Foreword

This year marks the Golden Jubilee of the founding of the Indian Academy of Sciences in April, 1934 by Professor C. V. Raman with the objective of promoting the progress and upholding the cause of science. Amongst the many activities undertaken by the Academy towards the achievement of its goals, publications have taken pride of place since the inception of the Academy and continue to do so. In addition to its Proceedings which have appeared without interruption since July, 1934, the Academy has published a large number of books, monographs, and in more recent years a number of specialist journals, now totalling eleven. The *Journal of Astrophysics and Astronomy*, the youngest of the Academy's journals, made its debut in September 1980 with an international Editorial Board whose Chairman was Professor M. K. Vainu Bappu. After his death in late 1982, we have attempted to maintain the high standards set by him for the journal and have also enlarged the Editorial Board.

To celebrate the completion of fifty years of scientific publishing by the Academy, all its journals are bringing out special issues. This is the first special issue of the *Journal of Astrophysics and Astronomy* containing specially invited articles from astronomers and astrophysicists of international repute. The next will appear in June, 1984. As shown by the papers in this issue, the Journal is open to the expression of views that are not always conventional. We believe with Zwicky that scientific "progress [lies in] the removal of prejudices and of absolute doctrines". The invited articles that follow also indicate well the wide range of topics the Journal intends to cover. On behalf of the Editorial Board, and the Academy, I would like to express our thanks to all the distinguished authors who responded so willingly to the invitation to contribute to these issues.

To mark the occasion, we have further attempted to make a significant improvement in the quality of the Journal. The type and printing have been improved and it appears now with a new cover design. Beginning with this issue the international distribution of the Journal will be handled by a professional agency*; it will also be covered by Current Contents/Physical, Chemical and Earth Sciences and the Science Citation Indexes.

Like all the other journals of the Academy, the *Journal of Astrophysics and Astronomy* levies no page charges. We shall maintain a strict control over the quality of its contents, ensure that it appears on time, and attempt to further reduce the publication lag. On this occasion, we would like to record our appreciation for the help rendered to the Journal by numerous authors, and a large number of referees from all over the world. We rely on them for their continued support.

V. Radhakrishnan
Chairman, Editorial Board

* Messrs J. C. Baltzer A. G., Scientific Publishing Company, Wettsteinplatz 10, CH-4058 Basle, Switzerland.

J. Astrophys. Astr. (1984) **5,** 3–11

The General Theory of Relativity: Why "It is Probably the most Beautiful of all Existing Theories"*

S. Chandrasekhar *Laboratory for Astrophysics and Space Research, Enrico Fermi Institute, University of Chicago, 933 East 56th Street, Chicago, Illinois 60637, U.S.A.*

By common consent, the general theory of relativity has a special aesthetic appeal to those who have studied it. I have chosen to quote Landau and Lifschitz from their *Classical Fields* in the title of my talk since their magnificent series of volumes, encompassing the whole range of physics, gives to their assessment a special authenticity. Others besides Landau and Lifschitz have applied the epithet 'beautiful to general relativity. Thus, Pauli, in his well-known article on 'The Theory of Relativity' in the *Encyclopädie der Mathematischen Wissenschaftien* (1921) has written

> This fusion of two previously quite disconnected subjects—metric and gravitation—must be considered as the most beautiful achievement of the general theory of relativity.

And in a similar vein, Dirac has written

> There was difficulty reconciling the Newtonian theory of gravitation with its instantaneous propagation of forces with the requirements of special relativity; and Einstein working on this difficulty was led to a generalization of his relativity—which was probably the greatest scientific discovery that was ever made.

In this lecture, I shall attempt to examine the origins and the reasons for the continuing belief that the general theory of relativity represents a beautiful scientific structure; and in this examination I shall try to be as objective as possible.

I

I shall begin with some remarks on the aesthetic impact which a discovery sometimes makes on the discoverer. That Einstein himself felt this aesthetic impact, when he finally arrived at his field equations, is evident from the concluding remark in his first

* Text of a Wolfgang Pauli Lecture given at Eidgenössische Technische Hochschule, Zürich (Switzerland), on January 9, 1984. A slightly different version was given as a Hans Bethe Lecture at Cornell University, Ithaca, New York (U.S.A.) on October 5, 1983.

preliminary announcement of his equations:

> Scarcely anyone who fully understands this theory can escape from its magic.

Einstein's reaction to his discovery is not very different from Heisenberg's reaction to his discovery that his matrix representation of the position and the momentum coordinates together with his commutation relation led to the correct energy levels of the simple harmonic oscillator. He has written

> . . . one evening I reached the point where I was ready to determine the individual terms in the energy table [Energy Matrix]. . . . When the first terms seemed to accord with the energy principle, I became rather excited, and I began to make countless arithmetical errors. As a result, it was almost three o'clock in the morning before the final result of my computations lay before me. The energy principle had held for all the terms, and I could no longer doubt the mathematical consistency and coherence of the kind of quantum mechanics to which my calculations pointed. At first, I was deeply alarmed. I had the feeling that, through the surface of atomic phenomena, I was looking at a strangely beautiful interior, and felt almost giddy at the thought that I now had to probe this wealth of mathematical structure nature had so generously spread out before me.

Heisenberg has recalled a meeting with Einstein soon after his discovery of quantum mechanics; and his remarks, at that meeting, as he has recounted them, illuminate the role of aesthetic sensibility in the discerning of great truths about nature:

> If nature leads us to mathematical forms of great simplicity and beauty . . . we cannot help thinking that they are "true," that they reveal a genuine feature of nature. . . . You must have felt this too: the almost frightening simplicity and wholeness of the relationships which nature suddenly spreads out before us and for which none of us was in the least prepared.

It may be argued that the aesthetic sensibility which may have guided Einstein or Heisenberg reflects only on their individuality; and that in any event the aesthetic appeal of a scientific insight is not really relevant to a judgement of its significance. One may indeed contend that the usefulness of any scientific discovery—be it theoretical or experimental—is to be measured only by its consequences. I shall not argue with those who take this pragmatic view. But it is fair to point out that the 'usefulness' of what one does is not always the prime motive for what one chooses to pursue. For example, Freeman Dyson has quoted Hermann Weyl as having said

> My work always tried to unite the true with the beautiful; but when I had to choose one or the other, I usually chose the beautiful.

I shall return later to give some examples of how, in the case of Hermann Weyl at any rate, his choice of the beautiful did eventually turn out to be true as well. But the task I now wish to set myself is to explain, as objectively as I can, why the general theory of relativity has had so strong an aesthetic appeal. In this attempt, I wish to be as serious as one is in literary or art criticisms of the works of Shakespeare or Beethoven.

II

At the outset one encounters a curious paradox. It is that the very beauty of the general theory of relativity is sometimes used as an argument for not pursuing it! Thus, Max Born has written

> It [the general theory of relativity] appeared to me like a great work of art, to be enjoyed and admired from a distance.

I am frankly troubled by Born's remark that the general theory of relativity is to be admired only from a distance. Is one to conclude that the theory does not require study and further development like the other branches of physical science?

I find it equally difficult to interpret a statement such as this of Rutherford:

> The theory of relativity by Einstein, apart from any question of its validity, cannot but be regarded as a magnificent work of art.

Apparently, beauty and truth are not to be confused!

For the present, I shall not concern myself with the question whether there can be beauty without truth. I shall turn instead to consider why a study of the general theory of relativity conduces in one a feeling not dissimilar to one's feelings after seeing a play of Shakespeare or hearing a symphony of Beethoven. But in attempting this task, it is useful to have some definite criteria for beauty in spite of the following view expressed by Dirac and shared by many:

> [Mathematical beauty] cannot be defined any more than beauty in art can be defined, but which people who study mathematics usually have no difficulty in appreciating.

I shall adopt the following two criteria for beauty. The first is that of Francis Bacon,

> There is no excellent beauty that hath not some strangeness in the proportion!

('Strangeness', in this context, has the meaning 'exceptional to a degree that excites wonderment and surprise.') And the second is that of Heisenberg:

> Beauty is the proper conformity of the parts to one another and to the whole.

III

That the general theory of relativity has some strangeness in the proportion, in the Baconian sense, is manifest. It consists primarily in relating, in juxtaposition, two fundamental concepts which had, till then, been considered as entirely independent: the concepts of space and time, on the one hand, and the concepts of matter and motion on the other. Indeed, as Pauli wrote in 1919, "The geometry of space-time is not given; it is determined by matter and its motion." In the fusion of gravity and metric that followed, Einstein accomplished in 1915 what Riemann had prophesied in 1854, namely, that the metric field must be causally connected with matter and its motion.

Perhaps the greatest strangeness in the proportion consists in our altered view of space-time with metric as the principal notion. As Eddington wrote: “Space is not a lot of points close together; it is a lot of distances interlocked.”

There is another aspect of Einstein’s founding of his general theory of relativity which has contributed to its uniqueness among physical theories. The uniqueness arises in this way.

