1

Historical notes on planet formation

Peter Bodenheimer

UCO/Lick Observatory, Santa Cruz

1.1 Introduction

The history of planet formation and detection is long and complicated, and numerous books and review articles have been written about it, e.g. Boss (1998a) and Brush (1990). In this introductory review, we concentrate on only a few specific aspects of the subject, under the general assumption that the Kant–Laplace nebular hypothesis provides the correct framework for planet formation. The first recognized "theory" of planet formation was the vortex theory of Descartes, which, along with related subsequent developments, is treated in Section 1.2. Magnetic effects (Section 1.3) were of great significance in the solution of one of the major problems of the nebular hypothesis, namely, that it predicted a very rapidly rotating Sun. The early histories of the two theories of giant planet formation that are under current debate, the disk gravitational instability theory and the core accretion-gas capture theory, are discussed in Section 1.4 and Section 1.5, respectively. In the final section, 1.6, certain specific examples in the history of the search for extrasolar planets are reviewed.

1.2 Descartes and von Weizsäcker: vortices

Descartes (1644) gave an extensive discussion on the formation of the Earth, planets, and major satellites, the main idea of which was that they formed from a system of vortices that surrounded the primitive Sun, in three-dimensional space. His picture did not involve a disk, the rotation axes of the vortices were not all in the same direction, and the physical basis for it is elusive. In a section of his book entitled "Concerning the creation of all of the Planets" he states:

Planet Formation: Theory, Observation, and Experiments, ed. Hubert Klahr and Wolfgang Brandner.
Published by Cambridge University Press.

"*. . . the extremely large space which now contains the vortex of the first heaven was formerly divided into fourteen or more vortices. . . So that since those three vortices which had at their centers those bodies that we now call the Sun, Jupiter, and Saturn were larger than the others; the stars in the centers of the four smaller vortices surrounding Jupiter descended toward Jupiter . . .*"

It is not specified how the vortices got there and how the planets formed from them, and the whole treatise may be regarded as more philosophical than scientific. Furthermore, he had to be very careful in what he said to avoid disciplinary action from the Church.

Three hundred years later, C. F. von Weizsäcker (1944) wrote an influential paper in which he envisioned a turbulent disk of solar composition rotating around the Sun, which consisted of a set of stable vortices out of which planets formed by accretion of small particles. Stability required that the vortices be counter-rotating, in the co-rotating coordinate system, to the direction of the disk's rotation. The radii of the various vortex rings corresponded roughly to the Titius–Bode law of planetary spacing. However, von Weizsäcker's main contribution was to recognize that in a turbulent disk there would be angular momentum transport as a result of turbulent viscosity, with mass flowing inward to the central object and with angular momentum flowing to the outermost material, which would expand. This realization solved one of the major problems of the Kant–Laplace nebular hypothesis – that the Sun would be spinning too fast. However the physical reality of the eddy system was criticized by Kuiper (1951) on the grounds that true turbulence involved a range of eddy sizes according to the Kolmogorov spectrum and that the typical lifetime of an eddy was too short to allow planet formation in it.

Vortices were not really taken seriously for a long time after that, until Barge and Sommeria (1995) showed that small particles could easily be captured in vortices and this process would accelerate planet formation. Although the presence of long-lived vortices was not rigorously proved, they suggested that particles could accumulate in the vortices to form the cores of the giant planets in 10^5 yr, thereby solving one of the major problems of the core accretion hypothesis. Later Klahr and Bodenheimer (2003), in a three-dimensional hydrodynamic simulation with radiation transfer, showed that under the proper conditions of baroclinic instability, vortices could form in low-mass disks. Further study of this process is strongly indicated.

1.3 Magnetic effects

Hoyle (1960) invoked magnetic braking to explain the slowly rotating Sun by transfer of angular momentum to the material that formed the planets. He claimed that purely hydrodynamic effects, such as viscosity, could not result in sufficient

transfer of angular momentum because the frictional effect requires that the disk material must be in contact with the Sun itself, and that therefore the Sun could be slowed only to the point where it was in co-rotation with the inner disk. He envisioned a collapsing cloud that is stopped by rotational effects and forms a disk. A gap opens between the contracting Sun and the disk, and a magnetic field, spanning the gap, transfers angular momentum from the Sun to the inner edge of the disk, forcing it outward. As the Sun contracts and tends to spin up, angular momentum continues to be transferred until the inner edge of the disk is pushed out far enough so that its orbital period is comparable with the present rotation period of the Sun. The temperature has to be ≈ 1000 K, to get magnetic coupling, and the field has to be ≈ 1 gauss. Beyond the inner edge of the disk he does not require the magnetic field to transfer angular momentum to the outer regions of the disk; viscosity would work in that case. The terrestrial planets form from refractory material that condenses out near the inner edge of the disk and becomes decoupled from the gas.

