Theme 9: Galaxies and Cosmology

Through most of the history of astronomy, the question of cosmology—the origin and history of the universe—was comparatively neglected, owing to a lack of useful evidence. Although speculation about cosmology is as old as humankind, the scientific study of cosmology dates back only a century: although there were some earlier musings, for example concerning the (lack of) stability of a static Newtonian universe, and the puzzle of the dark night sky in an infinite cosmos, they were not carried through into well-considered models. It is reasonable to argue that scientific cosmology originated with Einstein's theory of General Relativity in 1915, or perhaps from the first cosmological solutions of the field equations (Einstein's closed, static universe and de Sitter's empty universe) in 1917.

9.1 Early cosmological concepts

The basic concepts of cosmology can be expressed in terms of the universe's extent in time and space:

- Is the universe finite or infinite in time?
    - Is it infinitely old, or did it have a definite beginning—if the latter, how old is it?
    - Does it have an infinite future, or will it end—and if so, when?
- Is it static and unchanging, does it evolve unidirectionally, or is it cyclic?
- Is the universe finite or infinite in space?
    - And, if finite, does it have an edge?

The world's religions all offer answers to these questions, but are by no means unanimous: the Judaeo-Christian outlook favours a definite beginning in the not-too-distant past, and a catastrophic end in the unspecified future, whereas the Hindu view inclines towards an infinitely repeated cycle of creation and destruction (cycles, not necessarily infinite, also seem to be common in what little I know of Native American cosmologies). Early physicists such as Newton seem to have assumed for the purposes of calculation that the universe is static and eternal (this does not necessarily mean that they did not believe the Biblical timescales, since God could presumably have created the universe in approximately its present state).

Up to the 17th century, the spatial extent of the universe was assumed to be quite small: the stars were lights in the sky, and were located just sufficiently beyond the orbit of Saturn so as not to interfere with Saturn's motion. This is one reason why Aristarchos' heliocentric cosmos failed to attract support: the Greeks would have expected to see annual parallax, and it seemed extremely unnatural to place the stars far enough away to make that parallax negligible (Aristarchus was quite aware of the problem, and had—as we know from a quotation in Archimedes—the correct solution; it's just that the correct solution required a leap of faith about distances that Aristarchos' contemporaries were unwilling to take). In contrast, Kepler's solution to the planetary orbits, which required a heliocentric solar system, was too elegant to be dismissed—his tables of planetary positions *worked*—and was subsequently underpinned by Newton's universal laws of motion and gravitation; also, Galileo's telescopic resolution of regions of the Milky Way into small, faint stars strongly suggested that these stars were more distant than the naked-eye stars known since antiquity. Therefore, from the late 17th century onwards, there was general recognition that the stars were not attached to a spherical shell somewhere not too far beyond Saturn, but were scattered through a large three-dimensional space. Newton attempted to estimate the distance of Sirius by comparing its brightness with that of Saturn, and got an answer of the correct order of magnitude (10^6 AU, about a factor of 2 too large); other stars were fainter than Sirius, and therefore presumably more distant.

In Newtonian gravity, a finite, static universe is unstable: it will collapse rapidly towards its centre of gravity. Newton believed, wrongly, that this problem would be avoided if the universe were infinite, on the grounds that it would not then have a well-defined centre towards which to collapse. (This is true, but the equilibrium is unstable: if any region has even a slightly higher density, the surrounding regions will collapse towards it, increasing the local density and thus encouraging further collapse. Since the universe clearly contains regions of higher density, e.g. stars and stellar clusters, Newton's solution does not work.) Newtonian physicists therefore tended to assume an eternal, static, infinite universe.

This model is clearly disproved by the observational evidence of the dark night sky. In an infinite, static, unchanging universe, every line of sight should eventually intersect a star, and the night sky should have the brightness and temperature of an average stellar surface—about 3000 K. Rather obviously, it does not. This is known as **Olbers' Paradox**, after Wilhelm Olbers who described it in 1823, but is much older than this: it was known to Kepler in 1610. Indeed, Kepler offered the correct solution: that the (visible) universe is not infinite. (Olbers' solution, based on the presence of absorbing material between the stars, does not work: in an eternal universe, the absorbing material will heat up until it is itself at the same average temperature.) Since Newton is correct in asserting that a finite, static universe is unstable, the inevitable conclusion is that the universe is not static: it must be expanding or contracting (or, conceivably, rotating). Given the quality of the scientific talent on display in the 17th century, it is fairly astonishing that this did not occur to someone at the time: it would appear that the paradigm of a static universe was just too intuitively "right" to abandon.

After Römer's demonstration of the finite speed of light, a second class of explanations became viable: the universe may be infinite in space, but if it is not infinite in time then the light from more distant sources will not have reached us yet, and the sky can remain dark. Edgar Allan Poe, of all people, spotted this explanation in 1848, in his essay *Eureka*. However, by this point, creation of the entire infinite (or at least extremely large) universe in one fell swoop at some definite past time did not seem in accordance with physical law (the geologists' doctrine of uniformitarianism—that events in the past should be explained in terms of processes that we can see happening now). Therefore, it appears that from Newton's day to the close of the 19th century cosmology was more or less neglected (it is significant that the index of Agnes Clerke's *History of Astronomy during the 19th Century* contains no entries for "age", "cosmology", "origin" or "Olbers' paradox"—it does contain Olbers, but only with reference to his extensive work on comets and asteroids). One is reminded of Debye's view of the problem of beta decay, as quoted by Pauli in his famous letter about the neutrino: "it's better not to think of this at all, like new taxes."

9.2 The impact of General Relativity: Einstein, Friedman and Lemaître

The equations of General Relativity describe the behaviour of spacetime in the presence of matter: they are essentially a theory of gravity. Newtonian gravity was incompatible with Special Relativity because it entailed instantaneous action at a distance (i.e. the information carried by the gravitational force travelled faster than light). Einstein's key insight was that observers freely falling in a gravitational field accelerate, but do not experience any gravity (cf.

"weightless" astronauts in Earth orbit). Gravity behaves very like the "fictitious forces" introduced in classical mechanics when working in an accelerated frame of reference (e.g. centrifugal and Coriolis force in a rotating reference frame).

General Relativity interprets gravity as a consequence of the curvature of spacetime. This curvature affects the motion of objects; the mass of the objects in turn defines the curvature. As with Newtonian gravity, the equations are fairly difficult to solve except in special cases: the first proposed, in 1916, was Karl Schwarzschild's solution for a point mass, which describes a non-rotating black hole.