We can readily concede that Newton’s laws of gravitation require to be modified to allow for the finiteness of the velocity of light and to disallow instantaneous action at a distance. With this concession, it follows that the deviations of the planetary orbits from the Newtonian predictions must be quadratic in v/c where v is a measure of the velocity of the planet in its orbit and c is the velocity of light. In planetary systems, these deviations, even in the most favourable cases, can amount to no more than a few parts in a million. Accordingly, it would have been entirely sufficient if Einstein had sought a theory that would allow for such small deviations from the predictions of the Newtonian theory by a perturbative treatment. That would have been the normal way. But that was not Einstein’s way: he sought, instead, an exact theory. And his only guides in his search for an exact theory were the geometrical base of his special theory of relativity provided by Minkowski and the principle of equivalence embodying the equality of the inertial and the gravitational mass. The empirical equality of the inertial and the gravitational mass, assumed to be exact, is at the base of the Newtonian theory of gravitation; and Newton gave it its supreme place by formulating it in the opening sentences of his *Principia.* But the equality, as Weyl has stated, is an ‘enigmatic fact’; and Einstein wished to eliminate this enigma. The fact that Einstein was able to arrive at a complete physical theory with such slender guides has been described by Weyl as “one of the greatest examples of the power of speculative thought.” There is clearly an element of revelation in the manner of Einstein’s arriving at the basic elements of his theory. One feels, as Weyl has expressed, “it is as if a wall which separated us from Truth has collapsed.”

IV

The general theory of relativity thus stands beside the Newtonian theory of gravitation and motion, as the only examples of a physical theory born whole, as a perfect chrysalis, in the single act of creation of a supreme mind. It is this feature of the general theory of relativity, more than any other, that is normally in one’s mind when one describes the theory as a “great work of art to be admired from a distance.” But for a serious student of relativity, the aesthetic appeal derives even more from discovering that at every level of further understanding, fresh strangenesses in the proportion emerge always in conformity of the parts to one another and to the whole, even as an iridescent butterfly emerges from a chrysalis. I should like to give some illustrations of this feature of the theory. But to the extent they are illustrations, they may reflect my own perspective of the theory. I am sure that others will choose other illustrations.

My first illustration will relate to the solutions which the general theory of relativity provides as a basis for the description of the black holes of nature.

It is now a matter of common knowledge that black holes are objects so condensed that the force of gravity on their surfaces is so strong that even light cannot escape from them. The most elementary physical ideas combined with the most rudimentary facts

concerning stars, their sources of energy and their evolution, dictate their occurrence in very large numbers in the astronomical universe. This is not the occasion, and I do not have the time either, to elaborate on these astrophysical matters. I shall turn instead to what the general theory of relativity has to say about them. For this purpose, it is necessary to give a somewhat more precise definition of a black hole than I have given.

A black hole partitions the three-dimensional space into two regions: an inner region which is bounded by a smooth two-dimensional surface called the *event horizon*; and an outer region external to the event horizon which is *asymptotically flat*; and it is required that no point in the inner region can communicate with any point of the outer region. This incommunicability is guaranteed by the impossibility of any light signal, originating in the inner region, crossing the event horizon. The requirement of asymptotic flatness of the outer region is equivalent to the requirement that the black hole is isolated in space, which means only that far away from the event horizon the space-time approaches the customary space-time of terrestrial physics.

In the general theory of relativity we must seek solutions of Einstein's vacuum equations compatible with the two requirements I have stated. It is a startling fact that compatible with these very simple and necessary requirements, the general theory of relativity allows for stationary (*i.e.,* time-independent) black holes exactly a single, unique, two-parameter family of solutions. This is the Kerr family, in which the two parameters are the mass of the black hole and the angular momentum of the black hole. What is even more remarkable, the metric describing these solutions is simple and can be explicitly written down.

I do not know if the full import of what I have said is clear. May I explain.

As I have already stated, there are innumerable black holes in the present astronomical universe. They are macroscopic objects with masses varying from a few solar masses to millions of solar masses. To the extent they may be considered as stationary and isolated, they are all—every one of them—described exactly by the Kerr solution. This is the only instance we have of an exact description of a macroscopic object. Macroscopic objects, as we see them all around us, are governed by a variety of forces derived from a variety of approximations to a variety of physical theories. In contrast, the only elements in the construction of black holes are our notions of space and time. They are thus, almost by definition, the most perfect among all the macroscopic objects we know. And since the general theory of relativity provides a single unique two-parameter family of solutions for their description, they are the simplest objects as well.

As I have said on another occasion, Kerr's discovery of his solution is the only astronomical discovery comparable to the discovery of an elementary particle in physics; but in contrast to elementary particles, the black holes are pristine in their purity.

V

Again we need not be content with the discovery of the Kerr solution. We can study its properties in a variety of ways: by examining, for example, the manner of the interaction of the Kerr black-hole with external perturbations such as the incidence of waves of different sorts. Such studies reveal an analytic richness of the Kerr space-time which one could hardly have expected. This is not the occasion to elaborate on these

technical matters. Let it suffice to say that, contrary to every prior expectation, all the Standard equations of mathematical physics can be solved exactly and explicitly in the Kerr space-time. The Hamilton-Jacobi equation governing the motion of test particles, Maxwell's equations governing the propagation of electromagnetic waves, the gravitational equations governing the propagation of gravitational waves, and the Dirac equation governing the motion of electrons, all of them can be separated and solved explicitly in Kerr geometry. And the solutions predict a variety of physical phenomena which black holes must exhibit in their interaction with the outside world.

Let me illustrate by one particular process, discovered by Roger Penrose, which can take place in such interactions. It is that one can extract, under suitable conditions, the rotational energy of a black hole. When this phenomenon was first investigated, one found that such extraction of energy was accompanied by an increase in the surface area of the black hole. Generalizing this result, Hawking was able to prove an 'area theorem' to the effect that any interaction, experienced by a black hole, in which energy is exchanged, must result in an increase in its surface area. This fact suggests that the surface area of a black hole is in some sense analogous to thermodynamic entropy which has also the monotonic property of always increasing. By considering the quantum mechanics of pressure-free gravitational collapse, Hawking soon showed that this is more than an analogy and that one can, without ambiguity, define not only the entropy of a black hole but a surface temperature as well; and also that there is a flux of radiation from the surface of a black hole with a Planck distribution for the temperature that was assigned.

I stated earlier that one of the remarkable features of Einstein's formulation of general relativity was its bringing into a direct relationship the geometry of the space-time with its content of matter and motion. It is "this fusion of two previously quite disconnected notions" that Pauli found as the "most beautiful achievement of the general theory of relativity." We now find in Hawking's synthesis a still grander fusion of geometry, matter, and thermodynamics. There is clearly no lack in the strangeness in the proportion which a further study of relativity does not reveal.

VI

Let me consider one last illustration. It relates to certain singularity theorems proved by Penrose and Hawking. The theorems state, in effect, that the occurrence of singularities in space-times is generic to general relativity. Roughly speaking, what this statement means is that during the course of evolution of material objects, there exist 'points of no return' such that the trespassing of these points will necessarily lead, inexorably, to singularities. This theorem provides in fact the strongest reason for our present belief that our universe started with an initial singularity. The reason is that from the existence of the three-degree microwave-radiation, we can conclude that the universe retained its present homogeneity and isotropy when its radius was some one thousand times smaller than the present. It follows from this result and some additional astronomical facts that the universe was already then (or a little earlier)at a point of no return; and the inference of an initial past singularity cannot be avoided. The problems associated with the conditions just preceding the initial singularity thus become a necessary part of current investigations both in cosmology and in physics.

VII

So far I have considered the aesthetic appeal of the general theory of relativity in the manner of its founding and in the matter of its implications. But Poincaré, who has often emphasized the role of beauty in the motivations for scientific pursuits, has also stated that the "value of a discovery is to be measured by the fruitfulness of its consequences." I shall therefore consider some of the "fruitful consequences" of the general theory of relativity. Since astronomy is the natural home of general relativity, we must seek for its consequences in astronomy. I shall consider two such consequences. Both of them relate to certain crucial respects in which considerations of relativity have altered the astrophysicist's views relative to the stability of stars and stellar systems.

It is well known that in the framework of the Newtonian theory, the condition for the dynamical stability of a star, derives from its modes of radial oscillations and, that for stability the average ratio of the specific heats γ (defined as the ratio of the fractional changes in the pressure and in the density for adiabatic changes) must exceed 4/3. Alternatively, a star will become dynamically unstable if γ, or some average of it, is less than 4/3. This Newtonian condition is changed in the framework of general relativity: a star with an average ratio of specific heats γ, no matter how high, will become unstable if its radius falls below a certain determinate multiple of the Schwarzschild radius, $R_S = 2GM/c^2$ (where M denotes the mass of the star, G is the constant of gravitation, and c is the velocity of light). It is this fact which is responsible for the existence of a maximum mass for stable neutron stars. I may parenthetically point out that this important result is closely related to an early deduction of Karl Schwarzschild that a star in hydrostatic equilibrium must necessarily have a radius exceeding $\frac{9}{8} R_s$: this is the radius at which a star, with a ratio of specific heats tending to infinity, becomes unstable.