Actually Alfvén (1954) was the one who originally invoked magnetic braking, although his idea of how the Solar System formed was not considered very plausible. His theory did not involve a disk, but rather clouds of neutral gas of different compositions which fall toward the Sun from random directions, stopping at a distance where the ionization energy equals the infall kinetic energy (the so-called "critical velocity" effect). Once ionized, the material couples to the Solar magnetic field and angular momentum is transferred to it, forcing it outward and eventually into a disk plane. The elements with the lowest ionization potentials, such as iron and silicon, stop farthest out. The cloud, composed of hydrogen, along with elements of similar ionization potential such as oxygen and nitrogen, is envisioned to stop in the region of the terrestrial planets, while a cloud composed mainly of carbon stops at distances comparable to the orbital distances of the giant planets. The theory was criticized on the grounds that it did not explain the chemical composition of the planets, but in fact the crucial aspect of it was the magnetic braking.

Today it is known that even young stars are slowly rotating, and that the interface between disk and star in a young system is very complicated, involving accretion from disk to star as well as outflow in a wind. Modern theories (Königl, 1991; Shu *et al.*, 1994) show that the basic angular momentum-loss mechanism for the central star is magnetic transfer. However relatively large fields are required, of the order of 1000 gauss.

1.4 Gravitational instability

We now consider early developments in the theory of planet formation. The basic condition needed for the formation of a planet by gravitational instability in a gaseous disk goes back to Jeans (1929): in a medium of uniform density ρ and

uniform sound speed c_s a density fluctuation is unstable to collapse under self gravity if its wavelength λ satisfies the condition

$$\lambda^2 > \frac{\pi c_s^2}{G\rho}. \tag{1.1}$$

Although the physical assumptions leading to the derivation were inconsistent, this criterion still gives the correct approximate conditions for gravitational collapse.

Kuiper (1951) suggested that giant planets could form by this mechanism; he combined the Jeans criterion with the condition for tidal stability of a fragment in the gravitational field of the central star. He estimated that planets formed this way would have masses of the order of 0.01 Solar masses ($M_\odot$) and that the disk would need a mass of $\approx 0.1\ M_\odot$. He retained von Weizsäcker's idea of turbulence in the disk, suggesting that *"Turbulence may be thought of as providing the initial density fluctuations and gravitational instability as amplifying them,"* an idea that has been revived in the modern theory of star formation in a turbulent interstellar cloud.

Safronov (1960) and Toomre (1964) rederived the Jeans condition in a flat disk, including differential rotation, gravity, and pressure effects. If the sound speed is c_s, the epicyclic frequency κ, and σ the surface density of the disk (mass per unit area), then $Q = (c_s\kappa)/(\pi G\sigma) > 1$ for local stability to axisymmetric perturbations. Although Safronov was primarily interested in a flat disk of planetesimals, and Toomre was interested in a galactic disk of stars, and as stated, the derivation is valid only for axisymmetric perturbations, the "Toomre Q" is still a useful criterion even for stability to non-axisymmetric perturbations in gaseous disks. The critical value will depend on the equation of state and the details of the numerical code being used, but typically disks are stable if $Q > 1.5$.

Cameron (1969) later suggested that the protoplanetary disk, in the process of formation, could break up into axisymmetric rings which could then form planets by gravitational instability. There followed a series of evolutionary calculations for "giant gaseous protoplanets" which were assumed to have been formed by this mechanism (Bodenheimer, 1974; DeCampli and Cameron, 1979; Bodenheimer *et al.*, 1980). The general idea was that spherically symmetric condensations of approximately Jovian mass and Solar composition formed, in an unstable disk, with initial sizes of 1 to 2 AU, then contracted through an initial series of quasi-hydrostatic equilibria. The calculations involved the solution of the standard equations of stellar structure, including radiative and convective energy transport and grain opacities. The contraction phase lasts 2×10^5 yr for a protoplanet of 1.5 Jupiter masses (M_J) and 4×10^6 yr for a 0.3 M_J protoplanet (Bodenheimer *et al.*, 1980). These times depend on the assumed grain opacities; interstellar grains were used in this particular calculation. Once the central temperature heats to 2000 K,