Solutions of Einstein's equations intended to apply to the whole universe generally assume that the universe is **homogeneous** and **isotropic** on sufficiently large scales. This was not, in fact, very consistent with observation in 1915, since the prevailing view in the early 19th century was that the universe of stars consisted only of the Milky Way Galaxy, other nebulae being small objects within it, and the Milky Way is clearly neither homogeneous nor isotropic[1]. However, it had the immense advantage of producing soluble equations.

Like Newtonian gravity, the "obvious" formulation of Einstein's equations does not produce a static solution: it wants either to expand or to contract. Einstein, however, was convinced of the conventional wisdom of a static universe, and hence modified the equations by including the infamous cosmological constant, Λ. By suitably tuning Λ it is possible to construct a static model, although (as with Newton's infinite homogeneous model) it is unstable to small perturbations. Einstein published his static model in 1917. It is positively curved, which means that (like the surface of a sphere in two dimensions) it is finite but unbounded: if you move in a straight line, you will eventually return to your starting point. Also in 1917, **Willem de Sitter** (1872–1934) produced an apparently static solution to the Einstein equations, which however describes a universe with (rather unrealistically) zero matter content. (Indeed, if you introduce matter into a de Sitter space, it will be seen to expand.)

In the period 1922-24 the Russian mathematician **Alexander Friedman** (1888–1925) published a set of papers which described the three $\Lambda = 0$ solutions to Einstein's equations for a homogeneous, isotropic universe. These all predict an expanding universe (though a universe with positive curvature will later recollapse). Einstein was extremely critical of these solutions—initially he asserted, wrongly, that Friedman had made a mathematical error in deriving them—and this, coupled with Friedman's early death and status as a mathematician and meteorologist rather than an astronomer, meant that they received little attention in the astronomical community.

The expanding-universe solutions were rediscovered in 1927 by Abbé Georges Lemaître (1894–1966). Lemaitre was an astronomer rather than a mathematician, and was aware of the increasing evidence for expansion of the system of nebulae (the title of his 1927 paper refers explicitly to "the radial velocity of extragalactic nebulae"). Unlike Friedman, Lemaître made explicit reference to an initial extremely dense state for the universe (which he called the primeval atom): therefore, he can be regarded as the originator of the modern Big Bang model. Einstein was equally critical of Lemaître's work, but this time the observational evidence intervened conclusively in Lemaître's favour.

[1] Such a universe, with a finite distribution of stars embedded in an infinite universe, and presumably stabilised against gravitational collapse by orbital motion, does not suffer from Olbers' Paradox. However, as Einstein pointed out in his 1916 popular book *Relativity: the Special and General Theory*, it is not sustainable for infinite time, because the Milky Way is continually radiating away energy in the form of starlight.

9.3 Slipher, Hubble and an expanding universe

At the time when Einstein developed the theory of General Relativity, the status of the so-called spiral nebulae was not at all clear. There were two main schools of thought:

- ***Small Galaxy, extragalactic nebulae***
  Around 1800, William Herschel attempted to determine the shape of the Galaxy using "star gauging" (nowadays simply called "star counts". He concluded that it was a somewhat flattened system with the Sun near the centre. A century later, **Jacobus Kapteyn** (1851–1922) used essentially the same technique, and obtained an essentially similar result. Believers in the Kapteyn model of the Galaxy tended to accept that the nebulae were extragalactic systems probably similar to the Galaxy.
- ***Big Galaxy, intragalactic nebulae***
  In 1918, **Harlow Shapley** (1885–1972) used "Cepheid variables" (actually, what we would now call W Virginis variables and RR Lyrae stars) to determine the distances of globular clusters, and hence discovered that the centroid of their distribution was located far from the Sun (Shapley estimated "about 20,000 parsecs") in the direction of Sagittarius. Shapley therefore envisaged a large, fairly strongly flattened system with the Sun well off-centre—probably about halfway between the centre and the edge. This system was so large that Shapley was confident that nebulae were small systems located within the Galaxy.

In 1920, **Heber Curtis** (1872–1942) and Shapley engaged in the "Great Debate" about the status of the Galaxy and the nebulae, with Curtis defending the "small Galaxy, extragalactic nebulae" position. The Debate was not as conclusive as some later commentators would have one believe, largely because Shapley, who was in the running for the Directorship of Harvard College Observatory, did not wish to offend anyone influential and therefore gave a fairly low-level and uncontroversial talk. Much of the mythology of the Great Debate actually stems from papers published by the two speakers in 1921, which in fact contain much material not presented in the talks.

Meanwhile, **Vesto Slipher** (1875–1969) at the Lowell Observatory in Flagstaff, Arizona, was measuring the Doppler shift of spiral nebulae—a considerable technical challenge because the absorption lines in their spectra were much more difficult to measure than the bright emission lines of the gaseous nebulae. Slipher found that the spiral nebulae were mostly redshifted, and that the redshifts were very large compared to those of typical stars—several hundred km/s as opposed to a few tens of km/s. By 1917, Slipher had concluded from the pattern of his redshifts that the Sun, and by extension the Milky Way Galaxy, was moving relative to the centre of mass of the system of nebulae, and therefore that the nebulae were not part of the Milky Way. Owing to a near-total lack of distance estimates, however, he failed to realise the full significance of the fact that the vast majority of his Doppler shifts were *red* shifts (i.e. receding velocities).

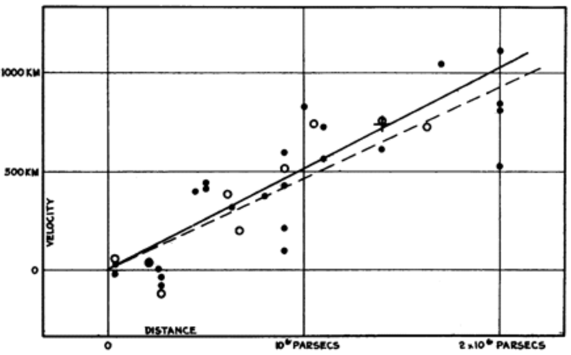
The significance of Slipher's redshifts was realised by **Knut Lundmark** (1889–1958) who in 1924 published a paper exploring the implications of the radial velocities of various classes of object in terms of a de Sitter universe (like most astronomers, he was obviously unfamiliar with Friedman's work). This paper includes a velocity-distance relation which predates Hubble's by five years (though it is nowhere near as convincing).

