large moon Charon may have formed in the aftermath of a similar impact.

1.4 Radioactive Dating of the Solar System

Determining the ages of individual stars from astronomical observations is a difficult exercise, and good constraints are normally only possible if the frequencies of stellar oscillations can be identified via photometric or spectroscopic data.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

Much more accurate age determinations are possible for the Solar System, via radioactive dating of apparently pristine samples from meteorites.

It is worth clarifying at the outset how radioactive dating works, because it is not as simple as one might initially think. Consider a notional radioactive decay A $\rightarrow$ B that occurs with mean lifetime τ. After time t the abundance n_A of "A" is reduced from its initial value n_{A0} according to

$$n_A = n_{A0} \mathrm{e}^{-t/\tau}, \tag{1.7}$$

while that of "B" increases,

$$n_B = n_{B0} + n_{A0}(1 - \mathrm{e}^{-t/\tau}). \tag{1.8}$$

We can assume that τ is known precisely from laboratory measurements. However, it is clear that we cannot in general determine the age because we have three unknowns (t and the initial abundances of the two species) but only two observables (the current abundances of each species). Getting around this roadblock requires considering more complex decays and imposing assumptions about how the samples under consideration formed in the first place.

For a simple example that works we can look at a rock containing radioactive potassium (^{40}K) that solidifies from the vapor or liquid phases during the epoch of planet formation. One of the decay channels of ^{40}K is

$$^{40}\mathrm{K} \rightarrow {}^{40}\mathrm{Ar}. \tag{1.9}$$

This decay has a half-life of 1.25 Gyr and a branching ratio $\xi \approx 0.1$. (The branching ratio describes the probability that the radioactive isotope decays via a specific channel. In this case ξ is small because ^{40}K decays more often into ^{40}Ca.) If we assume that the rock, once it has solidified, traps the argon and that *there was no argon in the rock to start with*, then we have eliminated one of the generally unknown quantities and measuring the relative abundance of ^{40}Ar and ^{40}K suffices to determine the age. Quantitatively, if the parent isotope ^{40}K has an initial abundance $n_\mathrm{p}(0)$ when the rock solidifies at time $t = 0$, then at later times the abundances of the parent isotope n_p and daughter isotope n_d are given by the usual exponential formulae that characterize radioactive decay:

$$\begin{aligned} n_\mathrm{p} &= n_\mathrm{p}(0)\mathrm{e}^{-t/\tau}, \\ n_\mathrm{d} &= \xi n_\mathrm{p}(0)\left[1 - \mathrm{e}^{-t/\tau}\right], \end{aligned} \tag{1.10}$$

where τ, the mean lifetime, is related to the half-life via $\tau = t_{1/2}/\ln 2$. The ratio of the daughter to parent abundance is

$$\frac{n_\mathrm{d}}{n_\mathrm{p}} = \xi\left(\mathrm{e}^{t/\tau} - 1\right). \tag{1.11}$$

A laboratory measurement of the left-hand-side then fixes the age provided that the nuclear physics of the decay (the mean lifetime and the branching ratio) is

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

accurately known. Notice that this method works to date the age of the rock (rather than the epoch when the radioactive potassium was formed) because minerals have distinct chemical compositions that differ – often dramatically so – from the average composition of the protoplanetary disk. In the example above, it is reasonable to assume that any ^{40}Ar atoms formed prior to the rock solidifying will not be incorporated into the rock, first because the argon will be diluted throughout the disk and second because it is an unreactive element that will not be part of the same minerals as potassium.

Radioactive dating is also possible in systems where we cannot safely assume that the initial abundance of the daughter isotope is negligible. The decay of rubidium 87 into strontium 87,

$$^{87}\mathrm{Rb} \rightarrow {}^{87}\mathrm{Sr}, \tag{1.12}$$

occurs with a half-life of 48.8 Gyr. Unlike argon, strontium is not a noble gas, and we cannot assume that the rock is initially devoid of strontium. If we denote the initial abundance of the daughter isotope as $n_\mathrm{d}(0)$, then measurement of the ratio $(n_\mathrm{d}/n_\mathrm{p})$ yields a single constraint on two unknowns (the initial daughter abundance and the age) and dating appears impossible. Again, the varied chemical properties of rocks allow progress. Suppose we measure samples from two different minerals within the same rock, and compare the abundances of ^{87}Rb and ^{87}Sr not to each other, but to the abundance of a separate stable isotope of strontium ^{86}Sr. Since ^{86}Sr is chemically identical to the daughter isotope ^{87}Sr that we are interested in, it is reasonable to assume that the ratio $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ was initially constant across samples. The ratio $^{87}\mathrm{Rb}/^{86}\mathrm{Sr}$, on the other hand, can differ between samples. As the rock ages, the abundance of the parent isotope drops and that of the daughter increases. Quantitatively,

$$\begin{aligned} n_\mathrm{p} &= n_\mathrm{p}(0)\mathrm{e}^{-t/\tau}, \\ n_\mathrm{d} &= n_\mathrm{d}(0) + \xi n_\mathrm{p}(0)\left[1 - \mathrm{e}^{-t/\tau}\right]. \end{aligned} \tag{1.13}$$

Eliminating $n_\mathrm{p}(0)$ between these equations and dividing by the abundance n_ds of the second stable isotope of the daughter species (^{86}Sr in our example) we obtain

$$\left(\frac{n_\mathrm{d}}{n_\mathrm{ds}}\right) = \left(\frac{n_\mathrm{d}(0)}{n_\mathrm{ds}}\right) + \xi\left(\frac{n_\mathrm{p}}{n_\mathrm{ds}}\right)\left[\mathrm{e}^{t/\tau} - 1\right]. \tag{1.14}$$

The first term on the right-hand-side is a constant. We can then plot the relative abundances of the parent isotope $(n_\mathrm{p}/n_\mathrm{ds})$ and the daughter isotope $(n_\mathrm{d}/n_\mathrm{ds})$ from different samples on a ratio–ratio plot called an isochron diagram, such as the one shown schematically in Fig. 1.4. Inspection of Eq. (1.14) shows that we should expect the points from different samples to lie on a straight line whose slope (together with independent knowledge of the mean lifetime) fixes the age. Two samples are in principle sufficient to yield an age determination, but additional data

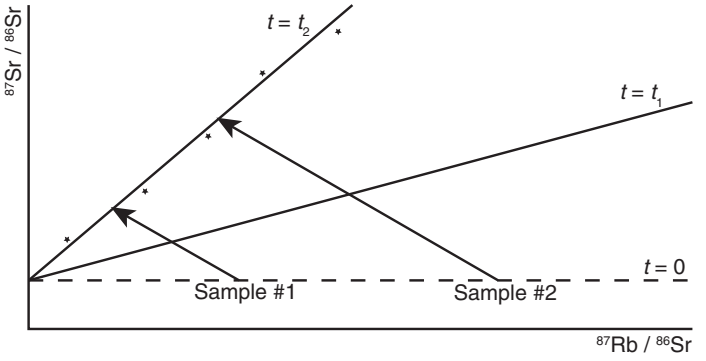

Figure 1.4 Ratio–ratio plot for dating rocks using the radioactive decay $^{87}\text{Rb} \rightarrow ^{87}\text{Sr}$. The abundance of these isotopes is plotted relative to the abundance of a separate stable isotope of strontium ^{86}Sr. When the rock solidifies, different samples contain identical ratios of $^{87}\text{Sr}/^{86}\text{Sr}$, but different ratios of $^{87}\text{Rb}/^{86}\text{Sr}$. The ratios of different samples track a steepening straight line as the rock ages.

provide a check against possible systematic errors. If the points fail to lie on a straight line something is wrong.

1.4.1 Lead–Lead Dating

The most important system for absolute determination of the age of Solar System samples, lead–lead dating, is based upon the measurement of lead and uranium isotopes. Lead has four stable and naturally occurring isotopes with mass numbers 204, 206, 207, and 208. Of these, ^{204}Pb preserves its primordial abundance, while the abundance of the heavier isotopes increases over time because these are daughter isotopes from the decay of uranium and thorium.The parent/daughter relationships are

$$\begin{aligned} ^{238}\text{U} &\rightarrow ^{206}\text{Pb} \quad (t_{1/2} = 4.47 \times 10^9 \text{ yr}), \\ ^{235}\text{U} &\rightarrow ^{207}\text{Pb} \quad (t_{1/2} = 7.04 \times 10^8 \text{ yr}), \\ ^{232}\text{Th} &\rightarrow ^{208}\text{Pb} \quad (t_{1/2} = 1.40 \times 10^{10} \text{ yr}). \end{aligned} \tag{1.15}$$

The relatively high abundance of the parent isotopes, coupled with the favorable half-lives, mean that this system is well suited to deliver ages of high accuracy and precision.

There are several ways to derive dates using the above system. Limiting attention to the two uranium decays gives a pair of equations analogous to Eq. (1.14):

$$\left(\frac{^{206}\text{Pb}}{^{204}\text{Pb}}\right) = \left(\frac{^{206}\text{Pb}}{^{204}\text{Pb}}\right)_0 + \left(\frac{^{238}\text{U}}{^{204}\text{Pb}}\right)\left[e^{t/\tau_1} - 1\right], \tag{1.16}$$

$$\left(\frac{^{207}\text{Pb}}{^{204}\text{Pb}}\right) = \left(\frac{^{207}\text{Pb}}{^{204}\text{Pb}}\right)_0 + \left(\frac{^{235}\text{U}}{^{204}\text{Pb}}\right)\left[e^{t/\tau_2} - 1\right]. \tag{1.17}$$

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

Here the ratios presented without subscripts are the present-day (measurable) values, while those with the subscript "0" refer to the initial values at the time when the system became closed. The mean lifetimes τ_1 and τ_2 refer to the decay of ^{238}U and ^{235}U. At this point we could, following the logic above, make two ratio–ratio plots (e.g. from the first equation ^{206}Pb/^{204}Pb against ^{238}U/^{204}Pb) and derive two independent ages from a single sample of primitive material. It turns out, however, that we can do a better job at reducing practical sources of error (associated, for example, with terrestrial contamination) by considering the two uranium/lead systems jointly. By taking the ratio of Eq. (1.17) to Eq. (1.16) we obtain a form in which the left-hand-side depends only on initial or current lead isotope ratios, while the right-hand-side depends only on the current uranium isotope ratio:

$$\frac{(^{207}\mathrm{Pb}/^{204}\mathrm{Pb}) - (^{207}\mathrm{Pb}/^{204}\mathrm{Pb})_0}{(^{206}\mathrm{Pb}/^{204}\mathrm{Pb}) - (^{206}\mathrm{Pb}/^{204}\mathrm{Pb})_0} = \left(\frac{^{235}\mathrm{U}}{^{238}\mathrm{U}}\right)\left(\frac{\mathrm{e}^{t/\tau_1} - 1}{\mathrm{e}^{t/\tau_2} - 1}\right). \tag{1.18}$$

Multiplying first by the denominator and then by (^{204}Pb/^{206}Pb) shows that a ratio–ratio plot of (^{204}Pb/^{206}Pb) versus (^{207}Pb/^{206}Pb) ought to give a straight line:

$$\left(\frac{^{207}\mathrm{Pb}}{^{206}\mathrm{Pb}}\right) = \underbrace{\left(\frac{^{235}\mathrm{U}}{^{238}\mathrm{U}}\right)\left(\frac{\mathrm{e}^{t/\tau_1} - 1}{\mathrm{e}^{t/\tau_2} - 1}\right)}_{\text{intercept}} + b(t)\left(\frac{^{204}\mathrm{Pb}}{^{206}\mathrm{Pb}}\right). \tag{1.19}$$

(Here $b(t)$ is an age-dependent constant whose exact form is not important for our purposes.) Measuring the intercept from a ratio–ratio plot involving the three lead isotopes, along with a current determination of the uranium isotope ratio, yields a measure of the sample age.

Connelly *et al.* (2017) review both the theory and the practical difficulties encountered in lead–lead dating of Solar System samples. Astonishingly, age determinations using this method are able to resolve events at the dawn of the Solar System, 4.57 Gyr ago, with a precision that is of the order of 0.2 Myr.

1.4.2 Dating with Short-Lived Radionuclides

Absolute chronometry provides an estimated age for the earliest Solar System solids of 4.5673 ± 0.0002 Gyr (Connelly *et al.*, 2017). This is the measurable quantity that, by convention, is described as the "age of the Solar System." Knowing the age of the Sun is useful for calibrating solar evolution models, but the exact absolute age is otherwise superfluous for studies of planet formation. More valuable are constraints on the time scale of critical phases of the planet formation process. For example, it is of interest to know whether the formation of the km-sized bodies called planetesimals was sudden or spread out over many Myr. Questions of this kind, which involve only *relative* ages, can of course be addressed given sufficiently accurate absolute chronometry. A complementary approach determines only relative ages from the decay products of short-lived isotopes. It is well established that the Solar Nebula initially contained radioactive

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

species with very short half-lives, including ^{26}Al which has a half-life of only 0.72 Myr. The likely origin of these isotopes is nucleosynthesis within the cores of massive stars followed by ejection into the surrounding medium via either a supernova explosion or Wolf–Rayet stellar winds. In either case the implication is that the Sun formed in proximity to a rich star forming region (or within a cluster, now dissolved) that also contained massive stars.