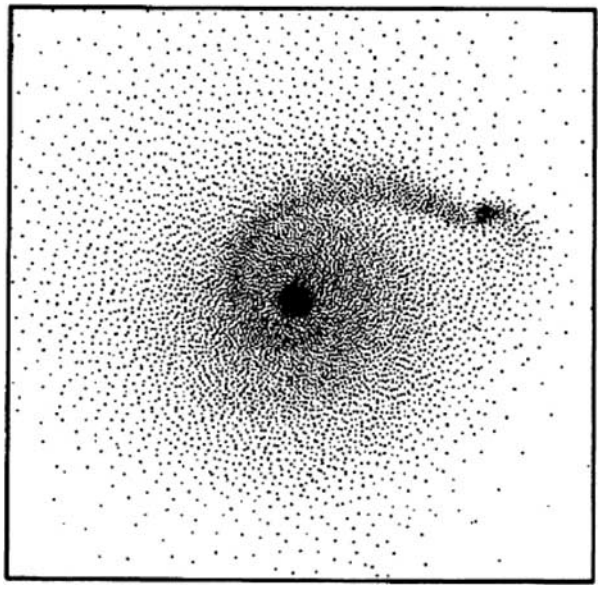

Fig. 1.1. Two-dimensional SPH simulation in the disk plane of gravitational instability in an isothermal disk with mass equal to that in the central star. The particle positions are shown after slightly more than one disk rotation at its outer edge, which lies at about 100 AU from the star. Reprinted by permission from Adams and Benz (1992). ©Astronomical Society of the Pacific.

molecular dissociation sets in, leading to hydrodynamic collapse on a timescale of less than a year, with equilibrium regained at a radius only a few times larger than those of Jupiter and Saturn. DeCampli and Cameron (1979) were the first to make an estimate as to whether a solid core (deduced to be present in both Jupiter and Saturn now) could form during the early quasi-static equilibrium phase, finding that a 1 Earth mass ($M_\oplus$) core was possible only if the protoplanet mass was less than 1 M_J. This calculation of settling of solid material toward the center was followed up by Boss (1998b), who argued that a giant gaseous protoplanet of 1 M_J could indeed form a core of a few $M_\oplus$, in line with present estimates of the core mass of Jupiter.

The first actual numerical simulation of gravitational instability in a gaseous disk in a situation relevant for planet formation was apparently done by Adams and Benz (1992), following a linear stability analysis by Shu *et al.* (1990). The calculation was done with a two-dimensional SPH code, the disk mass was equal to the mass of the central star, and the disk was assumed to be isothermal. The density was perturbed with an amplitude of 1% and an azimuthal wavenumber $m = 1$. The result was a one-armed spiral with a gravitationally bound knot (Fig. 1.1) of mass 1% that of the disk on an elliptical orbit; it was not determined whether the knot would survive for many orbits and evolve into a giant planet.

1.5 Core accretion: gas capture

Although the concept of a "planetesimal" had been discussed for a long time beforehand (Chamberlin, 1903), Safronov (1969) was the first to give a fundamental and useful theory for the accretion of solid objects. He states:

"... despite the complexity of the accumulation process and the fact that fragmentation among colliding bodies was important, the process of growth of the largest bodies (the planetary 'embryos') can be described quantitatively in an entirely satisfactory manner if we assume that their growth resulted from the settling on them of significantly smaller bodies and that they were not fragmented during these collisions."

Thus his fundamental equation for the accretion rate of planetesimals onto a protoplanetary "embryo" was relatively simple. In its modern form,

$$\frac{\mathrm{d}M_{\mathrm{solid}}}{\mathrm{d}t} = \pi R_c^2 \sigma \Omega \left[1 + \left(\frac{v_{\mathrm{e}}}{v}\right)^2\right], \tag{1.2}$$

where πR_c^2 is the geometrical capture cross-section, Ω is the orbital frequency, σ is the solid surface density in the disk, v_{e} is the escape velocity from the embryo, and v is the relative velocity of embryo and accreting planetesimal. The expression in brackets is known as F_{g}, the gravitational enhancement factor over the geometrical cross-section. An important requirement for a reasonable accretion timescale is that F_{g} be large. However Safronov typically takes it in the range 7 to 11.