This instability of relativistic origin, discovered some twenty years ago, plays a central role in all current discussions pertaining to the onset of instability during the course of evolution of massive stars prior to gravitational collapse.

There is another consequence of general relativity for the stability of neutron stars. The instability to which I now refer was discovered some ten years ago and derives from a dissipative phenomenon which general relativity naturally builds into the theory of non-axisymmetric oscillations of gravitating masses. The dissipation results from the emission of gravitational radiation with accompanying loss of energy and angular momentum. The manner in which this mode of dissipation of energy and angular momentum induces instability is in some ways similar to the manner in which viscous damping sometimes induces instability. It now appears, especially from the work of John Friedman, that this mode of instability sets a limit to the rotation of pulsars and bears on the stability of fast pulsars like the ones that have recently been discovered.

It is clear, then, that there are fruitful consequences of the general theory of relativity for the astronomer's view of the universe. He need not be content with admiring general relativity from a distance.

IX

I now turn to a somewhat more general question concerning the relation of truth to beauty in science.

I made a reference earlier to a statement of Weyl's to the effect that in his work he always tried to unite the true with the beautiful and that, when he had to make a choice he generally chose the beautiful. An example which Weyl gave was his gauge theory of gravitation, developed in his *Raum, Zeit, und Materie* (Space, Time, and Matter, 1918). Weyl became convinced that his theory was not true as a theory of gravitation; but he nevertheless kept it alive because it was beautiful But much later, it did turn out that Weyl's instinct was right after all: the formalism of gauge invariance was incorporated into quantum electrodynamics. A second example is provided by the two-component relativistic wave-equation of the massless neutrino. Weyl discovered this equation and the physicists ignored it for some thirty years because it violated parity invariance. And again it turned out that Weyl's instinct was right: he had discerned truth by trusting to what he conceived as beautiful.

A similar example is provided by Kerr's discovery of his solution. Kerr was not seeking solutions that would describe black holes. He was seeking instead solutions of Einstein's equation which had a very special algebraic property. But once he had discovered his solution, he could show quite readily that it did indeed describe a black hole. But its uniqueness for representing black holes was established only ten years later by Edward Robinson.

The foregoing examples provide evidence that a theory developed by a scientist with an exceptionally well-developed aesthetic sensibility can turn out to be true even if at the time of its formulation, it did not appear relevant to the physical world.

It is, indeed, an incredible fact that what the human mind, at its deepest and most profound, perceives as beautiful finds its realization in external nature.

What is intelligible is also beautiful.

We may well ask: how does it happen that beauty in the exact sciences becomes recognizable even before it is understood in detail and before it can be rationally demonstrated? In what does this power of illumination consist?

These questions have puzzled many since the earliest times. Thus, Heisenberg has drawn attention, precisely in this connection, to the following thought expressed by Plato in the *Phaedrus*:

> The soul is awestricken and shudders at the sight of the beautiful, for it feels that something is evoked in it, that was not imparted to it from without by the senses, but has always been already laid down there in the deeply unconscious region.

The same thought is expressed in the following aphorism of David Hume:

> Beauty in things exists in the mind which contemplates them.

Kepler was so struck by the harmony of nature as revealed to him by his discovery of the laws of planetary motion that in his *Harmony of the World*, he wrote:

> Now, it might be asked how this faculty of the soul, which does not engage in conceptual thinking and can therefore have no prior knowledge of harmonic relations, should be capable of recognizing what is given in the outward world. . . . To this, I answer that all pure Ideas, or archetypal patterns of harmony, such as we are speaking of, are inherently present in those who are capable of apprehending them. But they are not first received into the mind by a

> conceptual process, being the product, rather, of a sort of instinctive intuition and innate to those individuals.

More recently, Pauli, elaborating on these ideas of Kepler, has written:

> The bridge, leading from the initially unordered data of experience to the Ideas, consists in certain primeval images pre-existing in the soul—the archetypes of Kepler. These primeval images should not be located in consciousness or related to specific rationally formulizable ideas. It is a question, rather, of forms belonging to the unconscious region of the human soul, images of powerful emotional content, which are not thought, but beheld, as it were, pictorially. The delight one feels, on becoming aware of a new piece of knowledge, arises from the way such pre-existing images fall into congruence with the behaviour of the external objects. . . .

Pauli concludes with

> One should never declare that theses laid down by rational formulation are the only possible presuppositions of human reason.

It is clear that following these thoughts one is dangerously led into the path of the mystical. I shall desist following this path but conclude instead by quoting two ancient mottoes:

The simple is the seal of the true

and

Beauty is the splendour of truth.

J. Astrophys. Astr. (1984) **5,** 13–18

Type I Supernovae and Iron Nucleosynthesis in the Universe

I. S. Shklovskiĭ *Space Research Institute, USSR Academy of Sciences, Profsoyusnaya 88, 117810 Moscow, USSR*

Abstract. It is argued that the iron nucleosynthesis rate in the universe due to SNI outbursts is dependent on the mass function of star formation. Since the mass function depends on the chemical composition and since the masses of SNI precursors have upper limits, the iron nucleosynthesis rate was low at an earlier evolutionary epoch of the universe when mainly massive stars were formed. The iron nucleosynthesis rate should reach a maximum near $z \sim 0.5$. At such or similar value of z the well-known 'step' in the cosmic γ-ray background spectrum may be explained by the presence of γ-gray quanta accompanying the radioactive $^{56}\mathrm{Co} \rightarrow {}^{56}\mathrm{Fe}$ decay. An argument is presented against the identification of the hidden mass of the universe with black-hole remnants of 'type III' stars.

Key words: pregalactic stars—Supernovae—nucleosynthesis—cosmic γ-ray background

The idea that each outburst of a type I supernova (SNI produces about $1M_\odot$ of radioactive ^{56}Ni in the decay of which, with a half-life of 6.1d, radioactive ^{56}Co is generated transforming (half-life 77 d) in its turn into a stable isotope of iron ^{56}Fe, has a fairly long history (*cf.* Colgate & McKee 1969). The analysis of the spectra of SNI 1972e at a late stage of its evolution has yielded convincing arguments in favour of this idea (Kirshner & Oke 1975). These spectra have most reliably shown that beyond 50 d after the maximum, the SNI radiation in the visual band is determined by the blending of allowed and forbidden lines of Fe I and Fe II . According to Kirshner & Oke, the mass of ionized iron in the shell of this SN is about $10^{-2}\ M_\odot$. Meyerott (1980), and independently Shklovskiĭ (1981) have shown that in such a shell, iron should mainly be present as Fe III with the total mass of about $1M_\odot$.

Until recently, the hypothesis of 'radio-active nickel' faced a serious difficulty: X-ray spectroscopy methods failed to reveal an anomalously high iron abundance in the remnants of historic SNI (Tycno 1572, Kepler 1604, and 1006), though many attempts have been made. However, the IUE satellite observations of the fairly faint blue star SM onto which the central part of the SNR 1006 remnant is projected, helped to identify in its spectrum wide ($\Delta v \sim 5 \times 10^3$ km s^{-1}) and intense (apparently saturated) absorption lines of multiplets I, II, III of ionized iron. The total amount of iron in the ejected shell may reach 1 $M_\odot$ (Wu *et al.* 1983).

These observations show most convincingly that each SNI outburst does generate about $1M_\odot$ of iron. Meanwhile, observations in the optical and X-ray spectral regions do not imply any excess iron in the shells and remnants of larger-mass type II

Supernovae (SNII). Apparently SN of this type are responsible for the nucleosynthesis of such abundant nuclei as C, O, N and Si. We may thus postulate that iron nucleosynthesis in the universe owes its origin only to SNI outbursts. Though it is impossible at present to prove this postulate, it seems fairly well established empirically.

The postulate that only SNI are responsible for Fe nucleosynthesis enables several important cosmological conclusions. We begin by noting that the masses of stars exploding as SNI are comparatively small, or at least have an upper limit. It is a simple and well-known fact that SNI outbursts are not connected with the spiral structure of galaxies (Maza & van den Bergh 1976), that allows the conclusion that masses of SNI precursors are $< 7M_{\odot}$. We have argued recently that SNI represent the final evolutionary stage of stars whose core masses differ only slightly from the Chandrasekhar limit M_{Ch} (Shklovskiĭ 1983b)*. According to Paczynski (1970), such stars have initial masses from 3 to $7M_{\odot}$. Stars with masses over $7M_{\odot}$ form, during their evolution, cores whose masses exceed M_{Ch}. The evolution of these stars ends in an SNII explosion. Finally, stars of comparatively small mass—with a core mass smaller than M_{Ch}—end their evolution as white dwarfs prior to which their outer shell is detached producing a planetary nebula. Thus, iron nucleosynthesis occurs during the final stage of evolution of stars with initial masses within a comparatively narrow range of $3–7M_{\odot}$†. The question is when these stars might form.