Heating due to the radioactive decay of ^{26}Al is important for the thermal evolution of small bodies forming within protoplanetary disks, and the ionization produced from these decays may also be physically significant. For dating, the key point is that we can learn something about the relative ages of samples by measuring the abundance of daughter isotopes even when all of the parent species has long since decayed. For example, we could imagine measuring the abundance of ^{129}Xe, formed via the decay,

$$^{129}\mathrm{I} \rightarrow {}^{129}\mathrm{Xe}, \tag{1.20}$$

and comparing it to the abundance of a stable isotope of iodine ^{127}I. The ^{129}I decay has a a half-life of only 17 Myr. Accordingly, we would expect that samples that form early would incorporate a higher (^{129}I/^{127}I) ratio, and this would lead to a higher (^{129}Xe/^{127}I) ratio that is preserved at late times after all of the radioactive iodine is gone.

The main caveat to be borne in mind with radioactive dating (both absolute and relative) is that it depends upon there being no other processes besides radioactive decay that alter the abundance of either the parent or daughter isotopes. The radioactive age of a rock measures the moment it solidified only if there has been no diffusion, or other alteration of the rock, during the intervening period. Even if the rock itself is pristine, high energy particles (cosmic rays) can induce nuclear reactions that may change the abundances of some critical isotopes. Relative chronometry additionally relies on the assumed spatial homogeneity of the parent species within the disk. Because the half-lives of the species used for relative dating are short, there may not have been enough time for the short-lived parents to have been mixed throughout the gas that forms the disk, and the assumption of homogeneity can reasonably be questioned.

1.5 Ice Lines

Liquid water is not stable under the low pressure conditions of the gaseous protoplanetary disk. Water ice, which is abundant at low temperature, sublimates directly to vapor in regions of the disk where the temperature exceeds about 150–170 K. The radial location in the disk beyond which the temperature falls below this value is called the snow line. Analogous "ice lines" exist for other common chemical species. The carbon monoxide (CO) ice line corresponds to a temperature of 23–28 K, while the carbon dioxide (CO_2) ice line is predicted to be at 60–72 K.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

Theoretically, the location of the snow line is expected to change over time, due to variations in the stellar luminosity and in the rate at which gas is being accreted through the disk (as we will discuss in Chapter 3, accretion can be an important source of heat). For a solar mass, solar luminosity star, accreting at an observationally typical rate, the snow line is predicted to lie between 1 and 2 AU. Reasonable variations of these parameters lead to snow lines between about 0.5 and 5 AU (Garaud & Lin, 2007). Observationally, the location of the snow line is difficult to pin down from astronomical data on protoplanetary disks, and hence the main constraints come from laboratory measurement of meteoritic samples. Meteorites of the class known as carbonaceous chondrites, which are water-rich, have properties (such as their reflectance spectra) that match those of asteroids found only in the outer asteroid belt beyond about 2.5 AU. Conversely, the ordinary and enstatite chondrites, which contain negligible amounts of water, appear to originate from the inner asteroid belt at a distance from the Sun of around 2 AU. Consistent with this, the dwarf planet Ceres in the outer asteroid belt is known to have both surface and sub-surface water, and may contain a substantial interior water reservoir. The empirical conclusion is that compositional variations in the asteroid belt preserve evidence for a snow line in the early Solar Nebula at about $a \simeq 2.7$ AU.

The location of the snow line has important consequences for the composition and habitability of terrestrial planets. The standard definition of the "habitable zone" for terrestrial planets is based on requiring that liquid water is stable on the surface. Under Earth's atmospheric pressure, that translates into a surface temperature between 273 K $< T <$ 373 K. As noted above, however, the lower pressure in the disk gas means that prior to planet formation water would have existed only in the vapor phase at temperatures above $T \simeq$ 150–170 K. This leads to a somewhat surprising prediction; planets that formed in situ within the habitable zone are typically assembled from planetesimals that formed interior to the snow line in a region of the disk that would have been too hot for water-rich minerals to condense. Solar System evidence for a snow line at 2.7 AU supports this scenario for the Earth.

The idea that habitable planets form inside the snow line is consistent not just with Solar System meteoritic evidence, but with the first-order composition of the Earth. Surface appearances to the contrary, the Earth is not a water world. The mass of water in the ocean, atmosphere, and crust of the Earth is just $2.8 \times 10^{-4}\, M_\oplus$ (where the Earth mass, $M_\oplus = 5.974 \times 10^{27}$ g). The amount of extra water locked in the mantle is uncertain (especially early in the Earth's history) and could be substantial, but it is clear that the total water fraction is orders of magnitude below that found in ice-rich bodies beyond the snow line.

Several models have been proposed to explain where the Earth's water came from. The leading hypothesis is that water was delivered to the Earth from a reservoir of small bodies (asteroids, comets, or a now vanished class of objects) beyond the snow line. Alternatively, Solar System material at 1 AU may have

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

contained a small admixture of water (in the form of water molecules that can bind to some grain surfaces at relatively high temperature), and not have been as dry as is assumed in the standard model. The isotopic makeup of the Earth's water provides a constraint. The reference value for the deuterium to hydrogen (D/H) ratio of water in the Earth's ocean is 156 parts per million (ppm), in broad agreement with the mean 159 ± 10 ppm measured in carbonaceous chondrites (Morbidelli *et al.*, 2000). Data on comets are more limited, but values measured for Oort cloud comets are substantially higher (around 300 ppm). Hartogh *et al.* (2011) reported an Earth-like value of $\approx$160 ppm for the Jupiter family comet 103P/Hartley 2, but a much higher ratio of $\approx$530 pm was measured for 67P/Churyumov–Gerasimenko (the comet, also from the Jupiter family, which was the target of the *Rosetta* mission; Altwegg *et al.*, 2015).

Given the admittedly inconclusive evidence, the most likely source for the bulk of the Earth's water is asteroids from the outer part of the main belt. The mass of asteroids required is substantial. If our water arrived via asteroids with compositions similar to the carbonaceous chondrites, which have mass fractions of water of $\approx$10%, the total mass required would have been a few $10^{-3}\, M_{\oplus}$. Although this mass is negligible on the scale of the impacts that assembled the Earth and formed the Moon, it still exceeds the total mass in the present-day asteroid belt by an order of magnitude.

1.6 Meteoritic and Solar System Samples

Unique but sometimes hard to interpret information about conditions in the Solar Nebula comes from the study of Solar System samples, recovered from meteorites and from a handful of sample return missions. Meteorites originate from the asteroid belt, the Moon, and Mars, and have a composition that reflects the structure of their parent bodies. Undifferentiated bodies, which are generally smaller, never became hot enough to form a distinct core/mantle/crust structure. They give rise to chondritic meteorites, which once the most volatile elements are excluded have a bulk chemical composition that is similar to that of the Sun. Differentiated bodies, by contrast, are the progenitors of iron meteorites (made up of iron–nickel alloys) and achondrites, stony meteorites that are mostly made up of igneous rocks.

The chondritic meteorites are divided into numerous classes and are not, for the most part, entirely pristine. Their minerals often show evidence for changes that occurred on the parent body due to the action of heat and/or water. The carbonaceous chondrites are among the least altered (or most "primitive"), and together with samples returned directly from comets and asteroids represent the closest approximation to early Solar System material that can be studied in the laboratory.

The rock of chondritic meteorites is a mixture of three quite distinct components: refractory inclusions, chondrules, and matrix. The refractory inclusions – calcium, aluminium-rich inclusions (CAIs), and amoeboid olivine aggregates – have a

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

chemical composition and structure that suggest that they formed at temperatures in excess of 1350 K. Both absolute and relative dating support the idea that CAIs are the oldest known Solar System materials, with an age estimated at 4.5673 ± 0.0002 Gyr (Connelly *et al.*, 2017). The dispersion in the inferred ages of CAIs is small, suggesting that they may have formed under high temperature disk conditions, presumably close to the Sun, during the earliest stage of disk evolution.

Chondrules are a second high temperature component of chondrites. Chondrules are typically 0.1–1 mm spheres of igneous rock that make up a variable but often large fraction of chondritic meteorites (as much as 80% of the volume of ordinary and enstatite chondrites, and 0 to 70% of carbonaceous chondrites; Nittler & Ciesla, 2016). The most basic fact about chondrules, indicated by their spherical shape and mineral texture, is that they formed due to cooling of initially molten precursors. A great deal is known experimentally about what happens as igneous rocks cool and crystallize, and as a consequence there are surprisingly informative constraints on the conditions under which chondrules formed (Desch *et al.*, 2012; Nittler & Ciesla, 2016). It is generally accepted that chondrule precursors – the solid particles or material that were melted to become chondrules – formed under relatively low temperature conditions, because chondrules contain FeS, which would not be present at $T \gtrsim 650$ K. The precursor material was then heated rapidly, on a time scale of at most minutes, to a temperature of about 2000 K. Volatile elements, such as Na and K, may have been lost while the chondrule was molten, but did not experience the mass-dependent isotope fractionation that would be expected if they were surrounded by a low–pressure gas. This implies formation in a relatively large high density environment where evaporation and fractionation are suppressed by the development of a volatile-rich vapor. After melting, the various mineral textures observed in chondrules suggest cooling at rates between about 10 K per hour and 10^4 K per hour. Even the top end of this range is orders of magnitude below the cooling rate expected for a single small sphere radiating freely to space, implying again that the formation regions were large enough so that radiative transfer effects slowed the cooling rate. Both absolute and relative chronometry show that chondrules did not form before CAIs. Some chondrules may have formed at roughly the same time as refractory inclusions, but others formed several Myr later.

The third component of chondritic meteorites is matrix, small micrometer-sized particles of minerals such as crystalline magnesian olivine, pyroxene, and amorphous ferromagnesian silicates. Matrix materials contain a larger fraction of volatile elements, consistent with formation under lower temperature conditions than those needed for either CAI or chondrule formation. The average composition of chondrules and matrix is, in a colloquial sense, "complementary," in that the former is depleted and the latter enriched in volatiles. The hypothesis of chemical *complementarity* holds that this relationship also holds for individual chondrites.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

Complementarity in this stricter sense would provide evidence that chondrules and matrix formed from a common reservoir of disk gas.

Both CAIs and chondrules are found within meteorites whose parent bodies orbited within the asteroid belt at orbital radii not far from the inferred location of the snow line. At this radii there is almost an order of magnitude gulf between the expected disk temperatures and those implicated in the formation of these high temperature materials. For chondrules two main classes of models have been proposed to explain how the required high temperature can be realized at 3 AU from the Sun. One holds that chondrules formed as pre-existing solid particles passed through shock waves in the gas disk. The shock waves could have been either large-scale spiral shocks in the disk (although it is hard to see how these would form), or more localized bow shocks around orbiting planetary embryos. Alternatively, chondrules may have formed out of the cooling debris of planetesimal impacts. For CAIs the standard scenario is that these materials formed within the disk, probably closer to the Sun, and were transported radially before being incorporated into larger solid bodies.

The mystery of how CAIs and (especially) chondrules formed is one of the most important open problems in meteoritics and cosmochemistry. The wider relevance for planet formation cannot be determined without first knowing the answer. All that we know for sure is that the small mass currently in undifferentiated asteroids is made up of a high fraction by mass of chondrules. If the formation mechanism for chondrules involved long-lived disk processes then it is reasonable to assume that a large fraction of the solid material that forms terrestrial planets was once processed through a chondrule-like phase. Chondrules would then be generic precursors of planets. If, on the other hand, chondrules form from a process such as late planetesimal collisions, they are bystanders of only peripheral interest for understanding the main phases of planet formation.

Results from the *Stardust* sample return mission from the Jupiter family comet Wild 2 raise district but thematically related questions. Analysis of 1–10 μm particles collected from the comet showed that many of them are made up of olivine or pyroxene. These are crystalline silicates that can be produced from amorphous precursors via annealing at temperatures of 800 K and above (Brownlee *et al.*, 2006). This result implies that even the very cold regions of the disk where comets formed were somehow polluted with a fraction of material processed through high temperatures.

1.7 Exoplanet Detection Methods

The first extrasolar planetary system to be discovered was identified by Alex Wolszczan and Dale Frail via precision timing of pulses from the millisecond pulsar PSR1257+12 (Wolszczan & Frail, 1992). The system contains at least three planets

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

on nearly circular orbits within 0.5 AU of the neutron star, with the outer two planets having masses close to 4 $M_\oplus$ and the inner planet having a mass 0.02 $M_\oplus$ which is comparable to the mass of the Moon. Millisecond pulsars are a type of rapidly rotating neutron star that form in supernova explosions, and the mass loss during the explosion would unbind any pre-existing planets. The planets in the PSR1257+12 system must therefore have originated within a disk formed subsequent to the explosion. The observation that pulsar planets are not common (Kerr *et al.*, 2015) implies that the conditions leading to their formation could involve moderately rare events.

The precision required to find Jupiter mass planets in relatively short-period orbits via monitoring of the radial velocity of their host stars first became available in the 1980s (Campbell *et al.*, 1988). The first accepted detection, of 51 Peg b, a Jupiter mass body orbiting a solar-type star, was made using this technique by Michel Mayor and Didier Queloz in 1995 (Mayor & Queloz, 1995). The first two decades of exoplanet discovery were dominated by detections made via the radial velocity and transit methods, with smaller numbers of planets being found using microlensing and direct imaging techniques. An additional technique, astrometry, is expected to play a growing role in future planet discovery. Based on existing data it is known that planets have a broad range of masses, from sub-Earth mass up to the deuterium burning limit, and orbit all the way from 0.01 AU to 100 AU from their host star. No single planet discovery method is sensitive across this entire parameter space, large fractions of which remain unexplored by any method. A synthesis of multiple surveys, each with its own selection function in planetary mass, radius, and orbital period, is needed to build up an understanding of the population of extrasolar planetary systems.