Lundmark dealt with the lack of reliable distances by assuming that all the spiral nebulae were comparable objects and using their relative brightnesses to infer their distances in units of the distance to M31. (He had previously measured the distance to M31 using novae; in his 1924 paper he concludes that it is "of the order of half a million parsecs", which is very much the right order of magnitude—the modern estimate is 780 kpc.) Lundmark's paper, and the title of

Lemaître's theoretical work, show that the idea of an expanding universe was by no means novel by the end of the 1920s.

For Slipher and Lundmark, the missing information was the distances of the nebulae. This was a technological problem, as the necessary astronomical information had been available since 1912, when **Henrietta Swan Leavitt** (1868–1921), a member of "Pickering's harem", discovered the Cepheid period-luminosity relation. Shapley used Cepheids in his work on the structure of the Galaxy, but Cepheids beyond the Magellanic Clouds were too faint to resolve in the telescopes of the day. However, in 1917, the Hooker 100" telescope at Mt Wilson Observatory came into operation. The Hooker had more than double the light-collecting area of the next largest instrument (the 60" Hale telescope, also at Mt Wilson), and was to remain the world's largest optical telescope until the advent of the Palomar 200" in 1948.

Armed with the Hooker telescope, **Edwin Hubble** (1889–1953) and **Milton Humason** (1892–1971) were able to resolve Cepheid variables in several nearby galaxies. Combining Cepheid distances with estimates using novae and "blue stars involved in emission nebulosity", Hubble and Humason constructed a distance ladder which used these distances to calibrate the luminosity of the brightest stars in nebulae, and hence were able to estimate the distance of any nebula in which at least a few stars could be resolved. The resulting 1929 paper showed a reasonably clear correlation between redshift and distance; assuming a linear relationship, Hubble estimated a proportionality constant of about 500 km s^{-1} Mpc^{-1}. The linearity was solidified two years later when Hubble and Humason published a much more extensive data set, using Cepheids to calibrate brightest stars and brightest stars to calibrate average total brightness. The resulting plot was impressively linear, with a proportionality constant of 558 km s^{-1} Mpc^{-1}. Hubble and Humason acknowledge the possibility of systematic errors in this value, but believe that "the uncertainty in the final result is definitely less than 20 per cent and probably less than 10 per cent."



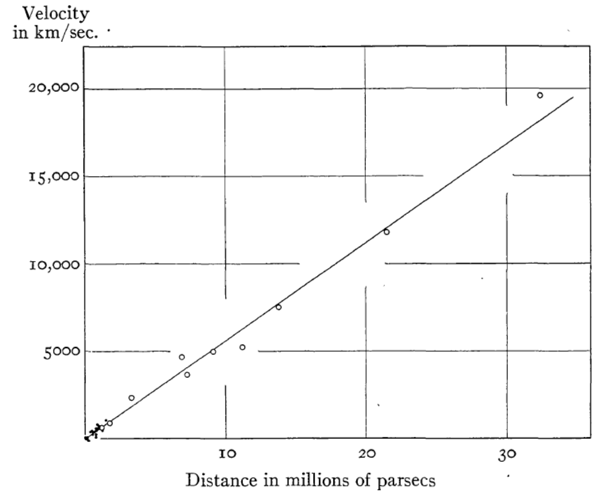


*Figure 9.1: Hubble's law by Hubble (PNAS **15** (1929) 168–173), above, and Hubble and Humason (ApJ **74** (1931) 43–80), right. Both plots give a Hubble constant of 500–550 km s^{-1} Mpc^{-1}.*

This paper established the proportionality between distance and redshift for the nebulae, now known as **Hubble's law** (which is slightly unfair to Humason). In the 1929 paper Hubble comments on the applicability of this relation to de Sitter's cosmology, and within a year or so Eddington had begun to publicise Lemaître's work (which had been languishing in an obscure Belgian journal). The expanding universe hypothesis had become solidly accepted.

Note that this *could* have been a classic example of Popperian hypothesis testing: Einstein could have believed his own equations and concluded that the universe must be expanding or contracting; Friedman or Lemaître could have published in less obscure journals and made more effort to convince the astronomical community. However, in practice the picture is much murkier. Hubble's own explanation of his motivation, in the introduction to the 1929 paper, is:

> *"Determinations of the motion of the sun with respect to the extra-galactic nebulae have involved a term of several hundred kilometers which appears to be variable. Explanations of this paradox have been sought in a correlation between apparent radial velocities and distances, but so far the results have not been convincing. The present paper is a re-examination of the question, based on only those nebular distances which are believed to be fairly reliable."*

This appears to indicate more interest in the solar motion relative to the nebulae—which had also been the focus of Slipher's 1917 paper—than in the redshift-distance correlation. It seems unlikely that the studies being carried out in the 1920s were aimed primarily at testing (unpopular) cosmological hypotheses. (It also indicates that Hubble knew what he was looking for: popular accounts which suggest that this was an astonishing and unforeseen development are written by people who haven't bothered to read the paper!)

9.4 Big Bang vs Steady State

The large value of *H* obtained by Hubble presents a problem. Interpreted in terms of an expanding universe, it implies an age for that universe of about 1.8 billion years. But radiochemical dating of the Earth's crust at around this time was already suggesting an age significantly greater than this. It is clearly not reasonable for the Earth to be older than the universe!

The obvious explanation, namely that Hubble's value for *H* was badly wrong, seems to have enjoyed remarkably little support. Jan Oort wrote a prescient paper in 1931 in which he argued that the value was (a) smaller and (b) extremely uncertain, but this does not seem to have won over the astronomical community, and in fact Hubble's value reigned supreme until 1951, by which time the age discrepancy was acute and very worrying. In the late 1940s, the time was therefore ripe for an alternative cosmological model.

In 1948, in nearly back-to-back papers (Hoyle, *MNRAS* **108** (1948) 372, and Bondi and Gold, *MNRAS* **108** (1948) 252), Hoyle on the one hand, and **Hermann Bondi** (1919–2005) and **Tommy Gold** (1920–2004) on the other, presented such an alternative theory. The **Steady State model** accepted the Hubble expansion, but posited that the overall appearance of the universe nonetheless remained constant over time, with new matter continually created to maintain constant density despite the expansion. In this model, all large-scale properties of the universe remain constant over time: the universe is infinite, eternal, and eternally expanding (note that the effective "cooling" of light from distant sources caused by the expansion redshift avoids any problem with Olbers' paradox).