The mass function $\psi(M)$ of stars newly formed from the diffuse medium depends on the chemical composition of the latter, or, on the percentage of heavy elements (by mass) Z, to be more exact. If $\psi(M) = AM^{-\alpha}$, then, as Terlevich & Melnik (1983) showed recently for galactic and metagalactic objects, the following empirical relation is valid:

$$\alpha = \log Z + 5.05. \tag{1}$$

This relation implies, *inter alia*, that if the diffuse medium from which stars form could be very poor in heavy elements (*e.g.* $Z \sim 10^{-5}$), mainly massive stars would form in it. Thus there would be no stars capable of exploding as SNI at the end of their evolution. Therefore, no iron nucleosynthesis would occur.

Recently a new interest has been shown in hypothetical stars of 'population III' ('zero' generation) which apparently preceded the formation of contemporary stars and galaxies (see *e.g.* Bond, Carr & Arnett 1983). As stars of this type should form from the primordial hydrogen-helium medium with negligible amounts of other elements, Z is very small, and therefore the mass function exponent α should be negative. This means that only stars of large (perhaps very large) mass could form at that epoch and no SNI outbursts or related iron nucleosynthesis would occur.

Stars with the masses corresponding to the precursors of SNI could form only after the interstellar diffuse medium became sufficiently enriched with heavy elements. It might possibly occur, for example, when zero-generation stars evolved and most of them exploded as SNII. According to Bond, Carr & Arnett (1983), the evolution of stars of very large mass should be accompanied by the formation of a considerable number of heavy elements (up to 10 per cent by mass). However, if zero-generation stars really did exist, they could hardly enrich the diffuse medium to an appreciable degree. The fact

* Recent estimates show that among the three bright nuclei of planetary nebulae in the Magellanic Clouds (the distance to which is well known and prevents possible mistakes) one nucleus has a mass of $1.2M_{\odot}$, which is sufficiently close to M_{Ch} (Stecher *et al.* 1982).

† SNI outbursts may also occur in old binary systems due to gas accretion onto a degenerate component after the mass of the latter reaches M_{Ch}.

that among oldest galactic populations (globular clusters, halo stars) there are stars extremely poor in metals ($Z < 10^{-4}$) should imply that the diffuse medium from which they formed had a similar low Z. But, in line with the assumption, the medium should have been enriched with heavy elements, by the end products of zero-generation stars. Hence it follows that only an insignificant part of the primordial diffuse medium ($\sim 10^{-1}$ to 10^{-2}) might have condensed into stars of 'zero' generation. For during the evolution of such stars most of their material should have been reworked into heavy elements and swept away into the interstellar space. But if the total mass of 'zero'-generation stars were essentially smaller than the mass of the interstellar medium, the 'hidden' mass of the universe cannot be explained by black-hole remnants of such stars. An impression sets in that stars of zero generation (population III) do not yet provide a clear possibility of the evolution of matter in the universe*. There is no need to introduce into cosmology those hypothetical objects to explain nucleosynthesis. We should then consider a continuous enrichment of iron in the universe, and the enrichment rate should be related to the variation of the mass function of star formation, which, in turn, depends on the continuous increase of Z.

The following circumstance should be emphasized. In the 'old' objects with a reduced abundance of heavy elements, the abundance ratio Fe/O is much lower than that of the Sun's. For example, according to Peimbert (1973) the abundance ratio Fe/O in the planetary nebula K 648 that belongs to the globular cluster M 15, is lower by about a factor of 10 compared to the solar value. Sneden, Lambert & Whitaker (1979) have showed that this ratio is lower than the solar value by a factor of 3 in stars of low metallicity. The extended X-ray halo of M 87 has a similar deficiency of Fe/O (Canizares *et al.* 1982), though the stellar halo has the same O/H as the Sun. We have recently interpreted these observations on the assumption that the SNI precursors have a low mass of 1.5–2$M_\odot$ (Shklovskiĭ 1983a). However, reliable observational and theoretical data indicates a much larger precursor mass of about 3–7$M_\odot$ (Clayton & Silk 1969). Hence the explanation we have given earlier for the smallness of abundance of old objects seems to be wrong.

Much more natural, in our opinion, is the assumption of a gradually changing mass function of newly generated stars, the change being the result of a continuous increase of heavy-element abundance in the interstellar medium. At the epoch when the process of star formation was just beginning, Z was very small and so was accordingly the index α in Equation (1). At this epoch almost no stars formed in the range of 3–7 $M_\odot$, and the iron enrichment of the interstellar medium was slow. However, the rate of production of comparatively massive stars having been high, the interstellar medium was being enriched with lighter elements (C, O, N and Si) at a high speed, inducing a change in the mass function and a growing number of stars in the mass range corresponding to the precursors of SNI. The Fe nucleosynthesis rate first became equal to and then larger than, that of lighter 'metals'.

Since Z increases continuously, the Fe nucleosynthesis rate should decrease after reaching the maximum at a certain redshift, say, z_1. It is of interest to estimate the epoch $T_1 = T_0\ (1 + z_1)^{-3/2}$ (T_0 being the age of the universe), when the rate reached the maximum. The rigorous mathematical consideration of this problem requires the knowledge of parameters such as the absolute value of the rate of star formation in

* In our view, the recent discovery of an infrared background radiation announced by Matsumoto, Akiba & Murakami (1983) and interpreted by them as the total radiation from zero-generation stars red-shifted by $z \sim 10$, is questionable and needs verification.

galaxies of various types, of the frequency and intensity of flashes of star-formation, *etc.* Values of these parameters are not even known to a first approximation. A method could, however, be suggested which—at least in principle—may help solve this problem. It is well known that γ-ray lines with the energy of ~ 1 MeV are emitted due to the radio-active decay in an SNI shell. Since even before half of ^{56}Co has decayed, the SNI shells become semi-transparent to radioactive γ-radiation, a considerable part of that radiation first enters the interstellar, and next the inter-galactic medium where the probability of their absorption is negligible. Thus it may be expected that the γ-ray lines, accompanying iron nucleosynthesis in SNI outbursts, will be present in the background cosmic hard-photon radiation.

According to Meyerott (1980), the surface density of an SNI shell is several g cm^{-2} after $t_{1/2}$ = 77 d (^{56}Co half-life). At this epoch, the absorption coefficient—the main contribution to which is from the Compton effect—equals 0.25 cm^2 g^{-1}, and hence the optical depth of the shell for γ-ray quanta generated within is τ_γ ($t = t_{1/2}$) ~ 1· Taking into account the attenuated power of radioactive γ-radiation, an assumption can be made that 0.3 of the quanta produced in the radioactive decay of ^{56}Co diffuse into the interstellar space, and then into the metagalaxy.

Thus a certain spectral feature should be expected in the background of the metagalactic isotropic γ-radiation. This idea was first suggested by Clayton & Silk as early as in 1969. However, accurate measurements of the γ-ray background had not yet been made at the time, and the nature of SNI phenomenon was much farther from understanding than it is now. Besides, Clayton & Silk believed that the rate of enrichment in the universe is either constant or increases inversely proportional to that of the universe. On the other hand, according to the above considerations, it reaches a fairly smooth maximum at $z = z_1$. This implies that there is a fairly wide spectral feature in the hard background radiation spectrum, that is, a radiation band. What are the chances of observing it?