Safronov actually wrote the above equation as

$$\frac{\mathrm{d}M_{\mathrm{solid}}}{\mathrm{d}t} = \frac{4\pi(1+2\theta)}{P}\sigma_0\left(1 - \frac{m}{Q}\right)R_{\mathrm{c}}^2, \tag{1.3}$$

where $\theta = (Gm)/(v^2 R_{\mathrm{c}})$ is known as the Safronov number, m is the embryo mass, Q is the present mass of the planet, σ_0 is the total initial solid surface density in the disk, and P is the orbital period. In connection with the m/Q factor he states:

"In the derivation of [the] formula ... for growth [times for terrestrial planets] it was assumed that the planetary zone was closed, or more precisely, that the total amount of solid material in the zone was conserved at all times and that its initial mass was equal to the present mass of the planet."

Thus effectively he has introduced the idea of the "minimum mass solar nebula" by requiring that the solid-surface density in the disk be just sufficient to correspond to the solid mass of the final planet.

Applying this assumption, he uses the equation to derive growth times. *"Within 100 million years the Earth's mass must have grown to 98% of its present value,"* consistent with modern estimates of the growth time of the Earth. Detailed numerical calculations (Wetherill, 1980) of the formation of the terrestrial planets starting from roughly 100 lower-mass objects with low eccentricities spread out over the

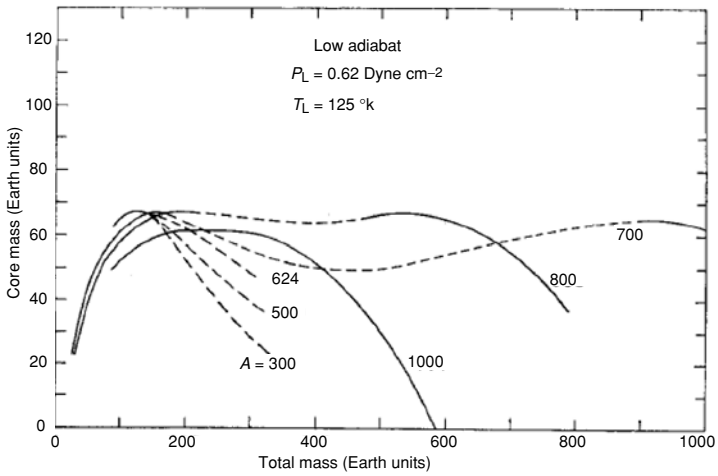

Fig. 1.2. The core mass as a function of total mass for a protoplanet consisting of a solid core plus a gaseous adiabatic envelope. Solid lines refer to hydrodynamically stable envelopes and dashed lines refer to unstable ones. A is a parameter determined only by the distance of the planet from the central star, while P_L and T_L refer to the pressure and temperature, respectively, at the outer edge of the planet. Reprinted by permission from Perri and Cameron (1974). ©Academic Press.

terrestrial-planet zone gave this timescale, along with approximately the correct number of objects.

However Safronov also noted: *"It would appear . . . that the distant planets (Uranus, Neptune, and Pluto) could not have managed to develop and use up all the matter within their zones within the lifetime of the solar system,"* a problem that is still not satisfactorily solved. In fact this statement actually led Cameron to pursue the gravitational instability hypothesis for the outer planets.

In connection with giant planet formation, Safronov mentions only briefly the process of gas capture: *"Effective accretion of gas by Jupiter and Saturn set in after they had attained a mass of about one to two Earth masses."* Cameron (1973) followed up on this remark with a statement to the effect that Jupiter could form by gas accretion once a solid core of about 10 $M_\oplus$ had been accumulated. He and Perri (1974) then made the first detailed calculation of a protoplanet consisting of a solid core and a gaseous envelope.