This is an unusual example of an astronomical hypothesis generated not by unexpected data but by a philosophical principle. Bondi and Gold in particular were very much influenced by **Mach's principle**, which states that the background distribution of distant galaxies is a necessary reference point for physical laws. Bondi and Gold argued that, if this is the case, it would be very surprising to find the same physical laws holding in the very early universe, when the average density of matter was presumably very different from the current value. Since spectroscopy of distant objects suggests that physical laws, or at least those governing atomic physics, actually are the same, it follows (argued Bondi and Gold) that the overall conditions should also be the same. They called this idea the **Perfect Cosmological Principle**; it is an extension of the usual

Cosmological (or Copernican) Principle, which says that the large-scale properties of the universe are the same everywhere in space (i.e. we do not occupy a privileged location in space).

Hoyle's motivation, at least as stated in his 1948 paper, is rather more pragmatic. He explicitly invokes the problem of the large Hubble constant: "[for a universe with k = 0], t must be about 4×10^{16} sec., which is about 1.3×10^{9} years. ... This conclusion is in discordance with astrophysical data, which strongly suggest that physical conditions have not changed appreciably over a period of about 5×10^{9} years. In this connection it may be noted that geophysical studies give about 2×10^{9} years for the age of the Earth."

The Steady State model evidently avoids the age problem. Although the average age of galaxies in a given region of space is $\sim 1/H$, any particular galaxy may be much older than this, and there is nothing to preclude our living in such a galaxy. It also avoids the mathematically intractable singularity at t = 0 in expanding-universe models. Its mathematical and philosophical elegance appealed to many theorists.

More to the point, the Steady State model is *testable*. Even with the exaggerated expansion rate given by Hubble's numbers, the level of matter creation required by the theory is far too small to be detectable experimentally, but its large-scale predictions are clear, and are very difficult to avoid: the theory's basic assumptions put extremely tight constraints on its formulation. This contrasts with the expanding-universe models, where there are a number of unknown and poorly understood parameters (the expansion rate itself, the way in which astrophysical objects and populations evolve over time, the average density—which is also unknown in the Steady State, of course, but in the Steady State model it doesn't *matter*, whereas in Friedman-Lemaître models it certainly does). The Steady State, with its testable predictions, encouraged observational astronomers to contribute to cosmological questions.

The main testable prediction of the Steady State is that *the overall appearance of the universe should not change over time*. Given the finite speed of light, this is equivalent to a statement that *the high redshift universe should look similar to the local universe*. An additional corollary is that all observed phenomena should be explicable by processes operating in the local universe—a sort of cosmic uniformitarianism. These predictions were, of course, not in conflict with any data available in the late 1940s (we have seen that at this time the radio astronomers interpreted their point sources as local phenomena within the Milky Way, and therefore of no cosmological significance).

The first significant change in the observational situation occurred almost immediately. Observing at Mt Wilson during the war, **Walter Baade** (1893–1960) had taken advantage of the dark skies produced by a blacked-out Los Angeles to resolve the central regions of M31, and therefore recognise the distinction between the red Population II of globular clusters and galactic bulges and the blue Population I of the solar neighbourhood and galactic discs. It was then realised that the "Cepheid variables" of the two populations were not identical, but differed in brightness by a factor of 4, with "Type I Cepheids" (now called classical Cepheids) being brighter than "Type II Cepheids" (W Virginis stars). Hubble had observed classical Cepheids in the disc of M31 and other galaxies, but calibrated them using W Vir stars in globular clusters: he had therefore[2] underestimated their distances by a factor of 2. At a stroke, the value of the Hubble constant was reduced from 500 to 250 km s^{-1} Mpc^{-1}, and much of the age discrepancy

[2] Actually, this is a bit of an oversimplification, since Shapley had used *Galactic* Cepheids to calibrate his globular cluster "Cepheids": the error should therefore have cancelled out. But the Galactic Cepheid calibration was wrong, for a variety of reasons: Galactic Cepheids were too far away for parallax, so their distances were poorly measured; their proper motions, needed for the distance estimate, were also poorly known; and interstellar absorption had been neglected.

between the Earth and the universe was resolved. This is not, of course, evidence against the Steady State, but it removes a major piece of evidence against the alternative theory.

The advent of radio source counts in the 1950s seemed to weigh against the Steady State model. As discussed earlier, the requirement that the space density of radio sources be constant throughout space and time (over a sufficiently large scale) is not consistent with an observed excess of faint sources, *if* it can be demonstrated that these faint sources are faint because of their distance. At the time, there were two arguments that could be used against the source count data:

- *Source confusion*
  A radio survey with poor resolution will overcount sources just above its threshold of sensitivity. This is because two sources with individual fluxes below the threshold can both lie in the same "pixel" of the survey, producing a phantom source above threshold. This was definitely an acute problem in the 2C catalogue: when the Sydney group produced a catalogue of southern sources, there was almost no correlation in the region of overlap between Sydney's faint sources and Cambridge's. The 3C catalogue was much better in this respect.
- *Two source populations*
  The argument advanced by Dennis Sciama was that an excess of faint sources can readily be explained if we assume that there are two distinct classes of radio source: luminous extragalactic sources and faint objects within our Galaxy. Other galaxies' populations of faint sources cannot be seen, because they are too distant, and it only takes a small fluctuation in the distribution of the local sources for them not to contribute at the highest fluxes. Therefore one observes a population of faint sources superimposed on the smooth −3/2 power law from the luminous sources. This argument is defensible as long as the proportion of *identified* faint sources remains small: as more faint sources are identified with extragalactic optical counterparts, the argument becomes progressively more difficult to sustain.

In 1964, the two-populations argument was still valid, although progress in source identifications would soon have posed a problem. The discovery of quasars in 1963 had added a further strike against the Steady State: here was a population of objects at high redshift that appeared to have no local analogue at all, in clear contradiction to the Steady State assumption that all epochs should be equivalent. The counter-arguments put forward by Steady State supporters required that the quasar redshifts be non-cosmological: although there are problems with this interpretation (the redshifts involved seem to be too large for explanation as gravitational effects and the lack of observed proper motions presents a problem for interpretation as Doppler shifts caused by local motion), these were not considered insurmountable in the mid-1960s, when the observational data on quasars were still quite limited. However, the clearest problem for the Steady State came in 1965 with the discovery of the **cosmic microwave background** (CMB).