Assume that the 'smeared' density of matter in the universe is ρ_0. The local density of γ-quanta produced when iron nuclei form via the ^{56}Co → ^{56}Fe radioactive decay will then be

$$n_\gamma = \frac{\rho_0}{m_{\mathrm{Fe}}}\,\delta\xi \tag{2}$$

where m_{Fe} is the mass of the iron nucleus, $\delta \simeq 10^{-3}$ is the present average cosmic Fe abundance, ξ is the fraction 0.3 of γ-ray quanta freely leaving the SN I shell. The intensity of this radiation calculated per unit energy interval is

$$I = \frac{n_\gamma c}{4\pi}\,\frac{E}{\Delta E}\,\mathrm{keV\,cm^{-2}\,sr^{-1}\,keV^{-1}} \tag{3}$$

where E is the energy of the quanta and ΔE the width of the spectral region. We assume that $\Delta E \sim E$ and $\rho_0 = 10^{-30}$ g cm^{-3}. Then

$$I \sim 10^{-2}\,\mathrm{keV\,cm^{-2}\,sr^{-1}\,keV^{-1}} \tag{4}$$

Observations have long ago shown a 'step' (bending) in the spectrum of the background cosmic γ radiation for E = 1 to 2 MeV (*cf.* Ramaty & Lingenfelter 1982b). The background intensity in the considered range is just equal to the value needed. The energies of γ-ray quanta of lines that occur in ^{56}Co radioactive decay are 0.845 (1), 1.26

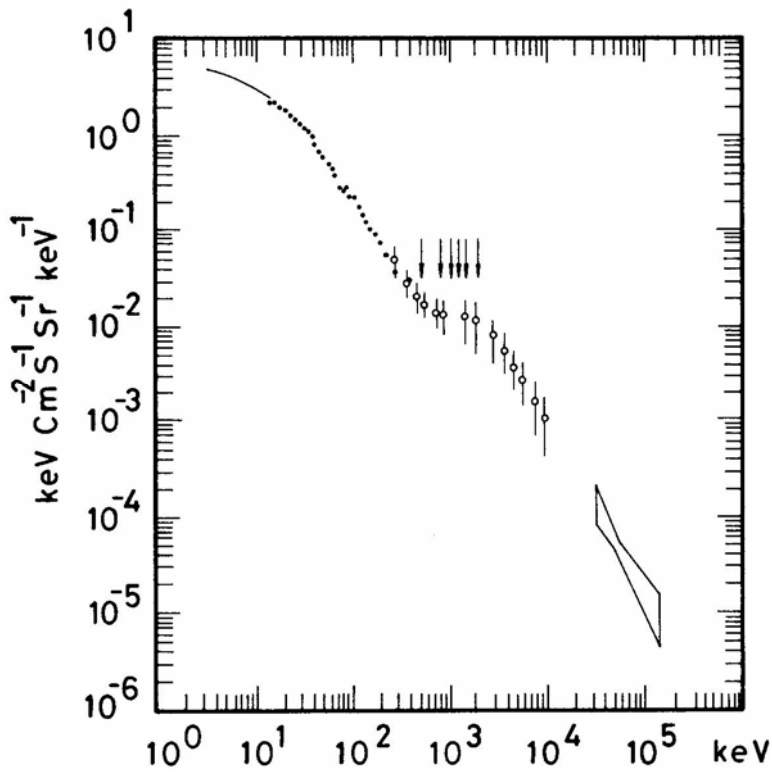


Figure 1. The diffuse X-ray and gamma-ray back ground spectrum. Nuclear lines of ^{56}Fe, 0.85, 1.26, 1.74, 2.01, 2.55 and 3.25 MeV which are radiated during the decay of ^{56}Co (Ramaty & Lingenfelter 1982a) are shown (arrows) corresponding to a cosmological redshift $z = 0.5$.

(0.5), 1.74 (0.2), 2.01 (0.1), 2.55 (0.2) and 3.25 (0.2) MeV. (The numbers in parentheses are the relative intensities of the respective lines.) If these lines are assumed to be responsible for the step in the spectrum of the background soft cosmic γ-radiation, they need to be redshifted by $z_1 \sim 0.5$. It should be kept in mind that the step is observed in the portion of the spectrum showing steep rise towards lower energies. Thus less intense though harder quanta make the largest contribution. The epoch of maximum Fe-production rate $T_1 \simeq T_0/(1 + z_1)^{3/2} = 0.5T_0$ corresponding to $z_1 \sim 0.5$; here T_0 is the present age of the universe. It is evident that the estimate of the intensity of γ-ray quanta accompanying iron nucleosynthesis is very crude. New high-quality observations of the cosmic background should be carried out in the range 0.8–3 MeV.

References

Bond, J., Carr, B., Arnett, D. 1983, *Nature*, **304,** 514.
Canizares, C., Clark, G., Jernigan, I., Markert, T. 1982, *Astrophys. J.*, **262,** 33.
Clayton, D., Silk, J. 1969, *Astrophys. J.*, **158,** L43.
Colgate, S. A., McKee, C. 1969, *Astrophys. J.*, **157,** 623.
Kirshner, R. P., Oke, J. B.1975, *Astrophys. J.*, **200,** 574.
Matsumoto, T., Akiba, M., Murakami, T. 1983, *Preprint*, Nagoya Univ.

Maza, J., van den Bergh, S. 1976, *Astrophys. J.*, **204,** 519.
Meyerott, R. 1980, *Astrophys. J.*, **239,** 257.
Paczynski, B. 1970, *Acta Astr.*, **20,** 47.
Peimbert, M. 1973, *Mém. Soc. R. Sci. Liége 6th Ser.*, **65,** 227.
Ramaty, R., Lingenfelter, R. E. 1982a, *A. Rev. nucl. part. Sci.*, **32,** 235.
Ramaty, R., Lingenfelter, R. E. 1982b, *Space Sci. Rev.*, **36,** 305.
Shklovskiĭ, I. S. 1981, *Sov. Astr.*, **25,** 578.
Shklovskiĭ, I. S. 1983a, *Nature*, **304,** 513.
Shklovskiĭ, I. S. 1983b, *Sov. Astr. Lett.*, in press.
Sneden, C., Lambert, D. L., Whitaker, R. W. 1979, *Astrophys. J.*, 234, 964.
Stecher, T., Maran, S., Gull, T., Aller, L., Savedoff, M. 1982 *Astrophys. J.*, **262,** L41.
Terlevich, R., Melnick, J. 1983, *Preprint*, ESO No. 264.
Wu, C-ch., Leventhal, M., Sarazin, C., Gull, Th. 1983, *Astrophys. J.*, **269,** L5.

J. Astrophys. Astr. (1984) **5,** 19–30

Measuring the Sizes of Stars*

R. Hanbury Brown *Chatterton Astronomy Department, School of Physics, University of Sydney, N.S.W. 2006, Australia*

Abstract. The Chatterton Astronomy Department aims to apply interferometers with very high resolving power to optical astronomy. The programme of the stellar intensity interferometer at Narrabri Observatory was completed in 1972 and since then the work has been directed towards building a more sensitive instrument with higher resolving power. As a first step a much larger intensity interferometer was designed but was not built because it was large, expensive and not as sensitive as desired. Efforts are now being made to design a more sensitive and cheaper instrument. A version of Michelson's stellar interferometer is being built using modern techniques. It is hoped that it will reach stars of magnitude +8 and will work reliably in the presence of atmospheric scintillation. It is expected to cost considerably less than an intensity interferometer of comparable performance. The pilot model of this new instrument is almost complete and should be ready for test in 1984.

Key words: stars, angular diameter—interferometry—atmospheric scintillation

1. Introduction

The long-term objective of the Chatterton Astronomy Department of the School of Physics (University of Sydney) is to apply interferometers with very high angular resolving power to optical astronomy. This has already been done with great success in radio-astronomy and there is good reason to believe that it would be equally fruitful in optical astronomy. We are apt to forget that the progress of astronomy, indeed of science, depends intimately on developing new tools and methods of observing the world around us.

The first objective of our programme was to measure the apparent angular diameters of the bright visible stars. We were, by-the-way, not the first people to try to do this; Galileo, for example, tackled the problem experimentally. He hung a fine silk cord vertically, and then measured the greatest distance from his eye at which this cord could be made to occult the bright star Vega. In this way he came to the conclusion that the angular diameter of Vega is about 5 arcsec. In doing this experiment Galileo was trying to find an answer to one of the most troublesome objections to the idea, suggested by Copernicus, that the Earth orbits the Sun. It had been pointed out that, if the Earth really does go around the Sun, then the bright stars should appear to move relative to fainter and more distant stars (orbital parallax). By measuring the angular size of Vega

* Text of an Academy Lecture delivered at the Raman Research Institute, Bangalore on January 27, 1984.

and by assuming that it is a similar body to the Sun, Galileo was able to show that the bright stars are so far away that their annual movement in the sky would not be detectable. For reasons which we now understand, Galileo's measurement gave an absurdly large value for the angular size of Vega; nevertheless it served his purpose.

Our own interest in measuring angular diameters is different. If, for example we measure the angular diameter θ of a star and we also know its distance d, then by simple trigonometry we can find its actual physical diameter $D = d\theta$. Alternatively we can find the flux of light F_λ emerging from its surface if we combine our measurement of θ with a measurement of the flux of light f_λ received from the star at the Earth, where

$$F_\lambda = 4f_\lambda/\theta^2 \tag{1}$$

The quantity F_λ gives us the actual flux of light radiated by unit area of the star's surface and is fundamental to the study of models of the star's atmosphere.

Another important piece of information about a star which we can find is the effective temperature T_e of its surface. To do this we measure the flux of light f_λ received over the whole of the spectrum and then compute T_e from,

$$\int_\lambda F_\lambda \mathrm{d}\lambda = \sigma T_e^4 \tag{2}$$

where σ is Stefan's constant.

2. The major difficulties in measuring angular size

The first major difficulty is to make an instrument with sufficiently high resolving power to measure the extremely small angles which are involved. For example, if we aim to measure a reasonable sample of main sequence stars we must measure angles of the order of 10^{-4} arcsec which, at optical wavelengths, necessarily involves building instruments with baselines of 100 m and more—and that is not easy.