1.7.1 Direct Imaging

The most straightforward way to detect extrasolar planets is to image the planet as a source of light that is spatially separated from the stellar emission. The main difficulty is the extreme contrast between a planet and its host star. A planet of radius R_p, orbital radius a, and albedo A intercepts and reflects a fraction f of the incident starlight given by

$$f = \left(\frac{\pi R_\mathrm{p}^2}{4\pi a^2}\right) A = 1.4 \times 10^{-10} \left(\frac{A}{0.3}\right) \left(\frac{R_\mathrm{p}}{R_\oplus}\right)^2 \left(\frac{a}{1\ \mathrm{AU}}\right)^{-2}. \tag{1.21}$$

Recalling that magnitudes are defined in terms of the flux F via $m = -2.5 \log_{10} F +$ const, one finds that Earth-like planets are expected to be 24–25 magnitudes fainter than their host stars. This faintness means that, quite apart from the very unfavorable contrast ratio between planet and star, moderately deep exposures are needed to have even a chance of directly imaging planets in reflected light. Alternatively, one may contemplate imaging extrasolar planets in their thermal emission. If we crudely

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

approximate the emission at frequency ν from a planet with surface temperature T as a blackbody, then the spectrum is described by the Planck function:

$$B_\nu(T) = \frac{2h\nu^3}{c^2}\frac{1}{\exp(h\nu/k_\mathrm{B}T) - 1}, \tag{1.22}$$

where h is Planck's constant, c the speed of light, and k_B the Boltzmann constant. The peak of the spectrum falls at $h\nu_\mathrm{max} = 2.8k_\mathrm{B}T$, which lies in the mid-IR for typical terrestrial planets (the wavelength corresponding to ν_max is $\lambda \approx 20$ μm for the Earth with $T = 290$ K). If the star, with radius R_*, also radiates as a blackbody at temperature T_*, the flux ratio at frequency ν is

$$f = \left(\frac{R_\mathrm{p}}{R_*}\right)^2 \frac{\exp(h\nu/k_\mathrm{B}T_*) - 1}{\exp(h\nu/k_\mathrm{B}T) - 1}. \tag{1.23}$$

For an Earth-analog around a solar-type star the contrast ratio for observations at a wavelength of 20 μm is $f \sim 10^{-6}$, which is some four orders of magnitude more favorable than the corresponding ratio in reflected light. This advantage of working in the infrared is, however, offset by the need for a larger telescope in order to spatially resolve the planet. The spatial resolution of a telescope of diameter D working at a wavelength λ is

$$\theta \sim 1.22\frac{\lambda}{D}. \tag{1.24}$$

A spatial resolution of 0.5 AU at a distance of 5 pc corresponds to an angular resolution of 0.1 arcsec, which is theoretically achievable in the visible ($\lambda = 550$ nm) with a telescope of diameter $D \approx 1.5$ m. At 20 μm, on the other hand, the required diameter balloons to $D \approx 50$ m, which is unfeasibly large to contemplate constructing as a monolithic structure in space.

These elementary considerations show that the difficulty of directly imaging planets depends strongly on their orbital radii. Giant planets orbiting at large or very large radii (tens of AU) are relatively easy to image, with known examples including the multiple planets of the HR 8799 system (Marois *et al.*, 2008). State-of-the-art direct imaging surveys employ a combination of telescope hardware and novel modes of observation to overcome the unfavorable contrast ratio between star and planet. Adaptive optics can recover close to diffraction-limited performance of large ground-based telescopes in the near-IR, and a coronograph can be used to suppress the dominant stellar contribution. Techniques such as angular differential imaging (Marois *et al.*, 2005), which combines observations that are rotated in the sky plane to reduce imaging artifacts caused by the point-spread-function of the optics, further improve the available dynamic range for planet discovery.

The challenge of imaging terrestrial planets orbiting within the habitable zone is altogether more formidable that that of detecting giant planets, but carries the payoff of being able to characterize their atmospheres through spectroscopy. The composition of the Earth's atmosphere is strongly modified by the presence of life,

and the overarching goal of proposals to image terrestrial planets is to search for biomarkers such as oxygen, ozone, or methane. In principle, the depth, resolution, and contrast required to first image and then obtain low resolution ($R \sim 10^2$) spectra of potentially habitable planets could be realized in three different ways:

- A highly optimized coronagraph on a space telescope of at least 4 m class, operating with near diffraction-limited performance in the ultraviolet (UV) to near-IR spectral range (0.1 μm $\lesssim \lambda \lesssim$ 2 μm). Larger apertures offer greater throughput and higher resolution, but become more challenging due to the high degree of wavefront stability that is required for planetary detection.
- A similar sized telescope that operates in tandem with a starshade. An optimally shaped starshade of 50–100 m diameter, flying $\sim 10^5$ km away from a space telescope, creates a narrow patch of extremely deep shadow that is highly effective at suppressing starlight and allowing planet detection and characterization.
- An interferometer, with optical elements on multiple free-flying spacecraft, operating in the mid-IR.

These architectures are all optically possible but technically very challenging, and which approach is selected for implementation first will depend more on engineering considerations of cost and technical risk than on any simple physical principle. Over the longer term it is likely that observations in both the visible and mid-IR wavebands will be needed, since detailed characterization of the possible existence of life on any nearby habitable planets will require spectroscopy over as wide a band as possible.

1.7.2 Radial Velocity Searches

A star hosting a planet describes a small orbit about the center of mass of the star–planet system. The radial velocity method for finding extrasolar planets works by measuring this stellar orbit via detection of periodic variations in the radial velocity of the star. The radial velocity is determined from the Doppler shift of the stellar spectrum, which in favorable cases can be measured at better than meter per second precision.

Circular Orbits

The principles of the radial velocity technique can be illustrated with the simple case of circular orbits. Consider a planet of mass M_p orbiting a star of mass M_* in a circular orbit of semi-major axis a. For $M_\mathrm{p} \ll M_*$ the Keplerian orbital velocity of the planet is

$$v_\mathrm{K} = \sqrt{\frac{GM_*}{a}}. \tag{1.25}$$

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

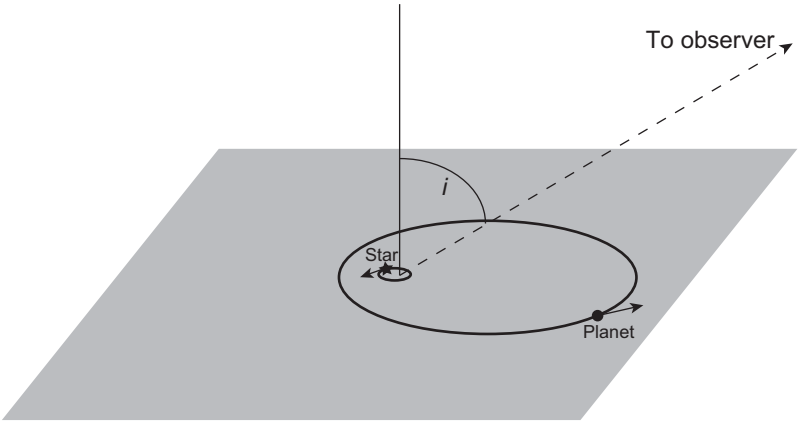

Figure 1.5 The reflex motion (greatly exaggerated) of a star orbited by a planet on an eccentric orbit. The inclination of the system is the angle i between the normal to the orbital plane and the observer's line of sight.

Conservation of linear momentum implies that the orbital velocity v_* of the star around the center of mass is determined by $M_* v_* = M_\mathrm{p} v_\mathrm{K}$. If the planetary system is observed at an inclination angle i, as shown in Fig. 1.5, the radial velocity varies sinusoidally with a semi-amplitude

$$K = v_* \sin i = \left(\frac{M_\mathrm{p}}{M_*}\right)\sqrt{\frac{GM_*}{a}}\sin i. \tag{1.26}$$

K is a directly observable quantity, as is the period of the orbit

$$P = 2\pi\sqrt{\frac{a^3}{GM_*}}. \tag{1.27}$$

If the stellar mass can be estimated independently but the inclination is unknown, as is typically the case, then we have two equations in three unknowns and the best that we can do is to determine a lower limit to the planet mass via the product $M_\mathrm{p} \sin i$. Since the average value of $\sin i$ for randomly inclined orbits is $\pi/4$ the statistical correction between the minimum and true masses is not large. The correction for individual systems is not normally known, however, and can be an important uncertainty, for example when analyzing the dynamics and stability of multiple planet systems.

Consideration of the solar radial velocity that is induced by the planets in the Solar System gives an idea of the typical magnitude of the signal. For Jupiter $v_* = 12.5\ \mathrm{m\,s^{-1}}$ while for the Earth $v_* = 0.09\ \mathrm{m\,s^{-1}}$. For planets of a given mass there is a selection bias in favor of finding planets with small a. In an idealized survey in which the noise per observation is constant from star to star, Eq. (1.26) implies that the selection limit scales as

$$M_\mathrm{p} \sin i|_\mathrm{minimum} = Ca^{1/2}, \tag{1.28}$$

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

with C a constant. Planets with masses below this threshold are undetectable, as are planets with orbital periods that exceed the duration of the survey (since orbital solutions are generally poorly constrained when only a fraction of an orbit has been observed).

Eccentric Orbits

For real applications it is necessary to consider the radial velocity signature produced by planets on eccentric orbits. The derivation of most of the required results is worked through in Appendix B, so here we just state the main results that are required.

Consider a planet on an orbit of eccentricity e, semi-major axis a, and period P. The orbital radius varies between $a(1+e)$ (apocenter) and $a(1-e)$ (pericenter). Suppose that passage of the planet through pericenter occurs at time $t_{\rm peri}$. In terms of these quantities, the *eccentric anomaly*[3] E is defined implicitly via Kepler's equation,

$$\frac{2\pi}{P}\left(t - t_{\rm peri}\right) = E - e\sin E. \tag{1.29}$$

Kepler's equation is transcendental and cannot be solved for E in terms of simple functions. However, it can readily be solved numerically. Once E is known, the *true anomaly* f is given by Eq. (B.16),

$$\tan\frac{f}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E}{2}. \tag{1.30}$$

The true anomaly is the angle between the vector joining the bodies and the pericenter direction. Finally, in terms of these quantities, the radial velocity of the star is

$$v_*(t) = K\left[\cos(f+\varpi) + e\cos\varpi\right], \tag{1.31}$$

where the longitude of pericenter ϖ is the angle in the orbital plane between pericenter and the line of sight to the system. The eccentric generalization of Eq. (1.26) in the same limit in which $M_{\rm p} \ll M_*$ is

$$K = \frac{1}{\sqrt{1-e^2}}\left(\frac{M_{\rm p}}{M_*}\right)\sqrt{\frac{GM_*}{a}}\sin i. \tag{1.32}$$

For a planet of given mass and semi-major axis, the amplitude of the radial velocity signature therefore increases with increasing e, due to the rapid motion of the planet (and star) close to pericenter passages.

Figure 1.6 illustrates the form of the radial velocity curves as a function of the eccentricity of the orbit and longitude of pericenter. Both e and ϖ can be measured given measurements of v_* as a function of time. Compared to the circular orbit

[3] The eccentric anomaly has a rather complex geometrical interpretation, but for our purposes all that matters is that it is a monotonically increasing function of t that specifies the location of the body around the orbit.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

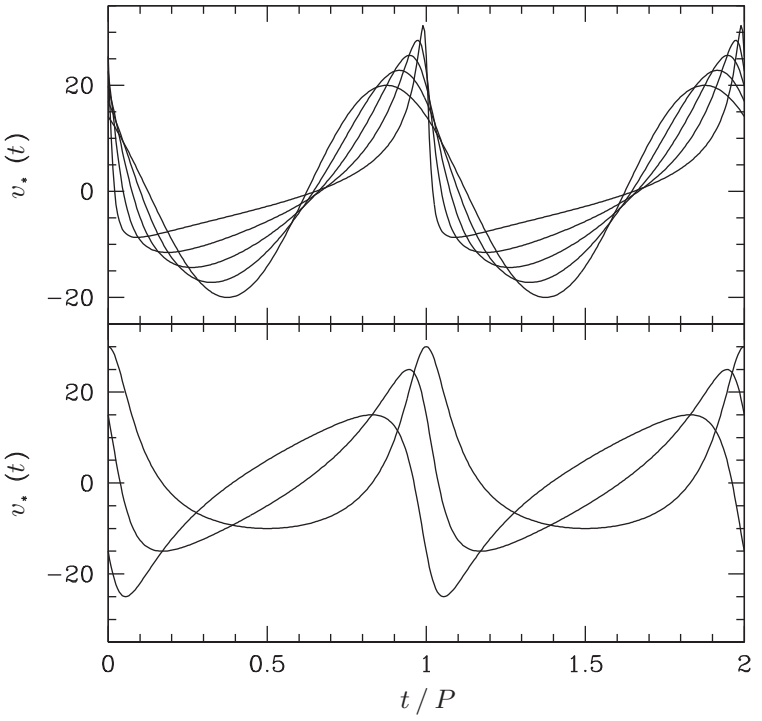

Figure 1.6 Time dependence of the radial velocity of a star hosting a single planet. The upper panel shows the stellar radial velocity when the planet has a circular orbit (the symmetric sinusoidal curve), and when the planet has an eccentric orbit with $e = 0.2, e = 0.4, e = 0.6$, and $e = 0.8$. In all cases the longitude of pericenter $\varpi = \pi/4$. The lower panel shows, for a planet with $e = 0.5$, how the stellar radial velocity varies with the longitude of pericenter.

of equal period, a planet on an eccentric orbit produces a stellar radial velocity signal of greater amplitude, but there are also long periods near apocenter where the gradient of v_* is rather small. These two properties of eccentric orbits mean that, depending upon the observing strategy employed, a radial velocity survey can be biased either in favor of or against finding eccentric planets. Such bias, however, is not a major concern for current samples, and the most important selection effects are those already discussed for circular orbits.