The CMB is an unambiguous prediction of a Hot Big Bang model[3]. In the hot, dense, ionised stage, one naturally expects radiation to be in thermal equilibrium with matter. Once the universe cools enough for neutral hydrogen atoms to form, it becomes transparent, and the radiation decouples from the matter, maintaining its thermal spectrum but cooling as the universe expands. The temperature at which neutral hydrogen forms is about 3000 K; estimating how

[3] By this time, the Big Bang model had acquired its name. It was christened by Fred Hoyle in the 1950s: he was presenting a radio show on cosmology and wanted a snappy name to contrast with Steady State. It is always presented as pejorative, but Ken Croswell (*The Alchemy of the Heavens*) quotes Hoyle as saying that it was not meant so.

much the universe has expanded since that time, one concludes that there should be a background of thermal radiation with a temperature of a few kelvin to a few tens of kelvin. This prediction was indeed made around 1950 by **Ralph Alpher** (1921–2007) and **Robert Herman** (1914–1997), members of George Gamow's research group.

The most famous paper of the Gamow group is **αβγ** (Alpher, Bethe and Gamow)[4], published in 1948 (the same year as the Steady State papers). This paper, the standard reference for the Hot Big Bang, is actually an almost entirely incorrect attempt to explain the abundances of the elements by starting from an extremely hot neutron gas. It doesn't work: the absence of any stable nuclide with mass 5 (^{4}He + (p or n)) or mass 8 (^{4}He + ^{4}He) prevents the build-up of heavier nuclei. However, the group wrote several other papers working out consequences of the model, and one of them includes this prediction of a thermal background at about 5K.

Unfortunately, a blackbody distribution at 5K is in the microwave and submillimetre region of the electromagnetic spectrum, which was not readily detectable in 1950. The prediction appeared untestable, and was promptly forgotten. In the mid-1960s, it was independently rediscovered by **Robert Dicke** (1916–1997) and his group at Princeton: they estimated the temperature as not more than 50 K, decided it should be observable in the microwave region, and designed a microwave receiver to look for it. Before this could be completed, they were scooped by **Penzias** and **Wilson** of Bell Labs, who famously found the CMB accidentally while trying to track down stray noise in their microwave horn antenna. If interpreted as thermal radiation, their signal had a temperature of about 3 K.

One point does not establish a blackbody distribution, and the Steady State theory could cope with a background distribution at an effective temperature of 3 K just fine: it can be generated by scattered background starlight. (Other diffuse backgrounds *are* generated in this way.) However, further points, mostly by **Wilkinson** (1935–2002) of the Princeton group, who had continued with his microwave observations despite having been robbed of his Nobel prize, all seemed highly consistent with a blackbody spectrum.

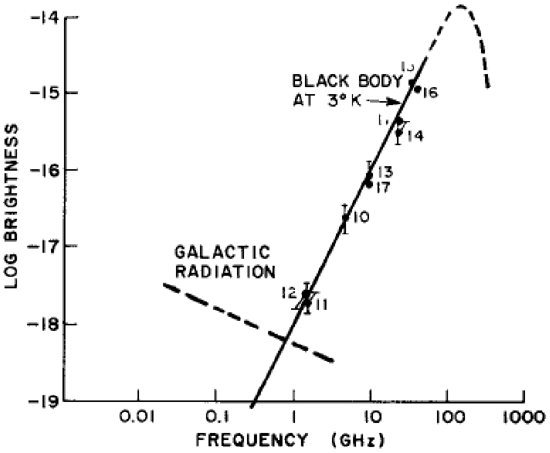


Figure 9.2: compilation of cosmic background measurements by Arno Penzias, IEEE Trans. Mic. Th. Tech. ***16*** *(1968) 608–611. Data from Penzias and Wilson (10, 12), Howell and Shakeshaft (11), Wilkinson and colleagues (13, 16, 17), Welch et al. (14), Ewing, Burke and Staelin (15). The measurements are highly consistent with a blackbody spectrum at ~3 K (in fact, the Princeton group were already claiming 2.7 K). Measurements around the peak were hindered by the fact that the atmospheric radio window closes at ~40 GHz: space-based instrumentation is needed to cover this region effectively.*

A pure blackbody spectrum is a natural and inevitable consequence of a Hot Big Bang: you have a condition of thermal equilibrium existing everywhere in the universe at one defined time in the past (the so-called "recombination era", when free protons and electrons first combined to make neutral hydrogen). This generates a blackbody spectrum which simply redshifts to lower temperatures as the universe expands. In contrast, there is no natural way to make a pure

[4] The pun is Gamow's, and is deliberate. The paper was actually written by Alpher and Gamow: Gamow (1904–1968) got Bethe to check the calculations and then added his name to the author list (without his knowledge, apparently) to produce the desired effect.

blackbody background in the Steady State model. If the background radiation is scattered starlight, it should be coming from many different galaxies with different redshifts—and a superposition of blackbody spectra with different temperatures is not itself a blackbody spectrum. In order to get a pure blackbody, you have to ensure that the blackbody spectrum is generated locally (at zero redshift) and not contaminated with more distant sources. While solutions of this type were put forward by Steady State loyalists, they were generally regarded as extremely contrived. Meanwhile, the identification of increasing numbers of radio sources, particularly quasars, as objects at high redshift (and the lack of corresponding objects at low redshift or blueshift) was increasing the pressure from the radio source counts. The combination was too much for the Steady State to withstand: with the exception of a few die-hards (including, sadly, Fred Hoyle, whose later life was regrettably marred by a succession of bad decisions; even the truly brilliant sometimes let emotional attachment get in the way of clear thinking), cosmologists abandoned the Steady State model en masse in the late 1960s. The Big Bang model was further strengthened by calculations of the yield of light nuclides by Wagoner, Fowler and Hoyle in 1967.