The second major difficulty is to make precise and reliable measurements in the presence of atmospheric scintillation. Turbulence in the atmosphere inevitably introduces fluctuations into the relative time of arrival of the starlight at separated points on the Earth. For small separations these fluctuations correspond roughly to changes in pathlength of about $10^{-6}\ D$ where D is the separation. What happens at greater spacings, such as 100 m, is not yet known and it would be very interesting to know.

Another effect of turbulence is to introduce temporal and spatial fluctuations into the amplitude and phase of the wavefront of the light from a star. The temporal fluctuations are known to have a frequency spectrum which extends up to 20 or 30 Hz or even more, depending on the wind speed. The spatial fluctuations have a characteristic size which depends upon the site, the weather, and the time of day; typically they have a characteristic length of 10 cm.

2.1 *Michelson's Stellar Interferometer*

The first successful attempt to measure the angular diameter of a star was made by Michelson and his colleagues in 1920 using an interferometer mounted on the 100-inch telescope at Mt Wilson. The number of stars which they could measure was severely

limited by the maximum possible separation (20 ft) between the two mirrors of their instrument which restricted their measurements to stars with angular diameters greater than 0.02 arcsec. Altogether they measured 6 stars, all of which were giants or supergiants because the resolving power of their interferometer was not sufficiently high to measure any common or main-sequence stars.

Following this work a determined effort was made by Hale and Pease to extend the measurements to fainter stars by building a larger instrument with a baseline of 50 ft. This larger instrument was completed but never worked satisfactorily. The difficulties appear to have been two-fold; firstly there were the mechanical difficulties of making the instrument sufficiently rigid and of guiding it precisely; secondly it was difficult to see the interference fringes by eye, let alone to measure them accurately, as they danced about under the influence of atmospheric scintillation. The whole programme was abandoned in 1937.

3. The stellar intensity interferometer at Narrabri observatory

The next successful attempt to measure the angular size of a star was made at Narrabri Observatory in New South Wales (Australia) between 1962 and 1972. The instrument (Hanbury Brown 1974)—an intensity interferometer—was based on a novel principle. It measured the correlation between the fluctuations in the output currents of two separated photoelectric detectors, one at each end of the baseline. These detectors were mounted at the focus of very large (6.7 m diameter) reflectors whose separation could be varied up to a maximum of 188 m. The instrument was capable of resolving angles of 2×10^{-4} arcsec and the faintest star which could be measured had a magnitude of + 2.5.

An intensity interferometer has the interesting and valuable property that the precision with which the paths in its two arms must be equalised is a function of the electrical bandwidth of the fluctuations and not, as in Michelson's interferometer, of the optical bandwidth. For example, the electrical bandwidth of the instrument at Narrabri was 100 MHz and therefore the two paths had only to be equalized with a precision of about 10 cm. This has two practical consequences, firstly it is comparatively simple to construct a very large instrument which will meet this tolerance; secondly atmospheric scintillation cannot affect the measurements significantly because the fluctuations in the pathlength of the starlight which they introduce are very much less than 10 cm. But we must pay heavily for these advantages by a loss of sensitivity; an intensity interferometer needs an enormous lot of light and is therefore limited to measuring bright stars.

In a programme lasting about 10 years the interferometer at Narrabri measured the angular diameters of 32 single stars in the spectral range O5 to F8. Several of these 32 stars are main sequence stars and are the first main sequence stars ever to be measured.

The measurements made at Narrabri were combined with photometric measurements of the flux f_λ to find the effective temperatures T_e of these 32 stars using the relations given by Equations (1) and (2). For this purpose the ultra-violet fluxes in the range 110–330 nm were measured by the Orbiting Astronomical Observatory (OAO-2) in collaboration with the University of Wisconsin and the longer wave fluxes were measured on the ground using conventional spectrophotometry. The results gave the first temperature scale for hot stars to be based entirely on measurements.

The interferometer at Narrabri was also used as a pilot instrument to explore and

demonstrate the possible uses of an interferometer with very high angular resolving power. However, because we were limited by its low sensitivity to measuring stars brighter than magnitude + 2.5, the number and variety of the objects on which we could work was severely limited. Nevertheless we managed to make some very interesting observations. For example, by observing α Vir (Spica) we showed how it is possible to find all the orbital parameters, and the distance, of a spectroscopic binary star. To explore the interesting, and potentially valuable, application of interferometers to the measurement of the angular size of stars in the light of narrow spectral lines, we measured the angular size of the Wolf-Rayet star γ Vel in the light of the continuum and in an emission line of ionised carbon. The results showed us that the angular size of the emission region surrounding the star is about 5 times that of the star itself. To demonstrate the application of an interferometer to the many problems of stellar rotation we measured the distortion in the shape of the rapidly rotating star α Aql (Altair). We also did a number of other experiments, including an attempt to measure the limb-darkening on α CMA (Sirius) and to observe a corona surrounding the hot star β Ori (Rigel).

We also did our best to understand as completely as possible the limitations of intensity interferometry. As one of the major advantages claimed for the technique is that the measurements are not significantly affected by atmospheric scintillation, we set out to test this claim using the light from α CMa (Sirius). We showed that the measurements of correlation were not noticeably affected by scintillation even when the star was scintillating violently at an angle of only 15 deg above the horizon.

When we had finished the programme of observing stars brighter than magnitude + 2.5 the intensity interferometer was dismantled and the observatory at Narrabri was closed. Regrettably it was not possible to modify the instrument to improve its performance by an amount which would have made the cost and effort worthwhile, and we needed all our resources to develop a new instrument. Too many observatories continue to exist largely because they are already in existence!

4. The next step

Long before the programme at Narrabri was finished we had started to think about the next step. First we made a detailed study of several possible astronomical programmes and reached the conclusion that—for stellar astronomy—any new instrument should be built to reach stars of magnitude +9 with an angular resolving power of about 10^{-5} arcsec. As it was obviously impossible to modify the existing instrument at Narrabri—even to approach the performance which we wanted—we designed a completely new intensity interferometer, making it as large as we thought anyone who was likely to finance it could afford.

The layout was radically different from that used at Narrabri. Four fiat coelostat mirrors were mounted on a straight railway track and reflected the light from the star into four fixed paraboloids each with a diameter of 15 m. The instrument was designed to operate simultaneously in 10 separate optical bands and the overall electrical bandwidth of the electronic correlator and phototubes was 1000 MHz. Based on our experience at Narrabri we estimated that it would reach stars of magnitude + 7.3 in an exposure of 100 h. Such an instrument would have cost about $A3 m to build (in 1972

dollars) and we could see no way of increasing its sensitivity to approach our ideal of + 9 without increasing its cost unreasonably.

There is little doubt that, had we built this larger intensity interferometer, it would have made and would still be making, a substantial contribution to stellar astronomy. Nevertheless it would have been very large and expensive and would not have reached the sensitivity that we really wanted. And so, before committing our small group to the many years' work which it would have taken to build such an instrument, we set out to find out whether there was a better way of doing the job.

At that time there were three contemporary developments which made us think. A small double-star 'Michelson' interferometer, which used 'active optics' to minimize the effects of atmospheric scintillation, was being developed by Richard Twiss at Monteporzio in Italy (Tango 1979). A 'speckle interferometer' was being developed by Antoine Labeyrie (1978) in France. The technique of using the Moon to occult visible stars was being developed by David Evans (1957) and his colleagues in the USA. We looked carefully at all these things and came to the conclusion that, although speckle interferometry is extremely interesting and offers superior sensitivity, and lunar occultation offers superior economy, neither of these techniques looked to us to be promising ways of making measurements with the high accuracy which we were seeking for our programme of stellar astronomy. We already know from our experience at Narrabri that the answers to many of the interesting questions about stars call for observational data of high precision; observations with an uncertainty of 10 per cent are of limited use, one really needs to achieve an accuracy of 1 or 2 per cent.

To cut a long story short, we decided that the most promising possibility was to modernise Michelson's stellar interferometer. In theory it offers a higher sensitivity than an intensity interferometer and it looked to us as though it should be significantly cheaper to build. The major uncertainty is, of course, whether or not it is possible to overcome the effects of atmospheric scintillation and the need for very high mechanical precision. As far as we could estimate it should be possible to overcome both these difficulties, at least to an adequate extent, by the use of some of the modern techniques which were not available to Michelson, such as narrow-band optical filters, photo-electric detectors, 'active optics', laser distance-measuring equipment and so on. But there was, of course, only one way to find out and that was to build an experimental model. And so, rather sadly, we put the designs of a larger intensity interferometer on one side and started to build a modernized version of Michelson's stellar interferometer.