The Perri–Cameron model was assumed to be in strict hydrostatic equilibrium, to have a solid core of a given mass, and a gaseous envelope, assumed to be adiabatic, of Solar composition extending out to the Hill radius. The idea was to find structures that were dynamically unstable, implying that the gaseous envelope would rapidly collapse onto the core and that gas accretion would continue on a short timescale. They found that the structure was stable for values of the core mass up to a critical mass, above which it was unstable. Figure 1.2 shows the results for a particular

choice of adiabat: the first maximum in the curve for a given value of the parameter A corresponds to the core mass where the configuration becomes unstable. The parameter A depends only on the distance from the Sun; relevant values of A for the formation of Jupiter and Saturn are in the range 300 to 500. In this range the figure shows that the critical core mass is about 70 $M_\oplus$, too high as compared with current core mass determinations for Jupiter and Saturn. With a different reasonable choice for the adiabat, the critical core mass turns out to be even higher, about 115 $M_\oplus$. Nevertheless, the result does not depend sensitively on the distance, in approximate agreement with the properties of the giant planets.

This type of model was improved considerably by Mizuno (1980). He again constructed models in strict hydrostatic equilibrium, with a solid core and a gaseous envelope, extending outward to the Hill sphere. The boundary conditions for density and temperature at the outer edge were provided by a disk model. The energy equation was solved, given an energy source provided by planetesimals accreting through the envelope and landing on the core, at a fixed rate, for example 10^{-6} $M_\oplus$ per yr. The model was assumed to be in thermal equilibrium in the sense that the luminosity radiated was just equal to the rate at which the accreting planetesimals liberated gravitational energy. Also, energy transport by radiation and convection in the envelope was taken into account, with opacity assumed to be provided by the gas as well as grains. There were two main results from this calculation:

(1) He found a critical core mass, above which no model in strict hydrostatic equilibrium was possible. For interstellar grain opacity the value turned out to be 12 $M_\oplus$ (Fig. 1.3). It was assumed at the time that rapid gas accretion would follow, but later calculations (Pollack *et al.*, 1996) showed that in fact the gas accretion could be quite slow, particularly for core masses less than 10 $M_\oplus$. The models simply switched from strict hydrostatic equilibrium to quasi-hydrostatic equilibrium in which gravitational contraction of the envelope is of some importance.
(2) The value of the critical core mass turned out to be almost independent of the distance of the planet from the Sun. The fact that the value for the critical mass, as well as the independence on distance, agreed quite well with the estimates of the core masses of the giant planets at the time, put the core accretion–gas capture model on a fairly solid foundation.

1.6 Planet searches

The theories of planet formation that have just been discussed were formulated on the basis of the observed properties of the giant planets in the Solar System. Nevertheless, over the same time period that these theories were being developed, the search was on for extrasolar planets. The method used was primarily astrometry, in which an attempt is made to measure the periodic shift of the position of a star