The choice of Big Bang vs Steady State is a good example of evidence-based decision making. Both theories made testable predictions (even if one prediction did get forgotten for 15 years!); the predictions were tested; the theory whose predictions were confirmed became established as the agreed paradigm. It is worth noting that the Steady State theory was actually more influential than the ultimately successful Big Bang in initiating the necessary observational work, because it made such clear and unambiguous predictions, and that some aspects of the "successful" Big Bang theory were unquestionably incorrect as initially proposed (the αβγ theory of nucleosynthesis is almost entirely wrong, as indeed is the starting point for that theory, i.e. that the initial hot dense plasma is a neutron gas). It is also true to say that the current "Standard Cosmological Model", though directly descended from the Friedman, Lemaître, Gamow et al. Hot Big Bang, would be unrecognisable to its progenitors: dark matter, inflation, and most recently dark energy have all been added to the original concept. Nevertheless, the reason that the Big Bang model remains the favoured theory, even in this much modified form, is the observational evidence: the CMB and the yield of light nuclides really are almost impossible hurdles for rival theories. In the modern era, the detailed analysis of the CMB anisotropies has been added to the list of observational evidence: again, the need to provide as good a description of the CMB power spectrum as that produced by the Standard Cosmological Model is an extremely stringent test of candidate alternative models, and one which no candidate has yet passed.

9.5 Hubble Wars

Hubble's original conviction that his number of 500−550 km s^{-1} Mpc^{-1} had an error of no more than 20% was undermined first by Baade's discovery of the two types of Cepheids, which reduced it by a factor of 2, and then by Allan Sandage's realisation that many of Hubble's "brightest stars" were not stars at all, but much brighter H II regions. By 1958, Sandage (*ApJ* **127** (1958) 513) had brought the value of H down to 75 km s^{-1} Mpc^{-1}, "with a possible uncertainty of a factor of 2." Ironically, this is within 5% of the currently accepted best value—but the history of H_0 between then and now is not a simple shrinkage of the error bars, but a wild oscillation, with Sandage one of the principal drivers.

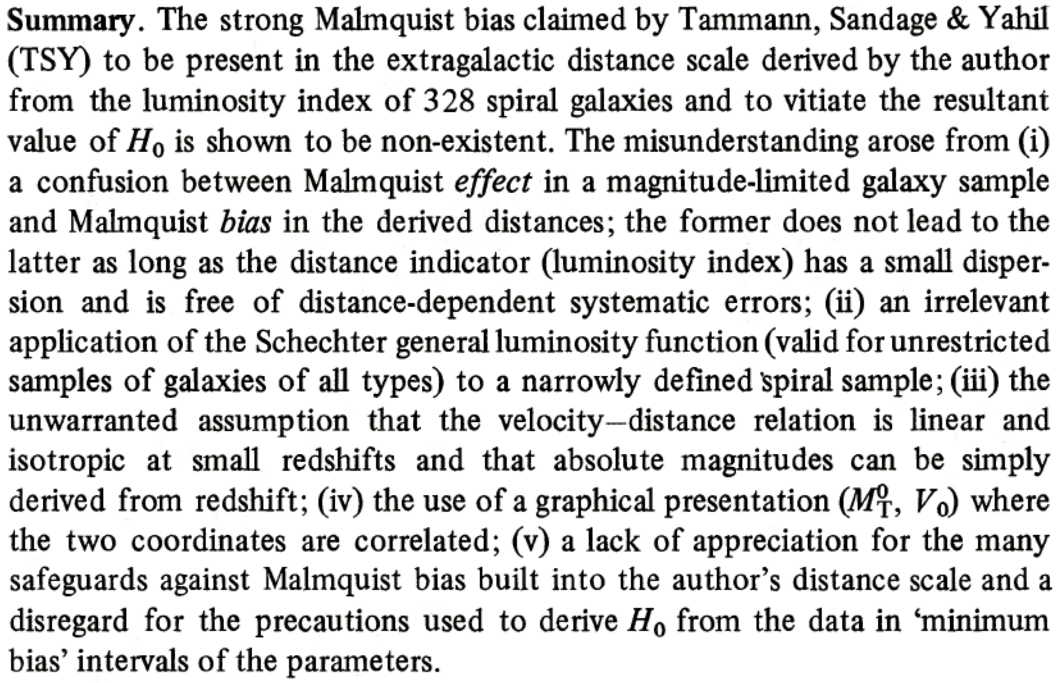
The problem with determining H_0 is not the redshift. With the exception of occasional pathological cases like the first few quasars, where the line identities are not obvious, redshift determination is extremely straightforward. The problem is the distance.

Distance determinations in astronomy are most commonly done by comparing apparent and absolute magnitudes: $m - M = 5 \log(d/10)$, where d is measured in parsecs. To apply this, you need a "standard candle"—an object whose absolute magnitude M you believe you know. The trouble is that if you get M wrong, it changes your distance estimate by a constant *factor* of $10^{\Delta M/5}$, where ΔM is the error in M. Since the equation you are trying to fit is $v = Hd$, multiplying all your values of d by $10^{\Delta M/5}$ does not destroy the straight line: it simply changes the apparent value of H. Hubble and Humason had an exemplary straight line, but their slope was out by a factor of about 8.

The errors which can enter into distance determinations are numerous. In cases where the absolute magnitude of the test object is related to some other observable quantity, e.g. the Cepheid period-luminosity law, it is often straightforward to determine the slope of the relation: the problem is the intercept, or zero-point. Ideally, one wants to be able to determine this from some absolute distance measure such as parallax, but this is frequently impossible, necessitating a multi-stage "distance ladder" in which the calibration is referred back to parallax not directly, but through a series of intermediate stages. An error in any of these stages causes an error in H. One often has to correct for interstellar absorption, either in the Milky Way or in the target galaxy: this can be difficult to estimate. If you are close to the limiting magnitude of your technique, you will tend to see those test objects which are slightly brighter than average, and miss those which are slightly fainter, thereby calculating an average brightness which is systematically too great (an effect known as **Malmquist bias**), and hence a distance which is too small. If you are working at distances which are not all that large (say, <100 Mpc), your galaxy's own peculiar motion may be of the same order as its Hubble velocity, producing scatter on your plot; because galaxies tend to occur in clusters, this scatter may not be random (for example, our Galaxy is falling in towards the Virgo cluster, so recession velocities measured in the direction of Virgo are smaller than they should be). If you are working at larger distances, you are looking at younger objects (because of the finite speed of light), and need to worry about how their evolution might affect their brightness—for example, younger globular clusters might be brighter than the Milky Way's 10 Gyr old specimens.