5. A modernized version of Michelson's stellar interferometer

As a first step we are building, and have almost completed, a small, experimental, interferometer in the grounds of the National Measurement Laboratory in Sydney (Davis 1979). The general layout is shown in Fig. 1. All the components are mounted on reinforced concrete plinths which are anchored in a monolithic layer of sandstone about 1 m below the surface of the ground. The mirrors which collect the light from the star are two small coelostats (C) (150-mm zerodur fiats) which are mounted on concrete plinths 1.35 m high and are separated by 11.4 m in a north-south direction. These coelostats are directed at the star by a computer which is corrected by a photoelectric star-guiding system. They reflect the starlight, via periscopes, into pipes which carry it

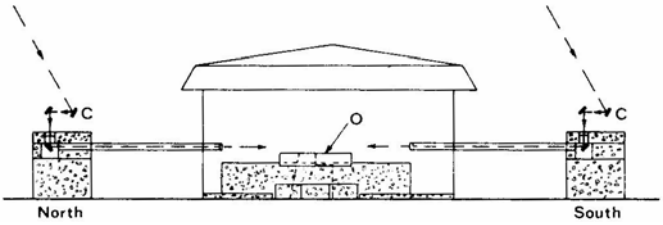


Figure 1. Cross-section of the pilot model of a modernised Michelson interferometer looking east.

to a central laboratory. The pipes are thermally insulated and can, if necessary, be evacuated.

In the central laboratory there is an optical table (O) mounted on a large concrete block. This table carries the optical system outlined in Fig. 2. The incoming beams of light from the coelostats are first reduced in diameter by a factor of 2.5 by the beam-reducing telescopes (BRT) which consist of two off-axis paraboloidal segments. The reduced beams then pass into the optical path compensators (OPLC) which consist of the retro-reflectors (R1, R2). These retro-reflectors are mounted on a very precise track and move under the control of a computer and a fringe-counting interferometer using a laser so that the pathlengths in the two arms of the equipment are equalised to a few microns. The beams then pass through the beamsplitters (G) which reflect roughly 5 per cent of the light into a lens which forms an image of the star on the quadrant detector (Q_g). Error signals from this quadrant detector are used to correct the pointing of the coelostats with a time-constant of several seconds, so that the beams are accurately

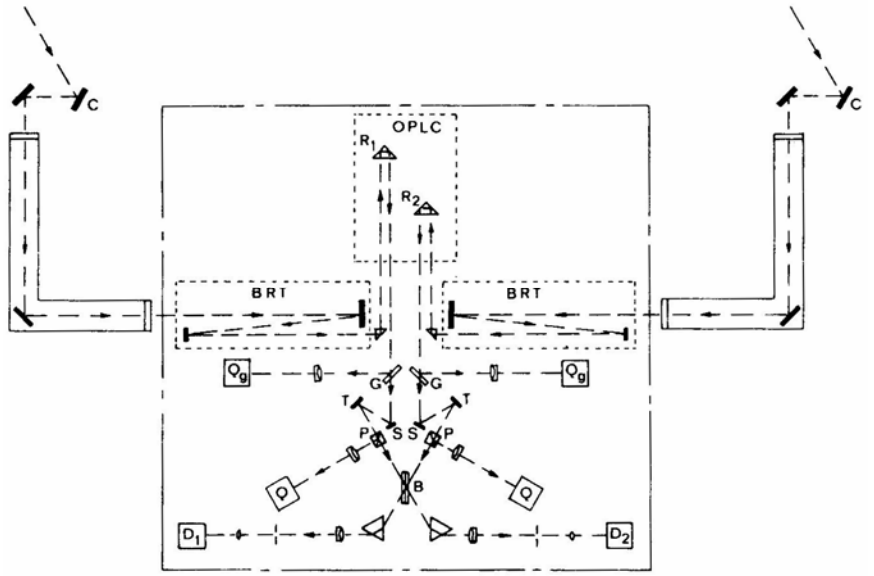


Figure 2. Outline of the optical system of the pilot model of a modernised Michelson interferometer.

aligned with the optical axis of the instrument except, of course, for the rapid angular variations of 1 or 2 arcsec due to atmospheric scintillation.

The beams are then reflected by the mirrors M and T into the polarizing beamsplitters P. The mirrors M are each mounted on a single cylinder of piezo-electric material so that, as discussed later, the relative phase of the two light beams can be varied at will. The 'active' mirrors T are each mounted on three piezo-electric cylinders in such a way that they can be tilted by electrical signals in any desired direction. The polarised beam-splitters P take all the light in one polarization and reflect it, *via* lenses, to the quadrant detectors Q. The error signals from these two quadrant detectors are then used to control the 'active' mirrors T so that they cancel, as far as possible, the angular scintillations in the two beams due to atmospheric scintillation. For a loss of 1 per cent in the measured fringe visibility the angular scintillation must be reduced to less than about 0.1 arcsec.

After passing through the polarizing beam-splitters P, the two beams—now polarized in only one plane—are incident on the neutral beam-splitter B where they interfere at zero angle, thereby forming sum and difference beams. These two beams then pass through matched prism spectrometers which transmit the light in a narrow optical band to the photon-counting detectors D1, D2. The spectral bandwidth can be varied over the range 0.1 to 1 nm and the sampling time of the two detectors is expected to lie in the range 1 to 10 ms.

In operation the two photon detectors (D1, D2) will measure the number of photoelectrons $n_1(\tau)$, $n_2(\tau)$ received in the short time interval τ(1–10 ms). A computer will then calculate the quantity,

$$\sum_{M} \{[n_1(\tau) - n_2(\tau)]^2 - [n_1(\tau) + n_2(\tau)]\} = |\Gamma|^2 \langle \cos^2 \phi \rangle \tag{3}$$

where, after normalization, $|\Gamma|$ is the modulus of the fringe visibility and ϕ is the relative phase of the starlight at the two coelostats. Assuming that this relative phase is random, then in a large number of samples we can replace $\langle \cos^2 \phi \rangle$ by 0.5. Alternatively we can, if necessary, drive the two mirrors M in such a way as to ensure that any effect of the relative phase of the light at the two coelostats on the measured value of $|\Gamma|$ is negligible.

6. How well do we expect this new instrument to work?

6.1 *The Need for Mechanical Precision and Rigidity*

The precision with which we must match the paths of light in the two arms of the instrument depends upon the optical bandwidth which we choose and that, in turn, governs the sensitivity. For a rectangular bandwidth $\Delta\nu$ and a path difference Δl, the loss of fringe visibility is given by,

$$\sin(\pi\Delta\nu\,\Delta l/c)/(\pi\Delta\nu\,\Delta l/c). \tag{4}$$

If therefore we wish to limit the loss of fringe visibility to 1 per cent then, at a wavelength of 400 nm $\Delta l \nmid 2 \times 10^7/\Delta\nu$ and, for a bandwidth of 2 nm, we must keep any path difference between the two arms to less than about 100 μm. Such differential pathlengths may be due to a whole host of factors, such as thermal expansion of the

instrument itself, earth movements and so on. We believe that they can be reduced to well below the required level by making the instrument mechanically symmetrical, by controlling its temperature, by mounting it on solid rock, by transmitting the light through evacuated pipes, by monitoring the pathlengths in the instrument with auxiliary interferometers and by calibrating it frequently on bright stars.

A second, perhaps more demanding requirement, is that the instrument should be sufficiently rigid so that the relative phase of the light in the two arms should not change during the sampling time τ of a single observation. This implies that any vibration in the instrument should not change the optical paths by more than a small fraction of the wavelength of light ($\lambda/40$) in a time of about 1 ms. We believe that this requirement can be met by making the coelostats and their associated mirrors massive, by shielding them from the wind and by choosing a suitable site perhaps on solid rock.

6.2 *The Effects of the Atmosphere*

Let us look first at the loss of fringe visibility caused by random fluctuations in the relative time of arrival of the light at the two coelostats. Theory suggests that, for a baseline of a few metres, these fluctuations are largely uncorrelated at the two coelostats and that they are given by,

$$\Delta l_{\mathrm{rms}} = 0.4\lambda\,(d/r_0)^{5/6} \tag{5}$$

where Δl is the fluctuating pathlength, d is the baseline and r_0 is the characteristic length of the scintillations. Taking a typical value of r_0 = 10 cm then, very approximately,

$$\Delta l \simeq 2 \times 10^{-6}\, d$$

and it follows that for an optical bandwidth of 2 nm, and a loss of fringe visibility of 1 per cent, we can use baselines of up to 50 m. At longer baselines it would be necessary either to restrict the optical bandwith, thereby losing sensitivity, or to develop a system of compensating automatically for the varying delay. A preliminary analysis suggests that it should be possible to do this by tracking the 'white-light fringes' and that such a system would have an adequate signal-to-noise ratio, although it must be remembered that the magnitude and time variation of these delays at optical wavelengths have never been measured at baselines greater than a few metres.