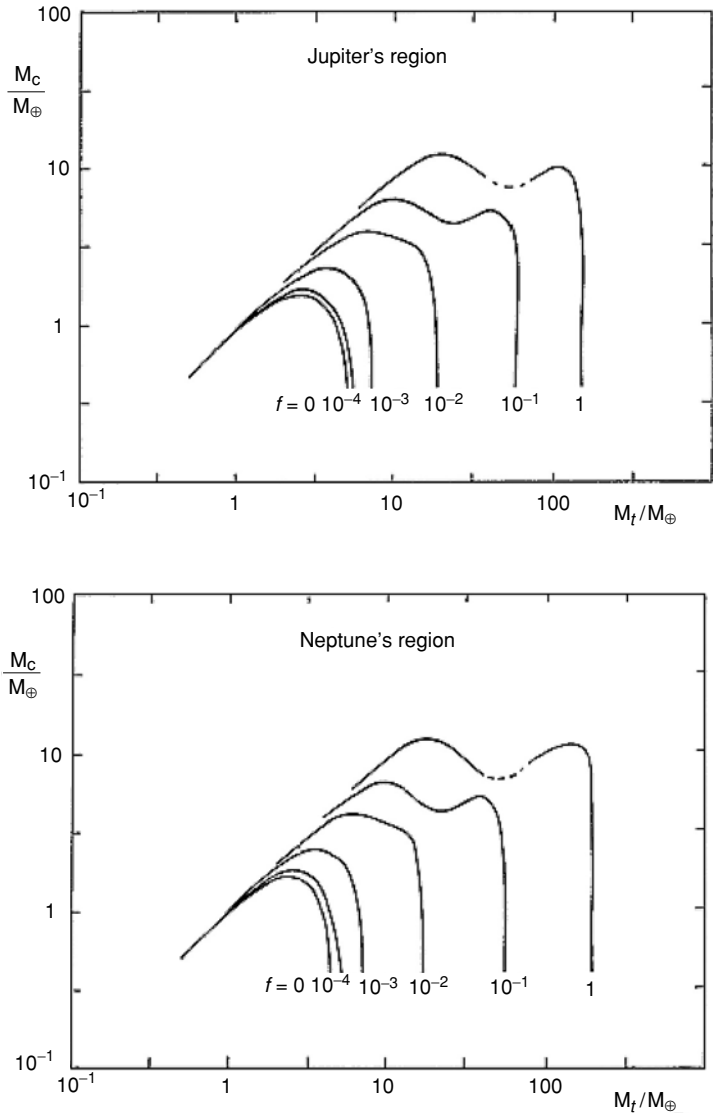

Fig. 1.3. The core mass as a function of total mass for a protoplanet consisting of a solid core plus a gaseous envelope calculated according to the equations of stellar structure. The masses are measured in units of $M_\oplus$, the Earth's mass. The parameter f corresponds to the ratio of the assumed grain opacities to the interstellar values. Solid lines refer to hydrodynamically stable envelopes and dashed lines refer to unstable ones. *Top*: giant planet formation at 5 AU. *Bottom*: giant planet formation at 30 AU. The first maximum on each curve corresponds to the critical core mass. Reprinted by permission from Mizuno (1980). ©Progress of Theoretical Physics.

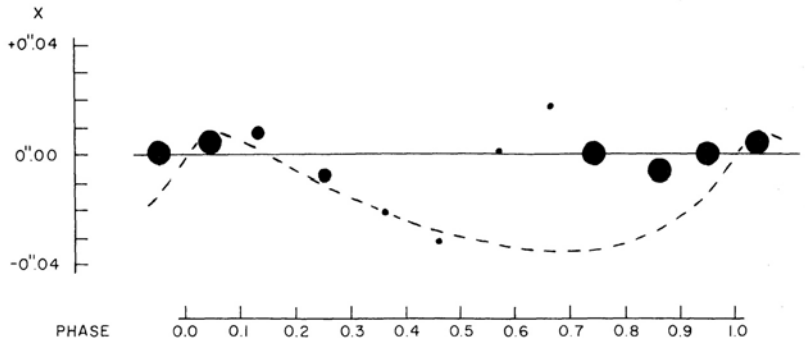

Fig. 1.4. The displacement of Barnard's star on the sky in the x-coordinate as a function of orbital phase, corresponding to a period of 24 years. The dashed line is a fit to the observations of van de Kamp (1963), while the filled circles are the measurements of Gatewood and Eichhorn (1973). The largest points have the highest degree of confidence. The solid line corresponds to zero displacement. Note that the deviation in x that is looked for is only a few hundredths of a second of arc, while the image of the star on a photographic plate is more like 1–2 seconds across. Reprinted by permission from Gatewood and Eichhorn (1973). ©American Astronomical Society.

on the sky, caused by the gravitational effects of an orbiting planet. However as reported by Black (1991): *"The history of searches for other planetary systems is littered with published detections that vanish under further scrutiny."*

The most famous example is Barnard's star, named after the astronomer who discovered its large proper motion over the period 1894–1916. Peter van de Kamp measured, as accurately as he could, the position of the star on the sky, trying to deduce the presence of a low-mass companion by observing the periodic motion of the star around the center of mass of the system. His initial observational program ran from 1916 to 1962 and produced more than 2400 images. However, the apparent size of the star's orbit turned out to be only about 1/100 the size of the stellar images on a photographic plate. Nevertheless, he claimed (van de Kamp, 1963) to have found a planet, and the result was announced in the News and Views section of *Nature*. The companion was supposed to have a mass of 1.6 M_J, orbiting a star of 0.15 $M_\odot$ with a period of 24 yr, an eccentricity of 0.6, and a semimajor axis of 4.4 AU. However the system was measured independently by Gatewood and Eichhorn (1973) with a series of 241 plates, and they did not confirm the presence of the planet (Fig. 1.4). Adjustments to the telescope configuration at the Sproul Observatory were thought to have introduced spurious displacements of the star.

Nevertheless van de Kamp (1975) reanalyzed his data, using only that which was taken after the major adjustments had taken place, and now claimed the presence of two planets, with masses of 1 M_J at a distance of 2.7 AU (11.5 yr period), and