In an ideal world, the astronomer corrects for the effects s/he can calculate, estimates the likely effect of those s/he cannot, and produces a reliable estimate of the systematic error of his/her final value. The fate of the Hubble constant in the years from 1958 to 1998 demonstrated that we do not live in an ideal world. A particular problem is that identifying and evaluating systematic errors is a difficult task, and can easily become biased: if you are "sure" that your measurement as initially analysed is too high, you will look for systematic effects that would make the true value lower, and *vice versa*. There is, therefore, a strong tendency towards concordance in an individual author's apparently independent measurements of H_0: he or she is likely to favour, consciously or unconsciously, assumptions about systematic errors that bring each measurement into line with his or her previous values.

This effect dominated the measurements of H_0 in the 1970s and 1980s, with Allan Sandage and colleagues consistently favouring low values of H_0 (around 50 km s^{-1} Mpc^{-1}) and **Gérard de Vaucouleurs** (1918–1995) and colleagues preferring larger values (around 100). Both camps consistently underestimated their errors, so that in the early 1980s we have Sandage quoting 50±7 and de Vaucouleurs 100±10—these are clearly not consistent with each other (the difference is over 4σ), nor with the currently accepted value of 69.3±0.7 (2.8σ discrepancy with Sandage and 3.1σ with de Vaucouleurs). The disagreement became very personal: here is the abstract from a 1983 paper by de Vaucouleurs (*MNRAS* **202** (1983) 367):



> **Summary**. The strong Malmquist bias claimed by Tammann, Sandage & Yahil (TSY) to be present in the extragalactic distance scale derived by the author from the luminosity index of 328 spiral galaxies and to vitiate the resultant value of H_0 is shown to be non-existent. The misunderstanding arose from (i) a confusion between Malmquist *effect* in a magnitude-limited galaxy sample and Malmquist *bias* in the derived distances; the former does not lead to the latter as long as the distance indicator (luminosity index) has a small dispersion and is free of distance-dependent systematic errors; (ii) an irrelevant application of the Schechter general luminosity function (valid for unrestricted samples of galaxies of all types) to a narrowly defined spiral sample; (iii) the unwarranted assumption that the velocity–distance relation is linear and isotropic at small redshifts and that absolute magnitudes can be simply derived from redshift; (iv) the use of a graphical presentation (M_T^0, V_0) where the two coordinates are correlated; (v) a lack of appreciation for the many safeguards against Malmquist bias built into the author's distance scale and a disregard for the precautions used to derive H_0 from the data in 'minimum bias' intervals of the parameters.

Despite the formal scientific language, this is clearly fighting talk. In fact, as shown by the figures above, both Sandage and Tammann and de Vaucouleurs were defending equally unrealistic values of *H* (and equally unrealistic error bars!). In a long review article in 1988, Michael Rowan Robinson (*Space Science Reviews* **48** (1988) 1-71) concluded that the best overall value at the time was 66±10 km s^{-1} Mpc^{-1}; it happens that 66 is exactly what you get if you take the weighted mean of 50±7 and 100±10! Rowan Robinson's value is consistent, within its errors, with the best current values, which are around 70 km s^{-1} Mpc^{-1}.

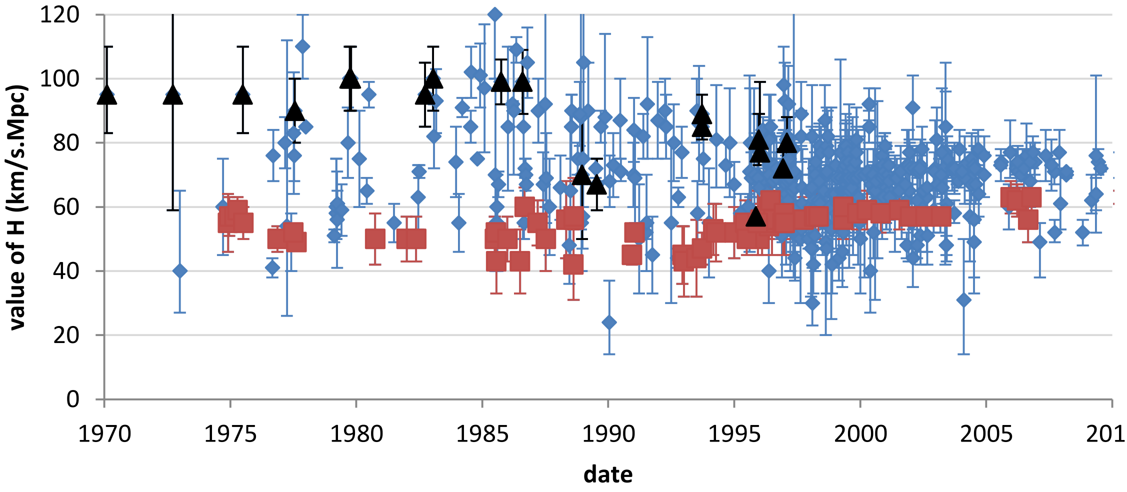


Figure 9.3: Hubble Wars. The plot is a compilation by the late John Huchra of published values of H_0. Squares indicate values by Sandage, Tammann et al.; triangles are by de Vaucouleurs or van den Bergh; diamonds are everyone else. Note the very strong tendency of the Sandage/Tammann values to be low, and the slightly less strong tendency of the de Vaucouleurs/van den Bergh values to be high.

The history of the Hubble constant in the later 20th century is a cautionary tale for experimentalists (and even more for theorists who are inclined to believe experimentalists too unquestioningly). There has never been the slightest suggestion that either de Vaucouleurs or Sandage has been dishonest in presenting and analysing the data; on the other hand, disinterested parties would find it hard to deny that both have allowed their expectations to affect their results. There is some ammunition here for those philosophers of science who argue that the expectations of experimenters colour experimental results sufficiently to render them invalid as tests of competing theories; however, it should be noted that the circumstances here (a measurement dominated by hard-to-quantify systematic errors) are atypical, and plenty of counterexamples exist where experimentalists did *not* get the results they expected. It is also true that in more recent years, estimates of the Hubble constant have become both less discordant and less polemical. The HST Key Project on the Hubble Constant, led by **Wendy Freedman** (b. 1957), used a variety of methods calibrated by HST observations of Cepheids in relatively nearby galaxies (the Cepheid distances themselves cannot be used because they do not extend far enough to be uncontaminated by local motions). They found a good level of concordance between different methods, in contrast to Rowan Robinson's 1988 study in which some methods appeared to have systematic offsets.