Noise Sources

The amplitude of the radial velocity signal produced by Jovian analogs in extrasolar planetary systems is of the order of 10 m s^{-1}. High resolution astronomical spectrographs operating in the visible part of the spectrum have resolving powers $R \sim 10^5$, which correspond to a velocity resolution $\Delta v \approx c/R$ of a few km s^{-1}. The Doppler shift in the stellar spectrum due to orbiting planets therefore results in a periodic translation of the spectrum on the detector by a few *thousandths* of a

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

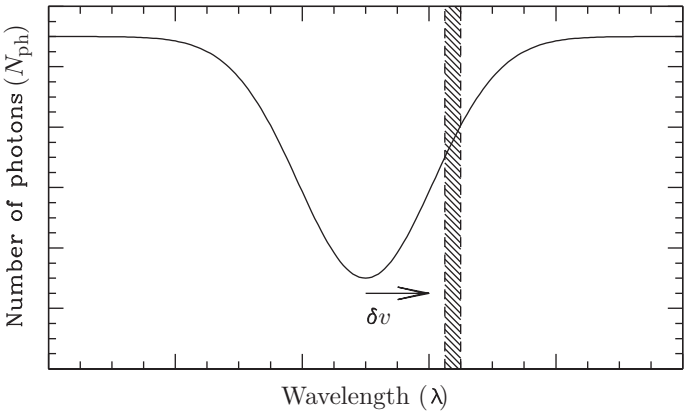

Figure 1.7 Schematic spectrum in the vicinity of a single spectral line of the host star. The wavelength range that corresponds to a single pixel in the observed spectrum is shown as the vertical shaded band. If the spectrum shifts by a velocity δv the number of photons detected at that pixel will vary by an amount that depends upon the local slope of the spectrum.

pixel.[4] Detecting such small shifts reliably requires both exceptional instrumental stability (extending, for long period planets, over periods of many years) and careful consideration of the potential sources of noise in the measurement of the radial velocity signal.

Shot noise (i.e. uncertainty in the number of photons due purely to counting statistics) defines the ultimate radial velocity precision that can be attained from an observation of specified duration. An estimate of the shot noise limit can be derived by starting from a very simple problem: how accurately can the velocity shift of a spectrum be estimated given measurement of the flux in a single pixel on the detector? To do this, we follow the basic approach of Butler *et al.* (1996) and consider the spectrum in the vicinity of a spectral line, as shown in Fig. 1.7. Assume that in an observation of some given duration, $N_{\rm ph}$ photons are detected in the wavelength interval corresponding to the shaded vertical band. If we now imagine displacing the spectrum by an amount (expressed in velocity units) δv the change in the mean number of photons is

$$\delta N_{\rm ph} = \frac{{\rm d}N_{\rm ph}}{{\rm d}v}\delta v. \tag{1.33}$$

Since a 1σ detection of the shift requires that $\delta N_{\rm ph} \approx N_{\rm ph}^{1/2}$, the minimum velocity displacement that is detectable is

$$\delta v_{\rm min} \approx \frac{N_{\rm ph}^{1/2}}{{\rm d}N_{\rm ph}/{\rm d}v}. \tag{1.34}$$

[4] For simplicity, we assume that one spectral resolution element corresponds to one pixel on the detector. A real instrument may over-sample the spectrum, but this practical point does not alter any of the basic results.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

This formula makes intuitive sense. Regions of the spectrum that are flat are useless for measuring δv while sharp spectral features are good. For solar-type stars with photospheric temperatures $T_{\rm eff} \approx 6000$ K the sound speed at the photosphere is around 10 km s^{-1}. Taking this as an estimate of the thermal broadening of spectral lines, the slope of the spectrum is at most

$$\frac{1}{N_{\rm ph}}\frac{{\rm d}N_{\rm ph}}{{\rm d}v} \sim \frac{1}{10\ {\rm km\,s^{-1}}} \sim 10^{-4}\ {\rm m^{-1}\,s}. \tag{1.35}$$

Combining Eqs. (1.34) and (1.35) with knowledge of the number of photons detected per pixel yields an estimate of the photon-limited radial velocity precision. For example, if the spectrum has a signal to noise ratio of 100 (and there are no other noise sources) then each pixel receives $N_{\rm ph} \sim 10^4$ photons and $\delta v_{\rm min} \sim 100$ m s^{-1}. If the spectrum contains $N_{\rm pix}$ such pixels the combined limit to the radial velocity precision is

$$\delta v_{\rm shot} = \frac{\delta v_{\rm min}}{N_{\rm pix}^{1/2}} \sim \frac{100\ {\rm m\,s^{-1}}}{N_{\rm pix}^{1/2}}. \tag{1.36}$$

Although this discussion ignores many aspects that are practically important in searching for planets from radial velocity data, it suffices to reveal several key features. Given a high signal to noise spectrum and stable wavelength calibration, photon noise is small enough that a radial velocity measurement with the meters per second (m s^{-1}) precision needed to detect extrasolar planets is feasible. The resolution of the spectrograph needs to be high enough to resolve the widths of spectral lines, but does not need to approach the magnitude of the planetary signal. In fact the intrinsic precision of the method depends first and foremost on the amount of structure that is present within the stellar spectrum, and the measurement precision will be degraded for stars whose lines are additionally broadened, for example by rotation.

Once precisions of the order of m s^{-1} have been attained, intrinsic radial velocity jitter due to motions at the stellar photosphere presents a second limit. The vertical velocity at the photosphere of a star is not zero, due to the presence of both convection and p-mode oscillations (acoustic modes trapped in the star). Although p-mode oscillations are of great intrinsic interest for helio- and asteroseismology, when under-sampled they represent noise for radial velocity searches for extrasolar planets. The effects of resolved stellar oscillations can be removed using appropriate observing strategies (Dumusque *et al.*, 2011), allowing radial velocity measurements with a precision that approaches the 10 cm s^{-1} needed to find Earth analogs.

1.7.3 Astrometry

Astrometric measurement of the stellar reflex motion in the plane of the sky provides a complementary method for detecting planets. From the definition of the

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

center of mass, the physical size of the stellar orbit is related to the planetary semi-major axis via $a_* = (M_\mathrm{p}/M_*)a$. For a star at distance d from the Earth the angular displacement of the stellar photo-center during the course of an orbit has a characteristic scale,

$$\theta = \left(\frac{M_\mathrm{p}}{M_*}\right)\frac{a}{d}. \tag{1.37}$$

Unlike radial velocity searches, which are biased toward detecting short period planets, astrometry favors large semi-major axes. A further difference is that astrometry measures two independent components of the stellar motion (versus a single component via radial velocity measurements), and this yields more constraints on the orbit. As a result there is no $\sin i$ ambiguity and all of the important planetary properties can be directly measured. Numerically the size of the signal is

$$\theta = 5 \times 10^{-4}\left(\frac{M_\mathrm{p}}{M_\mathrm{J}}\right)\left(\frac{M_*}{M_\odot}\right)^{-1}\left(\frac{a}{5\ \mathrm{AU}}\right)\left(\frac{d}{10\ \mathrm{pc}}\right)^{-1}\ \mathrm{arcsec}. \tag{1.38}$$

Even though the parameters adopted here are rather optimistic this is still a very small displacement, and none of the planets found in the first decade of discovery were identified this way. In principle, however, there are no fundamental obstacles to achieving astrometric accuracies of 1–10 μarcsec, which is good enough to detect a wide range of hypothetical planets. A $10M_\oplus$ planet at 1 AU, for example, yields an astrometric signature of 3 μarcsec at $d = 10$ pc.

1.7.4 Transits

A planet whose orbit causes it to transit the stellar disk can be detected by monitoring the stellar flux for periodic dips. At the most basic level transit events can be characterized in terms of their depth (the fraction of the stellar flux that is obscured), duration, and probability of being observed by a randomly oriented external observer. These quantities are readily estimated. The depth or amplitude of the transit signal is independent of the distance between the planet and the star, and provides a measure of the relative size of the two bodies. If a planet of radius R_p occults a star of radius R_*, with the geometry shown in Fig. 1.8, the fractional decrement in the stellar flux during the transit (assuming a uniform brightness stellar disk) is just

$$f = \left(\frac{R_\mathrm{p}}{R_*}\right)^2. \tag{1.39}$$

Gas giant planets have a rather flat mass–radius relation, so it is reasonable to use Jupiter's radius $R_\mathrm{J} = 7.142 \times 10^9$ cm as a proxy for all giant planet transits. The

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

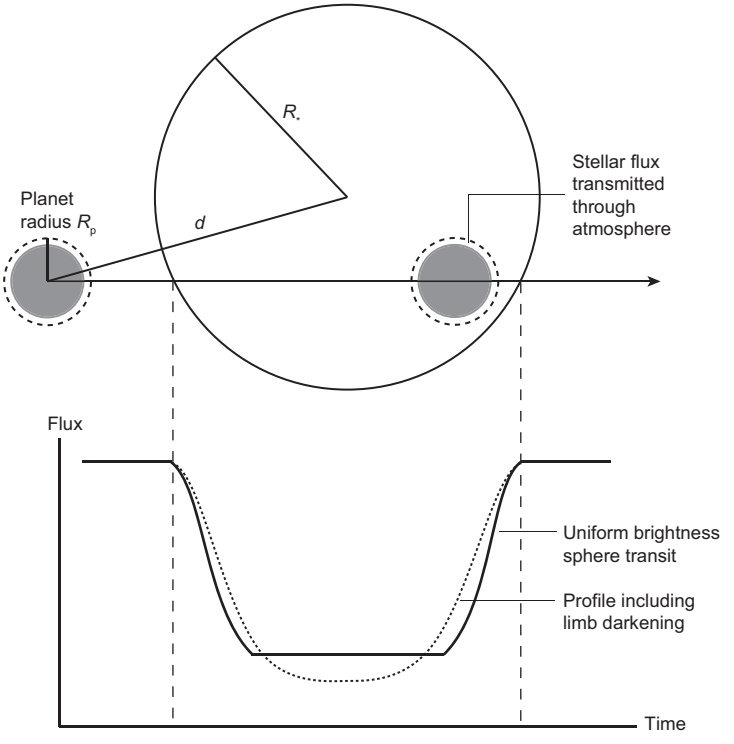

Figure 1.8 The geometry and light curve of a stellar transit by a planet (not to scale). The stellar and planetary radii are R_* and R_p respectively, and the distance between the projected centers of the bodies is d. The simplest model for the transit light curve corresponds to a perfectly opaque disk moving across a circle of uniform brightness. Deviations from this light curve occur due to the phenomenon of limb darkening. The photosphere is physically deeper looking toward the center of the disk, so we see hotter gas and greater intensity there than close to the limb. A small fraction of the starlight passes *through* the atmosphere of the planet, allowing measurement of the atmospheric constituents via transit spectroscopy.

amplitude of the signal for a gas giant transiting a solar-radius star is $f = 0.01$. For rocky planets with the same mean density as the Earth,

$$f = 8.4 \times 10^{-5} \left(\frac{M_\mathrm{p}}{M_\oplus} \right)^{2/3}. \tag{1.40}$$

The probability that a planet will be observed to transit follows from elementary geometric arguments. For a planet on a circular orbit with semi-major axis a, the condition that some part of the planet will be seen to graze the stellar disk is that the inclination angle i satisfies $\cos i \leq (R_* + R_\mathrm{p})/a$. If an ensemble of such systems has random inclinations then the probability of observing transits is

$$P_\mathrm{transit} = \frac{R_* + R_\mathrm{p}}{a}. \tag{1.41}$$

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

For a planet similar to the Earth, the probability is $P_{\rm transit} \approx 5 \times 10^{-3}$. If the geometry is favorable for observing transits, their maximum duration is roughly $2R_*/v_{\rm K}$. More accurately (Quirrenbach, 2006),

$$t_{\rm transit} = \frac{P}{\pi} \sin^{-1}\left(\frac{\sqrt{(R_* + R_{\rm p})^2 - a^2\cos^2 i}}{a}\right), \tag{1.42}$$

where P is the orbital period. A planet similar to the Earth whose orbit is seen edge-on ($i = 90^\circ$) transits its host star for 13 hours. Combining this result with the transit probability, one finds that if every star were to host an Earth-like planet, about 1 in 10^5 stars would be being transited at any given time.