There remain some issues. The two CMB anisotropy probes, WMAP and *Planck*, do not agree with the best "conventional" measurements: WMAP give 69.3±0.8, *Planck* 67.8±0.9, whereas Riess et al., *ApJ* **826** (2016) 56, give 73.2±1.7, which is 2.1σ from WMAP's result and 2.8σ from *Planck*'s); in general, the scatter in the values of H_0 from different methods is still somewhat higher than one would predict from their quoted errors (see, for example, Neal Jackson, *Living Reviews in Relativity* **18** (2015) 2). Future historians will no doubt still find much to comment on!

9.6 Inflation

Early cosmologists assumed that the universe was homogeneous and isotropic because it made the equations easier to set up and solve. However, by 1980 observational data on the CMB and the large-scale distribution of active galaxies were beginning to show that homogeneity and isotropy were actually an excellent approximation to the real universe. This was somewhat of an embarrassment.

In Friedman-Lemaître models, the horizon (the edge of the "visible" universe) expands faster than the universe itself, with the result that the observer is continually gaining access to previously unseen regions of the universe. But if these regions had not previously been able to exchange photons, there is no obvious reason for them to be at the same temperature (which, clearly, they are—at least to 1 part in 10^5). It was also puzzling that the geometry of the universe seemed to be so close to "flat", when theory indicated that flatness was an unstable equilibrium—i.e., any small deviation from flatness is rapidly magnified as the universe expands.

The currently preferred solution to this problem was first introduced in 1981 by **Alan Guth** (b. 1947), a particle theorist. He recognised that the particle physics which applies in the ultra-high-temperature environment existing a small fraction (~10^{-35}) of a second after the Big Bang might result in phase transitions as the high-energy Grand Unified Theory breaks down into the low-energy Standard Model as seen in the present universe. If the universe were to supercool below this phase transition, the energy release could drive a period of exponential expansion which would "inflate" a small, causally connected piece of spacetime to many times the size of the currently visible universe, thus solving the horizon problem; the inflation also dilutes any curvature present to negligible values, thus solving the curvature problem.

This theory has many of the expected features of a new paradigm: it is a single and basically simple idea which elegantly resolves several anomalies of the existing paradigm. The most interesting thing about it, however, is that (as formulated by Guth) it doesn't work, and Guth knows it doesn't work. The problem is the lack of a "graceful exit" from the era of exponential expansion: to match observations, one needs to have a smooth transition back into Friedman-Lemaître-style expansion, and Guth's model fails to provide one. One would naïvely expect that such a theory would never see the light of print; however, in practice, the neatness of the solution to the pre-existing problems was such that Guth wrote the paper, and the highly respected journal *Physical Review* published it (vol. **D23**, p347), in the hope that "some variation can be found which avoids these undesirable features but maintains the desirable ones." This hope was rapidly realised: in early 1982, **Andrei Linde** (b. 1948), who had been working along the same lines independently, published a "new inflationary universe scenario" (*Phys. Lett*. **108B** (1982) 389), which again uses an established theory in particle physics ("the Coleman-Weinberg mechanism of symmetry breaking") to produce a model of inflation which does indeed avoid the "undesirable features" of Guth's original model.

Inflation rapidly became a standard "add-on" to the Hot Big Bang model, despite a lack of direct confirmation: its solutions to the horizon and flatness problems were sufficiently compelling, and its roots in pre-existing particle physics theories sufficiently deep, that it largely avoided accusations of being an artificial, *ad-hoc*, modification to a failing theory. Since then, the good fit of the CMB power spectrum to expectations from inflation (in which the anisotropies are normal quantum fluctuations blown up to macroscopic size) has provided the model with more tangible support.

Inflation was the first notable example of theoretical ideas from particle physics being used to describe the very early universe (theoretical ideas from nuclear physics had of course been applied to the early universe for some time). This was the start of a trend which has greatly accelerated in recent years: the discipline of **particle cosmology**, which attempts to use the early universe (as recorded in the CMB anisotropies and similar evidence) as a laboratory for ultra-high-energy particle physics, is now well established.

9.7 Summary

Even more than astrophysics, modern cosmology is a young science: for the first 50 years of its existence (1915–1965), it suffered from a severe shortage of observational data. **Malcolm Longair** (b. 1941), former Astronomer Royal for Scotland, once said that when he graduated from university in 1963 "cosmology was a science of 2½ facts" (the facts in question being that the sky is dark at night, that the universe is expanding, and that cosmic populations evolve over time—the last being only half a fact, as at that point it was suspected, on the basis of the radio source counts and the discovery of quasars, but not proven). Two and a half facts are not much to hang a whole discipline on, and it is not surprising that the cosmology of 1963 would still have been fairly familiar to Einstein, Friedman and Lemaître.

This situation began to change in 1965, with the discovery of the cosmic microwave background, but as late as the 1980s one could still object that, though there were now more than 2½ facts, such facts as there were suffered from a lamentable lack of precision: Alan Guth in his 1981 paper is only prepared to assume (admittedly conservatively) that $0.1 < \Omega_0 < 10$, and (as we have seen) the uncertainty in H_0 at this time was about a factor of 2.

Cosmology in the modern era is very different, with many parameters of the model fixed to within a few percent. The driving force has been the analysis of anisotropies in the cosmic microwave background—see papers by WMAP and *Planck*—backed up by ground- and space-based observations of galaxy clusters, gravitational lensing, galaxy redshift distributions and much else besides. This is a strongly technology-driven advance: many of the studies that have contributed to it require space-based instrumentation, and those that do not are often dependent on advances in telescope and imaging technology, such as adaptive optics and very large CCD cameras. Remarkably, much of the underlying theoretical framework has not changed at all in the past century—though the cosmological constant went through a long period of hibernation between 1929, when its original motivation was removed by Hubble's redshift-distance relation, and 1998, when work on Type Ia supernovae pointed to an accelerating rate of expansion. The introduction of inflation in 1981 changed our picture of the very early universe, but has little effect (except for setting initial conditions) on cosmology beyond the first second after the Big Bang—Guth has aptly described it as a "prequel" to the classic Big Bang model.

The current picture of cosmology, with its unidentified cold dark matter and completely mysterious dark energy, cannot be described as elegant—but the theoretical framework is thoroughly testable, and the observational tests that it has passed are stringent. The cosmology of the last few decades is a good example of Kuhn's "normal science", with steady progress being made in our understanding of a solid underlying paradigm. Crises and revolutions may yet lie ahead, but at the moment this situation seems rather stable.