The predicted light curve for an idealized (noise-free) planetary transit can be computed using various approximate or empirical models for the brightness distribution of the stellar disk. The simplest model treats the planet as a perfectly occulting disk passing across the face of a star whose brightness or specific intensity is everywhere constant. Following Mandel and Agol (2002) we set up the problem as shown in Fig. 1.8. The planet and star have radii $R_{\rm p}$ and R_* respectively, and the instantaneous separation of the centers of the two bodies is d. We define auxiliary variables $z \equiv d/R_*$ and $p = R_{\rm p}/R_*$. In terms of these variables the fractional decrement f is given as

$$f(p,z) = \begin{cases} 0 & 1 + p < z, \\ \frac{1}{\pi}\left[p^2\kappa_0 + \kappa_1 - \sqrt{\frac{4z^2 - (1+z^2-p^2)^2}{4}}\right] & |1-p| < z \le 1 + p, \\ p^2 & z \le 1 - p, \\ 1 & z \le p - 1. \end{cases} \tag{1.43}$$

In these expressions,

$$\kappa_0 = \cos^{-1}\left[\frac{p^2 + z^2 - 1}{2pz}\right], \tag{1.44}$$

$$\kappa_1 = \cos^{-1}\left[\frac{1 - p^2 + z^2}{2z}\right]. \tag{1.45}$$

Unfortunately this model, which is simple enough to be written entirely in terms of elementary functions, is not very realistic. Real transit light curves do not show the perfectly flat bottoms with $f = p^2$ that would be predicted for a black disk crossing a star of uniform brightness. The problem is limb darkening. The line of sight to the star is perpendicular to the stellar surface at the center of the disk, but increasingly tangential as we move toward the limb. As a consequence, photons coming from the center of the star last scattered, on average, at a greater physical depth than those coming from near the limb. Greater physical depth corresponds to a higher temperature, so the central brightness is greatest and the limb regions are relatively darkened. There is no exact first principles description of limb darkening,

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

but a variety of empirical laws fit observational data well. Mandel and Agol (2002) provide more complex but still analytic light curves for some of these laws.

Noise sources for transit searches include photon shot noise, stellar variability (including a contribution from the asteroseismic modes), and, for ground-based experiments, atmospheric fluctuations. For stars of roughly solar radii ground-based experiments are an efficient means of finding planets with $R_{\rm p} \simeq R_{\rm J}$, while smaller planets with $R_{\rm p} \simeq R_{\oplus}$ are only detectable given the lower noise levels attainable from space. Ground-based surveys are sensitive to roughly Earth-radius planets around M-dwarfs, which are physically much smaller. An example is GJ 1132b, a 1.2 $R_{\oplus}$ planet orbiting a 0.18 $M_{\odot}$ star that was discovered from the ground (Berta-Thompson *et al.*, 2015). In addition to these true noise sources all transit surveys have to contend with false positives that are caused by planetary transit-like signals of alternate astronomical origin. A stellar eclipsing binary, for example, may source a signal that looks like a planet if its light is diluted by the presence of a third unresolved star within the photometric aperture. A combination of follow-up observations, yielding radial velocity or higher spatial resolution adaptive optics data, and statistical arguments are used to discriminate between such imposters and true planets.

Detection of repeated planetary transit signals provides an immediate measure of the orbital period, and of the ratio between the physical size of the planet relative to the star. In cases where multiple planets are observed to transit the same star, gravitational perturbations between the planets lead to small changes in the time at which transits occur. When detectable these *Transit-Timing Variations* (TTVs) provide constraints and in some cases measurements of the planetary masses (Agol *et al.*, 2005; Holman & Murray, 2005; Holczer *et al.*, 2016). Planetary masses can also be determined for systems in which transit data are supplemented by radial velocity measurements. In this case knowledge of the inclination removes the usual $\sin i$ uncertainty in the mass. Time-resolved measurement of the radial velocity signal *during* the transit also opens up the possibility of detecting the Rossiter–McLaughlin effect (originally discussed in the context of eclipsing stellar binaries; Rossiter, 1924; McLaughlin, 1924). The Rossiter–McLaughlin effect is a perturbation to the apparent stellar radial velocity that is caused by the fact that when a planet transits across the face of a rotating star, it obscures portions of the stellar photosphere that are either redshifted or blueshifted relative to the whole-disk average. Figure 1.9 illustrates the principle. Detection of the effect allows for a measurement of the projected obliquity, on the plane of the sky, between the stellar and orbital angular momentum vectors. This, in turn, informs and constrains theoretical models for the origin of planets in short-period orbits.

Transiting systems are also a rich source of informative data on planetary atmospheres. In some cases one can detect the secondary eclipse when the planet moves behind the star, and in others it is possible to measure the out-of-eclipse phase modulation as the planet orbits. Depending upon the wavelength of the observations,

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

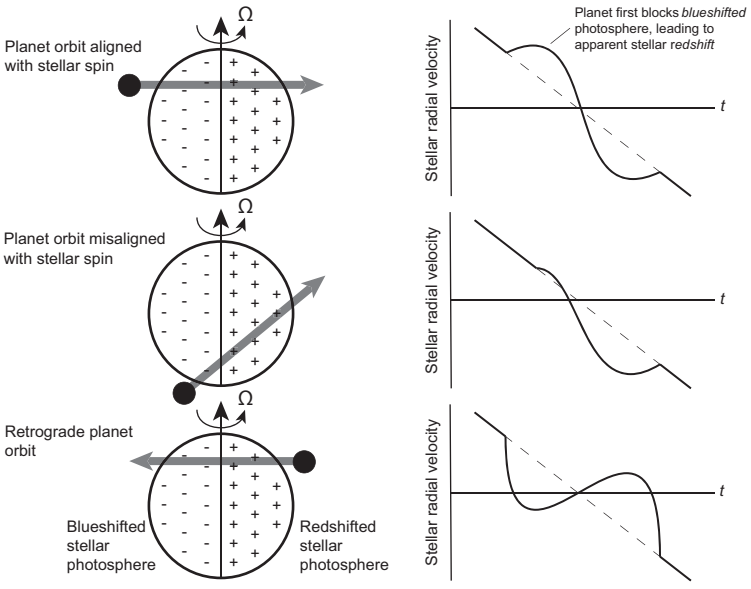

Figure 1.9 Illustration of the Rossiter–McLaughlin effect. Selective obscuration of the photosphere of a rotating star introduces a time-dependent perturbation to the measured radial velocity signal as a planet transits the star. The form of the perturbation depends upon whether the planet first blocks a redshifted or blueshifted piece of photosphere, and hence probes for misalignment between the stellar and orbital angular momentum vectors.

data of this type can constrain the albedo of the planet or the size of the difference between the temperature of the day side and that of the night side. Even more powerful are observations of the apparent size of the planet at different wavelengths, due to light that passes through the atmosphere at the planet's limb. This is a small effect. For a planet whose atmosphere has temperature T, mean molecular weight μ, and surface gravity g, the exponential scale-height is

$$H_{\rm p} = \frac{k_{\rm B} T}{\mu g}, \tag{1.46}$$

where $k_{\rm B}$ is the Boltzmann constant. Suppose that we observe a planet at two wavelengths, one where the atmosphere is entirely transparent (so that the apparent planetary size is given by that of the solid surface) and one where an atmospheric absorber is optically thick up to n scale-heights above the surface. The transit depth measured at the two wavelengths will differ by an amount

$$\begin{aligned} \delta f &= \frac{(R_{\rm p} + n H_{\rm p})^2}{R_*^2} - \frac{R_{\rm p}^2}{R_*^2} \\ &\simeq \frac{2 n R_{\rm p} H_{\rm p}}{R_*^2}. \end{aligned} \tag{1.47}$$

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

For an Earth analog with $H_p = 8$ km orbiting a solar-type star $\delta f \sim 2 \times 10^{-7}$, and detecting this effect is close to 1000 times harder than simply seeing the transit itself. Larger signals, which are detectable today, occur for larger or hotter planets, and for those orbiting smaller stars. In these cases measurement of the apparent planetary radius as a function of observing wavelength provides a handle on the atmospheric composition.

1.7.5 Gravitational Microlensing

Gravitational microlensing is a powerful method for detecting extrasolar planets that is based upon indirectly detecting the gravitational deflection of light that passes through the planetary system from a background source. The foundations of the method were laid by Einstein, who derived the general relativistic result for light bending by gravitating objects. A photon that passes a star of mass M_* with impact parameter b is deflected by an angle

$$\alpha = \frac{4GM_*}{bc^2}. \tag{1.48}$$

If two stars lie at different distances along the same line of sight, consideration of the geometry illustrated in Fig. 1.10 implies that the image of the background star (the "source") is distorted by the deflection introduced by the foreground star (the "lens") into a ring. Writing the distance between the observer and the lens as d_L, the observer–source distance as d_S, and the lens–source distance as d_{LS}, the angular radius of the so-called *Einstein ring* is

$$\theta_E = \frac{2}{c}\sqrt{\frac{GM_* d_{LS}}{d_L d_S}}. \tag{1.49}$$

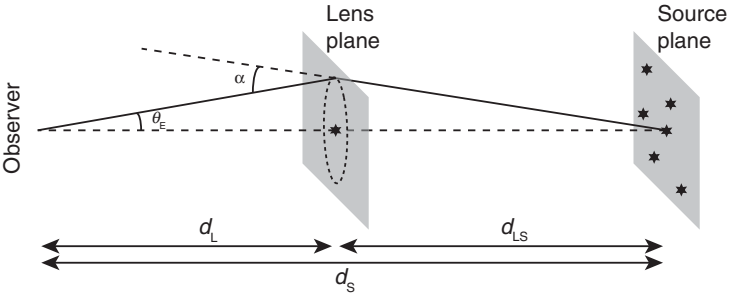

Figure 1.10 Microlensing geometry. Light from a background star is deflected by a small angle α when it passes a foreground star of mass M_*. If the alignment between the foreground and background star is perfect, then by symmetry the observer sees the image of the background star distorted into a circular Einstein ring around the foreground star. The ring has an angular radius θ_E. When lensing is used as a tool to detect planets, the distances involved are such that the ring is unresolved. The observable is the change in brightness of the background star as the two stars first align, and then move apart.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

If the alignment between the source and the lens is nearly but not quite perfect, the axial symmetry is broken and the source is lensed into multiple images whose angular separation is also of the order of θ_E.

For planet detection, a typical configuration of interest observationally is that where the source star lies in the Galactic bulge ($d_S = 8\,\mathrm{kpc}$) and the lens star is in the disk ($d_L \approx 4\,\mathrm{kpc}$). The Einstein ring radius for a solar mass lens with these parameters is $\theta_E \sim 10^{-3}$ arcsec, so the multiple images cannot be spatially resolved. A property of lensing, however, is that it conserves surface brightness. The area of the lensed image is greater than it would have been in the absence of lensing, and hence the flux from the lensed images exceeds that of the unlensed star. This means that if the alignment between the source and the lens varies with time, monitoring of the light curve of the source star can detect unresolved lensing via the time-variable magnification caused by lensing. Light curves in which a single disk star lenses a background star are smooth, symmetric about the peak, and magnify the source achromatically. If the proper motion of the lens relative to that of the source is μ, the characteristic time scale of the lensing event is $t_E = \theta_E/\mu$. For events toward the Galactic bulge this time scale is about a month, scaling with the lens mass as $\sqrt{M_*}$.

Microlensing light curves are altered when the lens star is part of a binary. For star–planet systems in which the mass ratio $q = M_p/M_* \ll 1$ the conceptually simplest case to consider is when the planet orbits close to the physical radius of the Einstein ring at the distance of the lens. If one of the multiple images of the source passes close to the planet during the lensing event the planet's gravity introduces an additional perturbation to the light curve. To an order of magnitude, the time scale of the perturbation is just $t_p \sim q^{1/2} t_E$, while the probability that the geometry will be such that a perturbation occurs is $P \sim Aq^{1/2}$, where A is the magnification of the source at the moment when the image passes near the planet. A second channel for planet detection occurs in rare high magnification events, during which planets close to the Einstein ring modify the light curve near peak regardless of their orbital position. This channel is valuable observationally since high magnification events can be anticipated based on photometric observations made well before the peak, allowing for detailed monitoring of those events that are most favorable for detecting planets.

Determining the properties of planets from the analysis of microlensing light curves is a difficult inverse problem, but the above discussion is enough to explain the attractive aspects of the method. First, the physical radius of the Einstein ring for disk stars lensing background stars in the Galactic bulge corresponds to a few AU. Lensing is therefore well suited to detect planets at relatively large orbital radii (roughly corresponding to the location of the snow line in the Solar System) which are challenging to detect via radial velocity or transit methods. Second, the weak $\sqrt{M_P}$ dependence of the time scale on planet mass allows for a wide range of planets to be detected. In particular, Jupiter mass planets yield a perturbation time scale of around a day, while Earth mass planets have a characteristic time scale of about an hour. Both time scales are quite accessible observationally. Moreover,

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

when a planetary perturbation occurs its magnitude can be significant, even for Earth mass planets. Low mass planets can therefore be detected from the ground via gravitational microlensing, though the superior precision and uninterrupted viewing possible from space affords large advantages.

1.8 Properties of Extrasolar Planets

A central result of exoplanet studies is the extraordinary diversity of observed extrasolar planetary systems. Figure 1.11 illustrates this property by plotting the orbital radii of planets in some celebrated systems on a logarithmic scale. Solar System planets are restricted to about two orders of magnitude in semi-major axis. Extrasolar planets have been found across four orders of magnitude in orbital radius (and are particularly abundant close-in), though whether this full range is populated in any individual system remains unknown. Planets around low mass stars are common, as are planets and planetary systems that violate the clear dichotomy between terrestrial and giant planets that is a feature of the Solar System.

It is sometimes implied that the discovery of extrasolar planetary systems that do not resemble the Solar System came as a complete surprise to observers and

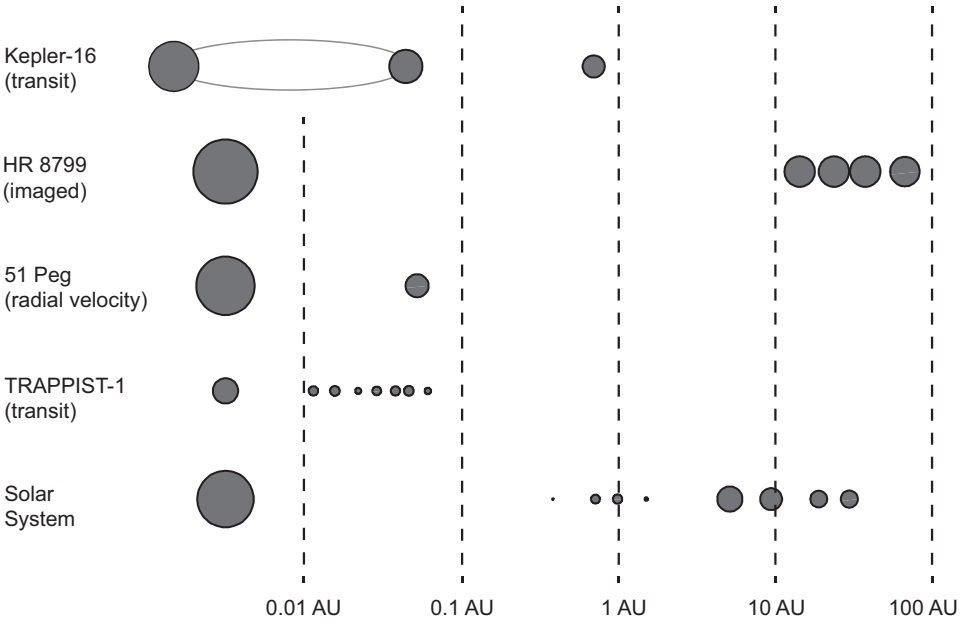

Figure 1.11 Observed planetary systems display diverse orbital architectures. Systems of moderately low mass planets around low mass stars – TRAPPIST-1 is a spectacular example (Gillon *et al.*, 2017) – are common, but there are also systems with circumbinary planets (illustrated here is the Kepler 16 system; Doyle *et al.*, 2011), with massive planets at orbital radii substantially larger than in the Solar System (HR 8799; Marois *et al.*, 2008), and with hot Jupiters (51 Peg b; Mayor & Queloz, 1995).

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

theorists alike. This is an over-statement. Otto Struve, in proposing a program of high-precision radial velocity measurements, suggested that it is "not unreasonable that a planet might exist at a distance of 0.02 AU" (Struve, 1952), and theorists expected that planets ought to be common (e.g. Wetherill, 1991) and identified in advance some of the mechanisms that lead to orbital migration (Goldreich & Tremaine, 1980). It is the extent of the differences between the Solar System and many extrasolar planetary systems that constitutes a genuine surprise. It has prompted both a re-evaluation of the effects of previously known planet formation processes, and a search for new ones.

1.8.1 Parameter Space of Detections

Figure 1.12 shows the distribution of semi-major axes for a sample of extrasolar planets with known masses. This plot, and most like it, is neither complete nor unbiased, and it omits most transit-detected planets as these often lack mass estimates. It suffices, however, to illustrate the main populations of known extrasolar planets.

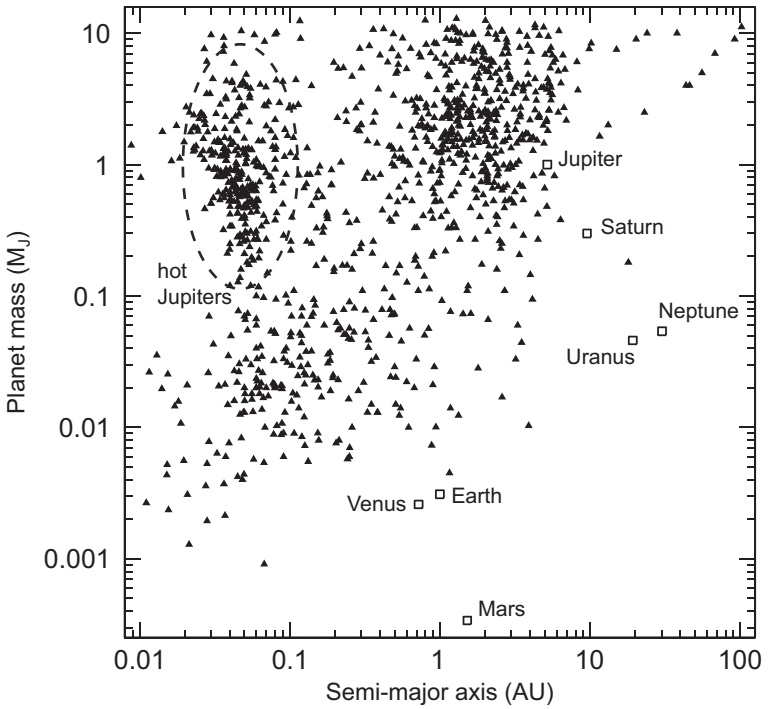

Figure 1.12 The distribution of a sample of known extrasolar planets in semi-major axis and planet mass (either M_p or $M_p \sin i$, depending upon the available observations). Selected Solar System planets are shown as the open squares. Data from the *NASA Exoplanet Archive*.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

In the realm of giant planets with masses equal to or larger than Saturn, a population of "warm" and "cold" Jupiters extends from a few tenths of an AU out to at least Jupiter's orbital radius. Knowledge of the true outer extent of this population is currently limited by selection effects. The masses of extrasolar giant planets range up to the deuterium burning limit (an order of magnitude larger than in the Solar System), but there are relatively few brown dwarfs with AU-scale orbits around solar mass stars, a property known as the "brown dwarf desert." Statistical studies show that giant planets with orbital periods $P <$ few yr orbit approximately 10% of Sun-like stars. Over the observed range of orbital radii the probability density is roughly uniform in log P. Low mass giants are more common than high mass ones. Cumming *et al.* (2008) estimated that $\mathrm{d}N_\mathrm{p}/\mathrm{d}\ln M \propto M^\alpha$ with $\alpha \simeq -0.3 \pm 0.2$ for this population.

Closer in the "hot Jupiters" have clearly distinct physical and dynamical properties, and possibly a different formation history, than their colder brethren. These planets orbit within about 0.1 AU of their hosts, and have correspondingly high effective temperatures. They have near circular orbits and are expected to be in tidally synchronized rotation states. Hot Jupiters are present around approximately 1% of solar-type stars, and are less common around M dwarfs (Dawson & Johnson, 2018).

Lower mass planets, although under-represented in Fig. 1.12, are substantially more common. Statistics of this population were derived from the results of the *Kepler* mission, which monitored ≈150 000 stars of spectral type FGKM for transits for 4 years. Roughly half of Sun-like stars have planets with radii $1-4\ R_\oplus$ with orbital periods $P \leq 1$ yr. Closely spaced systems with multiple planets in these relatively short-period orbits are common (Winn & Fabrycky, 2015).

A small number of stars are known to host massive planets with orbital radii substantially larger than any of the Solar System's giant planets. An example is HR 8799, a 1.5 $M_\oplus$ star that is still young enough to have a debris disk. HR 8799 has at least four planets, with masses of ∼5–10 M_J and orbital radii between 15 and 70 AU (Marois *et al.*, 2008).

Important parts of the parameter space of possible planetary systems remain poorly surveyed. Knowledge of the abundance of planets with masses or radii comparable to the Earth is extremely limited beyond ≈1 AU, while beyond 10 AU information even on giant planets is scant. A full picture of the architecture of typical extrasolar planetary systems is also lacking. At the time of writing most known close-in low mass planets were discovered using *Kepler* data, and orbit relatively faint stars that are hard to monitor for long-period radial velocity signals. As a consequence, knowledge of the relationship between low mass and giant planets in extrasolar planetary systems is more limited than one might expect given the amount of good data on each individual population. Basic questions, such as the average number of planets that orbit stars of different masses, remain open.

1.8.2 Orbital Properties

The eccentricity of giant extrasolar planets is a function of their orbital radius. The closest-in planets, including most of the hot Jupiters with $a \leq 0.1$ AU, have circular or low eccentricity orbits. A qualitatively similar result is found for binary stars, and in both regimes it is attributed to the circularizing influence of tides. At orbital radii where the effects of tides can be neglected, giant extrasolar planets have a broad eccentricity distribution which includes highly eccentric planets that have no Solar System analogs. HD 80606b, for example, has an eccentricity $e \simeq 0.93$ which in the Solar System would be more characteristic of comets than any larger bodies. Figure 1.13 shows the semi-major axes and eccentricity for a sample of planets discovered as a result of radial velocity searches. Excluding the hot Jupiters we see that small to moderate eccentricities $e = 0{-}0.3$ are common, and that there is a tail that extends to high values. Restricting the sample to about 400 planets with $0.3\,M_{\rm J} \leq M_{\rm p} < 10\,M_{\rm J}$ and $0.1\,{\rm AU} < a < 5\,{\rm AU}$ the mean eccentricity is $\langle e \rangle \simeq 0.25$, while the median $\langle e \rangle \simeq 0.19$.

The detailed interpretation of exoplanet eccentricities is complex, and various models have been proposed to explain the data. Most invoke gravitational dynamics,

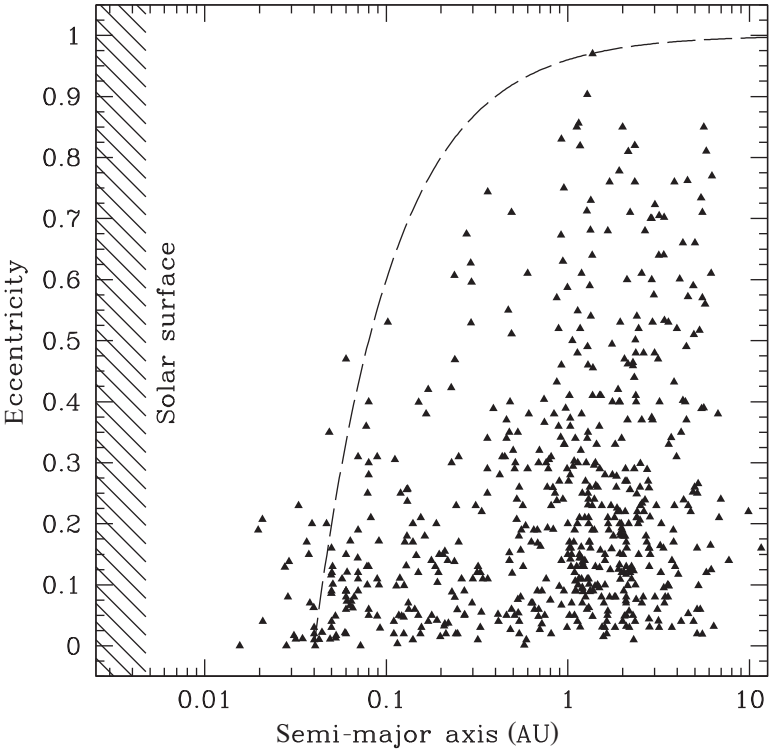

Figure 1.13 The distribution of a sample of extrasolar planets discovered via radial velocity searches in semi-major axis and eccentricity. The dashed line shows the eccentricity of a planet whose pericenter distance is 0.04 AU.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

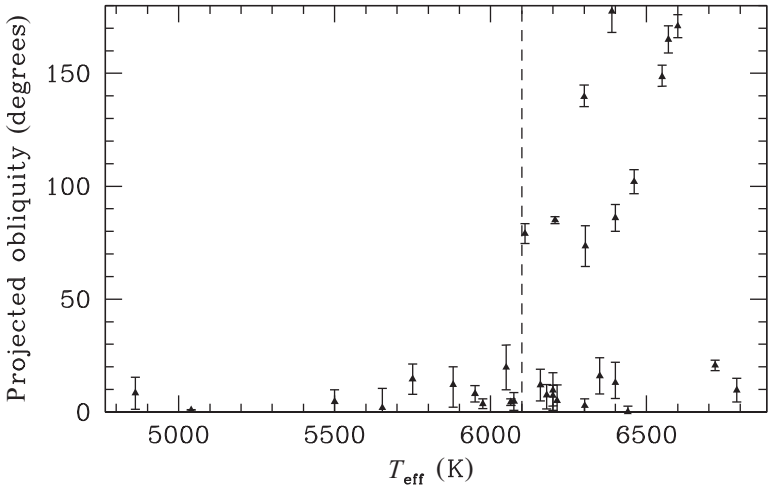

Figure 1.14 The sky-projected angle between the stellar spin axis and the orbital plane of close-in massive planets. The sample includes hot Jupiters with orbital period $P < 7$ days, masses between 0.3 and 3 $M_{\rm J}$, and quoted measurement errors of less than 10° (data from Wright *et al.*, 2011a).

taking place after the main phase of planet formation, to excite the eccentricity of initially near-circular giant planets. The observed distribution may also be affected by interactions (also mediated by gravity) between planets and gas that occur before or during the dispersal of the protoplanetary disk.

Measurements of the Rossiter–McLaughlin effect (Fig. 1.9) provide a constraint on how well planetary orbits align with the equators of their host stars. The observable quantity is not the misalignment itself but rather the *projected obliquity*, that is the angle between the vectors describing the orbit and the stellar spin projected on the sky plane. Figure 1.14 shows measurements of the projected obliquity for a sample of hot Jupiters orbiting stars of varying effective temperature $T_{\rm eff}$. A subset of hot Jupiters have high projected obliquities, which implies that some giant planets orbit their stars in highly tilted, polar, or even retrograde orbits. The observed distribution has a striking dependence on the stellar effective temperature. Low projected obliquities are the norm for $T_{\rm eff} \lesssim 6100$ K, while a wide range of obliquities characterize hot Jupiters around hotter stars (Winn *et al.*, 2010).

A small number of dynamical processes, including the Kozai–Lidov effect which we will discuss in Chapter 7, are able to tilt the orbits of planets that form initially in the plane defined by the stellar spin. They were fingered as prime suspects as soon as the first examples of misaligned hot Jupiters were identified. Alternatively, the disk plane itself could be tilted away from the equatorial one in some fraction of systems, leading to a primordial origin for the misalignments (Bate *et al.*, 2010; Batygin, 2012). In either case the fact that the threshold in effective temperature coincides with known structural differences in stars suggests that stellar rotation and tidal dissipation play a role in shaping the late-time distribution.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

Definitive determination of whether specific extrasolar planetary systems with multiple planets are resonant or not is often not possible. In many cases where a pair of planets are close to a commensurability the available data are ambiguous as to whether a true resonance exists. Nonetheless, some clear examples of resonant systems are known. Among massive extrasolar planetary systems three of the planets in the Gliese 876 system occupy a Laplace-type resonance in which the orbital periods are in the ratio of 4:2:1 (Rivera *et al.*, 2010; Millholland *et al.*, 2018). A number of resonances and resonant chains are also observed among lower mass planetary systems, including the TRAPPIST-1 system (Gillon *et al.* 2017).

As in the Solar System the existence of resonant configurations among extrasolar planets is strong evidence for the importance of dissipative processes at an earlier time in the systems' history. Statistically, however, resonant configurations are not the most common state, though there is a marked dependence on planet mass. Wright *et al.* (2011b) analyzed a sample of 43 well-characterized multiple planet systems that were discovered from radial velocity searches. Among this sample – dominated by massive planets – roughly a third showed evidence for low-order period commensurability. This fraction is substantially greater than would be expected by chance. In contrast the distribution of period ratios for the mostly lower-mass planets discovered by the *Kepler* mission, shown in Fig. 1.15, shows no preference for resonant configurations. Instead there is a modest enhancement in the number of systems with period ratios just *larger* than would be expected for low-order mean-motion resonances (Fabrycky *et al.*, 2014).

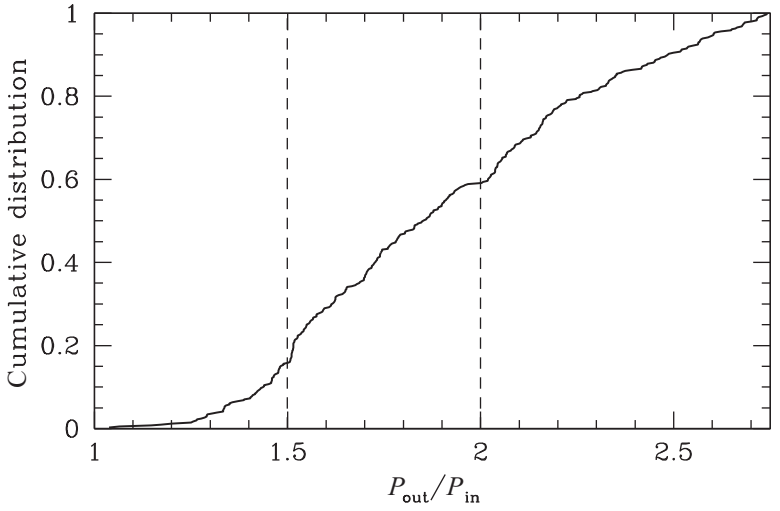

Figure 1.15 The cumulative distribution of period ratios for pairs of extrasolar planets detected via transit. Period ratios indicative of low-order mean-motion resonances are uncommon, but there is a modest excess of pairs with period ratios slightly larger than commensurabilities such as 3:2 and 2:1 (data from Fabrycky *et al.*, 2014).

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

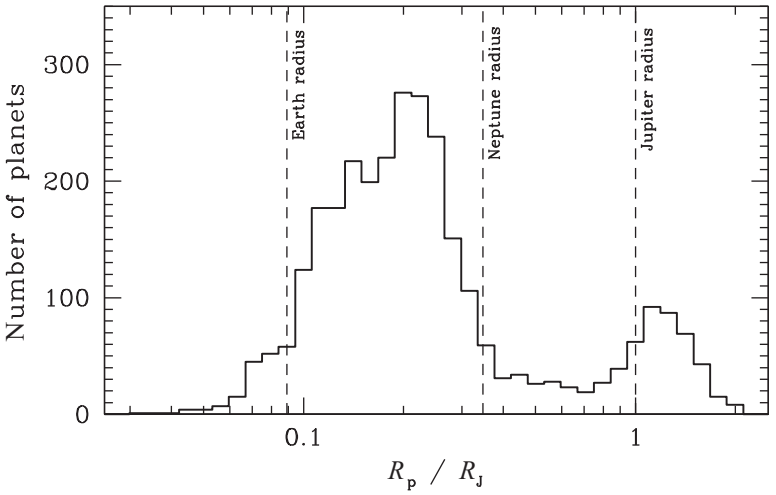

Figure 1.16 The distribution of radii for approximately 3000 extrasolar planets that were discovered via the transit method (using data from the *NASA Exoplanet Archive*). The data have not been corrected for completeness, and becomes increasingly incomplete for small planets. "Small" planets – those with radii intermediate between the Earth and Neptune – are substantially more abundant in transit surveys than gas giants. Many stars host detectable planets with radii that are intermediate between the Solar System's terrestrial and ice giant planets.

1.8.3 Mass–Radius Relation

The distribution of radii for planets discovered via the transit method is shown in Fig. 1.16. The data used to make this plot were not corrected for bias and incompleteness, but it is clear that the distribution is bimodal, and that "small" planets with radii of $R_p = 0.1{-}0.4\ R_J$ are substantially more abundant than gas giants with $R_p \simeq R_J$. What is *not* seen is a clear break between planets with $R_p \approx R_\oplus$, which Solar System experience would suggest are rocky terrestrial worlds, and planets with $R_p \approx R_{Nep}$, which in the Solar System are ice giants. Many stars host planets with radii that are midway between the Solar System's terrestrial and ice giant classes. The basic possibilities are that these are scaled-up versions of terrestrial planets ("super-Earths") or scaled-down versions of ice giants ("mini-Neptunes").

The physical nature of extrasolar planets of different sizes can be determined with greater confidence for the subset of planets whose masses (and hence densities) are also known. Unfortunately knowing the density does not yield a unique answer as to what the planet is made of. Planets with entirely different compositions can have identical masses and radii, making classification necessarily ambiguous. Several structures are well motivated by the existence of Solar System analogs and/or formation considerations.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

- **Gas giants** with masses roughly in the range between $M_{\rm Sat}$ and 13 $M_{\rm J}$. Theoretical models of gas giant structure show that the predicted radius is a weak function of mass, so to leading order gas giants ought to have $R_{\rm p} \simeq R_{\rm J}$. At the next order of approximation the radius is a decreasing function of the total mass of heavy elements within the planet, independent of whether that material is concentrated in a core or dispersed throughout the envelope.
- **Rocky planets.** There is no clear lower limit to the mass of rocky (or icy) worlds, though bodies smaller than a few hundred km in size are typically irregular in shape and fall into the class of minor bodies. There is also no absolute upper limit, though once we reach masses of 5–10 $M_\oplus$ capture of an envelope dense enough to increase the radius measured via transit becomes increasingly likely. Zeng *et al.* (2016) provide a semi-empirical model for the mass-radius relation of broadly Earth-like planets,

$$R_{\rm p} = (1.07 - 0.21\, f_{\rm core}) \left(\frac{M_{\rm p}}{M_\oplus} \right)^{1/3.7} R_\oplus, \tag{1.50}$$

  valid for the range between 1 and 8 Earth masses. In this expression $f_{\rm core}$ is the core mass fraction (for the Earth $f_{\rm core} \simeq 0.33$). Smaller radii are possible for iron-rich planets (Mercury-like and super-Mercury worlds), with the minimum physically plausible radii being attained for objects made up of 100% Fe.
- **Low mass core/envelope planets.** A variety of structures are possible within the generic class of a rocky or icy core (typically of a few to about ten $M_\oplus$) surrounded by a sub-dominant gaseous envelope. End members are "gas dwarfs," where the envelope is made up of hydrogen and helium, and objects where the envelope is heavily enriched with water and other species such as NH_3 and CH_4. Planets with these structural features would be seen as low density objects with masses of a few to about ten Earth masses.
- **Water worlds.** The high predicted abundance of water beyond the snow line in protoplanetary disks opens up the possibility of planets with a water dominated bulk composition. Such planets have a mass–radius relation that is similar to rocky planets, but 30–40% larger in size.

Figure 1.17 shows the observed masses and radii of a sample of extrasolar planets for which both quantities have been measured with moderate to good precision. Among giant planets the weak predicted dependence of radius on mass is clearly evident. A small number of giant planets have anomalously small radii, indicative of a large bulk heavy element abundance. Among the hot Jupiters, many planets have anomalously *large* radii, exceeding in some cases the prediction for a planet that lacks a significant mass of heavy elements. These radius anomalies correlate with the incident stellar irradiation, giving support to the hypothesis that some fraction of the incident stellar energy finds its way into the planets' convective zones. Energy input into the convection zone is known to stall contraction or inflate

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

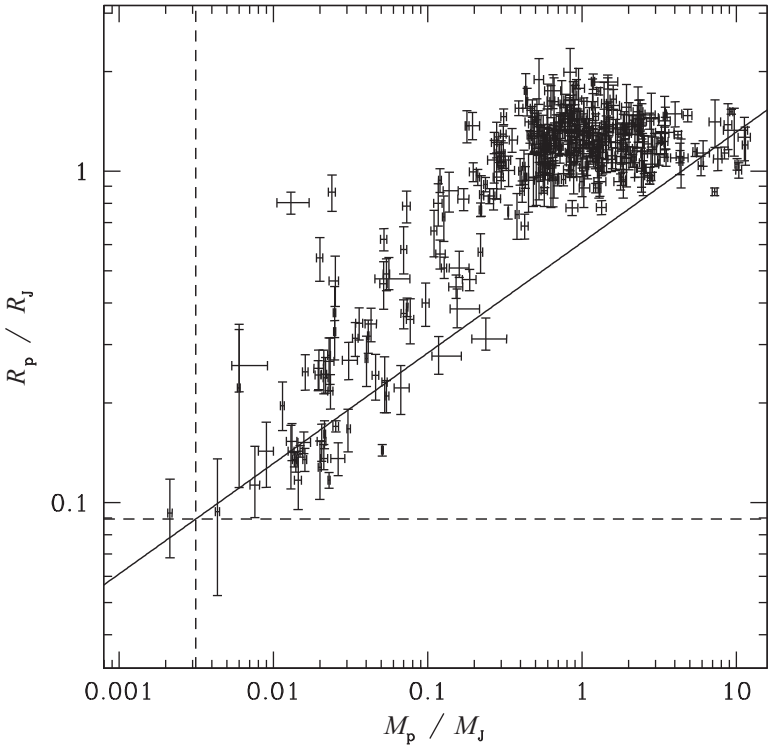

Figure 1.17 The distribution of a sample of extrasolar planets with well-determined mass and radius estimates.

giant planets, though the mechanism that injects stellar energy deep into the planet remains unclear.

Mass and radius data for smaller planets are sparser, especially for planets with masses similar to or smaller than that of the Earth. The evidence is consistent with the idea that many planets with $R_p \lesssim 1.5\ R_\oplus$ are rocky super-Earths, while larger planets have a range of radii that reflect variations in the amount of gas within their envelopes (Rogers, 2015). Not very much is known about the detailed composition of envelopes in the mini-Neptune size range, nor can it be excluded that some observed planets have water dominated compositions.

1.8.4 Host Properties

Important constraints on planet formation models come from the observation that the abundance of planets correlates with properties of the host star. A key observation was the discovery by Debra Fischer and Jeff Valenti that the fraction of stars with observed giant planets increases with the metal abundance measured spectrosopically in the stellar photosphere (Fischer & Valenti, 2005). Figure 1.18 shows an updated version of this correlation (Sousa *et al.*, 2011). For planets (and

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

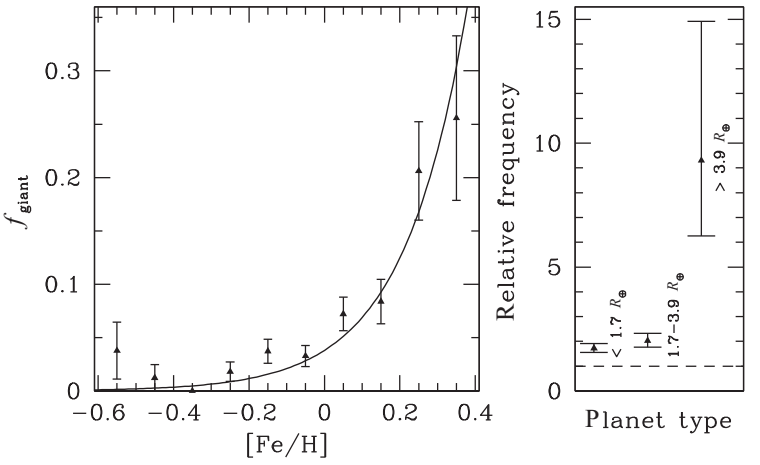

Figure 1.18 Left panel: the fraction of a sample of stars that host known giant extrasolar planets (using data from Sousa *et al.*, 2011). Right panel: the enhancement to the observed planet frequency for planets of different radii, corresponding approximately to super-Earths, mini-Neptunes, and gas giants. Note that astronomical convention is to measure the abundance of heavy elements relative to hydrogen, with the resulting ratio being normalized to the ratio found in the Sun. A logarithmic scale is used, denoted by square brackets. Thus, a star with "[Fe/H] = 0" has the same fractional abundance of iron as the Sun, whereas one with [Fe/H] = 0.5 is enriched in iron by a factor ∼3 as compared to the Sun.

a small number of more massive objects) with masses between 0.1 $M_{\rm J}$ and 25 $M_{\rm J}$ the detected planet frequency is found to rise steeply with metallicity, following a scaling law $P_{\rm planet} \propto Z^{2.58}$.

The planet–metallicity correlation was discovered from, and is strongest for, giant planets. Using a sample of around 400 *Kepler*-discovered planets, Wang and Fischer (2015) determined the extent to which being metal rich ([Fe/H] $>$ 0.05) increased the planet occurrence rate as opposed to being metal poor ([Fe/H] $<$ −0.05), for different planet sizes. The boost that a high metal content gives to planet abundance, shown as the right panel in Fig. 1.18, is substantially larger for planets with radii characteristic of gas giants than for smaller gas dwarf or terrestrial planets.

Because most of the heavy elements in a collapsing molecular cloud core end up in the star rather than in planets, the stellar photospheric metal abundance is reasonably indicative of the bulk enrichment of the protoplanetary disk that star once had. Under the assumption that the initial gas mass of disks is not a strong function of metallicity, the current stellar metallicity is then a marker for the mass of solid material that was once available to form planets. It is then certainly not surprising that higher stellar metallicities lead to a greater probability of planet formation, or to a greater number of planets in systems where more than one planet forms, as is observed. As we will discuss later, the fact that the planet–metallicity correlation is strongest for giant planets appears to be consistent with the idea that

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

giant planet formation involves a time-limited threshold step – to form a giant planet a massive *core* needs to assemble in a limited period of time before the gas in the protoplanetary disk disperses.

1.9 Habitability

Several factors appear to contribute to the hospitable environment that the Earth offers for life. The surface temperature and pressure allow for the presence of liquid water across much of the surface, and these conditions are maintained over time due to geological processes (volcanism and plate tectonics) that stabilize the climate. Convective motions within the Earth's core sustain a magnetic field, which reduces the rate at which water is lost from the upper atmosphere due to collisions with high energy solar wind particles. Together with a moderate global abundance of water, established at early times, the low loss rate means that the current surface conditions – with both major oceans and large landmasses – have remained roughly constant over billions of years. Dynamical effects may also be important. Our large Moon, for example, stabilizes the Earth's obliquity against perturbations from the other planets, and the Solar System's architecture does not lead to large excursions in the Earth's eccentricity.

A full accounting of the planetary conditions that are necessary and sufficient for life to survive is not known, and it is a more difficult task still – bordering on guesswork – to enumerate the characteristics of planets where life might form in the first place. It is therefore customary to adopt a deliberately limited definition of the "habitable zone" as the range of orbital distances where a planet of given mass, orbiting a star of specified mass or spectral type, can sustain liquid water on its surface. To leading order this requires that the incident stellar flux be similar to the 1360 W m^{-2} received by the Earth, and the steep scaling of stellar luminosity with stellar mass implies that the habitable zone has a strong dependence on host star mass. As shown in Fig. 1.19 a planet around a 0.2 $M_\odot$ star needs to orbit at roughly $a \simeq 0.1$ AU to experience similar insolation as the Earth, while for the lowest mass hydrogen-burning objects the habitable zone is at just a few hundredths of an AU.

A crude estimate of the width of the habitable zone can be obtained by ignoring atmospheric physics and assuming that a planet reradiates intercepted stellar radiation as a single temperature blackbody. For a planet with radius R_p, albedo A, and an orbital distance a from a star of luminosity L, the temperature given these assumptions would be determined by

$$4\pi R_\mathrm{p}^2 \sigma T^4 = \frac{L(1-A)}{4\pi a^2} \pi R_\mathrm{p}^2, \tag{1.51}$$

where σ is the Stefan–Boltzmann constant. Identifying the outer and inner edges of the habitable zone with the freezing and boiling points of water at the Earth's atmospheric pressure we estimate the relative width of the habitable zone to be

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

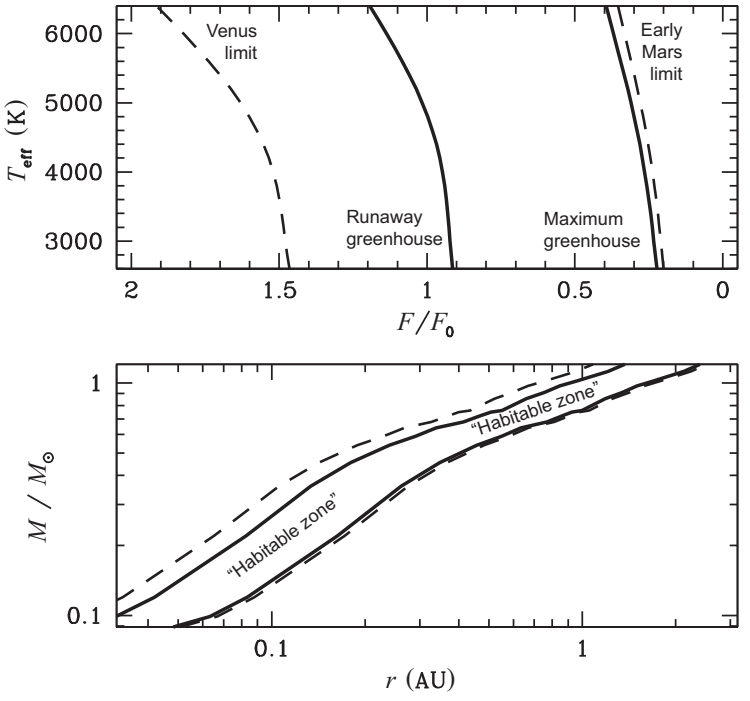

Figure 1.19 Estimates of the extent of the habitable zone, where liquid water can stably exist on the surface of a roughly Earth mass terrestrial planet. The upper panel, based on calculations by Kopparapu *et al.* (2013), shows theoretical estimates based on limiting cases of the greenhouse effect (solid lines) along with the extensions that follow if we assume that Mars and/or Venus were once but are no longer habitable (dashed lines). The width of the zone is expressed for stars of different effective temperature $T_{\rm eff}$ in terms of the incident planetary flux normalized to that of the Earth. The lower panels shows the same estimates converted into orbital radii for main-sequence stars of given mass.

$$\frac{a_{\rm out}}{a_{\rm in}} = \left(\frac{273\ {\rm K}}{373\ {\rm K}}\right)^{-2} \approx 1.9. \tag{1.52}$$

Although the neglect of the greenhouse effect means that the temperatures predicted by Eq. (1.51) are inaccurate, this estimate of the width of the zone is not too far off from actual calculations.

The use of climate models to map out the extent of the habitable zone was pioneered by Kasting *et al.* (1993). A climate model predicts the surface temperature of a planet given a specified atmospheric composition and stellar radiation field.[5] Geophysical assumptions are needed to define the limits of plausible

[5] In practice, calculations of the habitable zone often use an "inverse" climate model in which the surface temperature is specified and the model is used to compute the stellar flux that would yield that temperature.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

atmospheric compositions, which change over time. On the Earth the amount of CO_2 in the atmosphere is regulated by the carbonate–silicate cycle. Carbonate minerals react with SiO_2 under the high temperature conditions found in subduction zones to liberate CO_2, which is vented into the atmosphere via volcanic activity. Burton *et al.* (2013) estimate the magnitude of this source term to be about 0.5 MT yr^{-1}. Although dwarfed by current anthropogenic emissions this natural source of CO_2, if unchecked, would significantly increase the atmospheric budget of greenhouse gases and raise the global temperature on a time scale of the order of a Myr. The counter-balancing sink is provided by weathering. Atmospheric CO_2 reacts with rainwater to form carbonic acid, H_2CO_3, which dissolves silicate rocks on land. The weathering products flow into the oceans, where they ultimately reform carbonate minerals. Crucially, the weathering rate is an increasing function of the temperature, so it is possible (though not guaranteed) to reach an equilibrium state in which volcanic sources of CO_2 balance the weathering sink. In the extreme case where a planet becomes entirely ice covered (a "snowball Earth" state) the cessation of weathering and resulting build-up of CO_2 provides a mechanism to return the surface to habitability on a relatively short time scale.

The formation of calcium carbonate by living organisms in the oceans is an important part of the Earth's carbonate–silicate cycle, whose quantitative details would certainly be different on other planets. It seems reasonable, however, to assume that a qualitatively similar negative feedback loop acts to stabilize the climate of terrestrial planets that, like the Earth, have active plate tectonics, a large reservoir of carbon, and substantial but not dominant amounts of water. Within this class of planets the limits to habitability can be identified with catastrophic failures of climate stabilization mechanisms. As the stellar flux decreases for planets at larger orbital distance, the expected drop in the surface temperature is partially offset by the development of successively thicker CO_2 atmospheres which provide warming via the greenhouse effect. This stabilization fails once the atmosphere reaches a "maximum greenhouse" limit, beyond which CO_2 starts to condense out and is unable to provide additional warming. Calculations by Kopparapu *et al.* (2013) for a solar-type star place the maximum greenhouse limit at $F \simeq 0.35 F_0$ (where F_0 is the flux received by the Earth), corresponding to an orbital radius of 1.7 AU. Moving in the other direction, a closer-in analog of the Earth can maintain a stable but hotter climate until it reaches the "moist greenhouse" limit, when the stratosphere becomes dominated by water. Water can then be lost from the planet via a combination of dissociation of water molecules and escape of the light H atoms. The time scale for water loss becomes less than the age of the Earth for a surface temperature of $T \approx 340$ K, so this limit is more stringent than the simple requirement that the temperature remain below the boiling point of water. The oceans are expected to evaporate in their entirety at the "maximum

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

greenhouse" limit, beyond which CO_2 can accumulate further in the absence of efficient sink processes. The nominal moist and runaway greenhouse limits for the Solar System are at 0.99 AU and 0.97 AU (Kopparapu *et al.*, 2013), although the cooling effect of clouds (which are not included in these calculations) means that the Earth is not as close to the inner edge of the habitable zone as these estimates would suggest.

Figure 1.19 shows how the inner and outer boundaries of the habitable zone are predicted to scale with stellar mass. The leading order effect, which dominates the plot of $a_{\rm out}(M_*)$ and $a_{\rm in}(M_*)$ in the lower panel, is the scaling of stellar luminosity with mass. Smaller but still significant effects occur because of the shift of the peak of the stellar spectrum with mass. Cooler stars emit a greater fraction of their bolometric luminosity in the near-IR as compared to the visible, and this near-IR flux is both scattered less (Rayleigh scattering is proportional to λ^{-4}) and absorbed better (by species such as H_2O and CO_2) by planetary atmospheres. An Earth analog around a lower mass star would therefore have a lower albedo, and both the inner and outer edges of the habitable zone would be pushed outward. The magnitude of the shift, expressed in terms of the threshold flux F/F_0, is about 20–25% for stars ranging in mass between 0.1 $M_\odot$ and 1 $M_\odot$.

Opportunities to assess the validity of theoretical estimates of the habitable zone are for now extremely limited. Mars cannot sustain liquid water on its surface today, and hence is not within the habitable zone by our definition. The presence of river valleys and canyons on the Martian surface, however, strongly suggests that liquid water flowed on Mars in the early history of the planet, and may have been stable until atmospheric loss under the low gravity conditions led to cooling. This line of reasoning is consistent with the theoretical expectation that Mars' orbit lies close to the outer edge of the habitable zone. Much less can be said about the inner edge of the habitable zone. Venus is assuredly not habitable today, but there is no direct way of knowing whether the planet had liquid water at an early time when the solar luminosity was smaller. Empirically, then, the inner edge of the habitable zone lies beyond the orbit of Venus, but could be significantly closer to the Sun than the theoretical estimate shown in Fig. 1.19.

It is worth emphasizing again that the conventional definition of the habitable zone does not necessarily match up to the plain English meaning of habitability. Planets within the habitable zone may be inhospitable to life, for example because they possess vanishing quantities of water, while bodies far outside the zone could in principle harbor habitable sub-surface oceans. (There is strong evidence for such reservoirs being present on both Jupiter's moon Europa and Saturn's moon Enceladus.) For these reasons, the habitable zone is best viewed as defining a region of exoplanet parameter space that can reasonably be prioritized in follow-up atmospheric characterization studies focused on the identification of biomarkers such as oxygen.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press

1.10 Further Reading

The properties of the planets, moons, and minor bodies of the Solar System are discussed in depth in any planetary science text. A good example is *Planetary Sciences* by I. de Pater and J. J. Lissauer (updated second edition 2015, Cambridge: Cambridge University Press).

The *Handbook of Exoplanets* (editors H. J. Deeg and J. A. Belmonte, 2018, Cham: Springer) includes review articles on different exoplanet detection methods along with analyses on the properties of the exoplanet population.

https://doi.org/10.1017/9781108344227.002 Published online by Cambridge University Press