There are also the spatial fluctuations in the phase and amplitude of the starlight. In general, the wavefront of the light arriving at the coelostats will not be plane nor will it be normal to the direction of the star; it will be tilted and curved and the relative phase and amplitude of the waves at the two coelostats will vary rapidly and at random.

There will be a loss of fringe visibility if the relative phase of the light at the two coelostats varies significantly during a sampling interval τ. If this loss is not to exceed 1 per cent then any variation in the relative phase must not exceed about 10 deg in a sampling interval. It follows that, for wind speeds of a few metres per second and a typical scintillation scale of 10 cm, the sampling time cannot be greater than a few milliseconds. As far as the fluctuations in intensity are concerned, it can be shown that their effect on the measurements can be removed entirely if each elementary observation is normalised by the total number of photons counted in that interval.

The next effect which we must consider is that of the fluctuations in the tilt of the wavefront or, in other words, in the apparent direction of the incoming light from the star. If this tilt is not corrected the two beams of light will not interfere at zero angle in

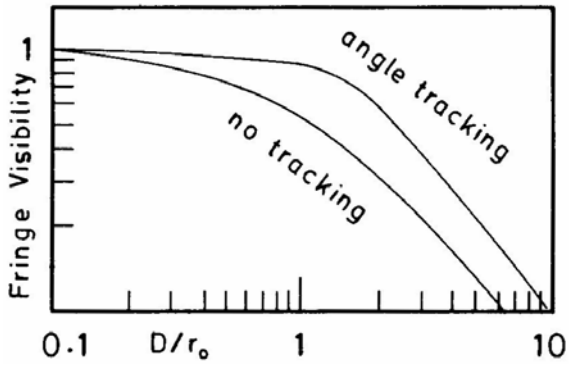


Figure 3. The theoretical loss of fringe visibility as a function of the ratio of the mirror diameter D to the size of the atmospheric scintillations r_0.

the beam-splitter B and the measured value of $|\Gamma|$ will be reduced. Given a theoretical model of the atmospheric fluctuations this loss can be calculated as a function of the ratio of the diameter of the mirror D to the characteristic size of the scintillations r_0. The curve marked 'no tracking' in Fig. 3 represents the results of one such calculation and shows how vulnerable this type of interferometer is to atmospheric scintillation.

To a large extent this alarming loss of fringe visibility can, so we expect, be reduced by the use of 'active optics'. As we have seen in Fig. 2 the two small mirrors T are servo-controlled to maintain the two beams of light at the correct angle relative to the optical axis within a fraction of an arcsec, and to control these mirrors we have taken all the light in one polarization and focussed it on the quadrant detectors. Such a scheme removes the average tilt of the two beams and should greatly reduce the loss of fringe visibility. This is shown by the curve marked 'angle-tracking' in Fig. 3 which has been calculated on the assumption that the average tilt of the wavefront has been reduced to zero. The remaining loss of fringe visibility, which increases with D/r_0, is due to the curvature of the wavefront.

The curves in Fig. 3 allow us to choose the size of the coelostat mirrors. They show clearly that if we wish to restrict the loss of fringe visibility due to curvature of the wavefront to 1 per cent then, without 'angle-tracking', we are limited to the use of very small mirrors indeed, and therefore to an instrument of low sensitivity. For example, if we take a typical value of r_0 = 10 cm, then without angle-tracking our mirrors must not be significantly larger than 0.1 r_0 in diameter, which means that they must be absurdly small. If on the other hand we use angle-tracking then, for a 1 per cent loss in fringe visibility, the size of the mirrors can be increased to about 0.25 r_0 or 2.5 cm. Even so, the sensitivity of the instrument would be too low and for that reason we propose to make the (projected) diameter of the coelostats mirrors about 10 cm and, as discussed later, to correct for the loss of fringe visibility by using a value of r_0 measured continuously by an independent interferometer. In this way we believe that it will be possible to correct for a loss of fringe visibility of the order of 10 per cent with adequate precision, thereby making it possible to use coelostats of at least 10 cm diameter most of the time.

6.3 *The Sensitivity*

One of the main attractions of a Michelson interferometer is that, compared with an intensity interferometer, it is more sensitive. For the instrument shown in Fig. 2 it can

be shown that at low photon rates the signal to noise ratio is given by,

$$S/N_{\mathrm{rms}} = n|\,\Gamma\,|^2\,(\tau T_0/2)^{1/2} \tag{6}$$

where τ is the sampling time, T_0 is the total time of observation and n is the mean counting rate of photoelectrons in one channel. If we take the diameter of the coelostats as 10 cm, τ = 2 ms, the optical bandwidth as 2 nm at a mean wavelength of 550 nm, the overall transmission of the atmosphere and the optics as 0.35, the quantum efficiency of the photodetectors as 0.2, and if we assume that the lowest signal-to-noise ratio with which we can usefully work is 10/1 in one hour, then Table 1 shows the limiting magnitude for an A0 star at the zenith. Column 2 has been calculated for the instrument outlined in Fig. 2, which has only one pair of photon detectors, and shows that the limiting magnitude for that simple configuration is only + 7.6 which falls significantly short of our target of + 9. It should, however, be comparatively easy to increase the sensitivity by adding several pairs of detectors in separate spectral channels and column 3 shows that, by using only 10 separate channels, we should be able to reach stars of magnitude + 9.

It is, however, by no means certain that the sensitivity of the instrument would be limited by the signal-to-noise ratio in the photon-counting process. It seems more likely that it would be limited in practice by the signal-to-noise ratio in the angle-tracking system. The 'signal' in that system corresponds to the zero order or average tilt of the incoming light, and the 'noise' to a combination of the higher order components of the curvature of the wavefront and the statistical noise in the photoelectron stream at the output of the quadrant detectors. Inevitably this 'noise' introduces unwanted 'dither' of the tilting mirrors and there is a corresponding loss of fringe visibility. Tango–& Twiss (1980) have made a detailed analysis of this loss and have shown that it will be about 1 per cent for stars of magnitudes + 5 and will increase to about 10 per cent for stars of magnitude + 8.

In principle it should be possible to correct for this loss by measuring the residual fluctuations in the incoming beams after they have been reflected from the tilting mirrors, but how accurately this can be done remains to be found out by experiment. In the meantime it looks as though the sensitivity of the interferometer may be limited, perhaps to stars of about magnitude + 8, by the angle-tracking system.

6.4 *The Precision of the Results*

As we have seen the measured fringe visibility will be reduced by two effects, curvature of the wavefront and angular noise in the angle-tracking system. The final precision and reliability of the results will depend on how accurately we can correct for the losses due to these two effects.

Table 1. The limiting magnitude of a modernized Michelson stellar interferometer for a signal/noise ratio of 10/1 in 1 h.

Optical bandwidth of each channel $\Delta\nu$	Limiting magnitude	
	1 optical channel	10 optical channels
2 nm	+7.6	+8.9
10 nm	+9.4	+10.6

The first effect, the loss due to curvature of the wavefront, is a function of the ratio of the characteristic size of the scintillations r_0 to the diameter D of the coelostats. In principle it should be possible to correct for this loss if we know r_0 and, to that end, we are building a small auxiliary interferometer to measure r_0 continuously. The curves in Fig. 3 suggest that, if we restrict our observations to atmospheric conditions when $D/r_0 < 1$, then we shall be able to reduce the uncertainty in our measures of fringe visibility to 1 per cent if we can measure r_0 with an uncertainty of about 5 per cent:

The second effect, the loss of fringe visibility due to noise in the angle-tracking system, is expected to be significant only for the fainter stars and may, as noted above, prove to be the principal factor which determines the limiting magnitude. To what extent it can be corrected remains to be determined by experiment.

To sum up, our preliminary analysis suggests that the instrument which we are building will measure fringe visibilities with an uncertainty of a few per cent (less than 5 per cent) for stars brighter than about magnitude + 6. For fainter stars this uncertainty will increase and may reach unacceptable levels for stars of magnitude + 8 or + 9.

7. Conclusion

The observing programme of the Stellar Intensity Interferometer at Narrabri Observatory was restricted to stars brighter than magnitude +2.5 and was completed successfully in 1972. Since then we have been trying to design a more sensitive instrument to carry on and extend this work to fainter stars.

As a first step towards this goal we designed an intensity interferometer with a sensitivity about 100 times greater (limiting magnitude + 7.3) and a resolving power 10 times greater (baseline 2 km) than that of the original instrument at Narrabri. It was both large and costly ($3 m in 1972) and we could see no way of increasing its sensitivity to approach our target of + 9 without unreasonable expense. The only thing which can be said in its favour is that we were confident that it would work!

Before committing ourselves to building such a large instrument we looked carefully at all the possible alternatives and decided to try to improve Michelson's stellar interferometer. We have designed a modernized version of Michelson's interferometer which, so we hope, will be able to measure bright stars with a precision of a few per cent even in the presence of atmospheric scintillation. If all goes well it should reach stars of magnitude +8. We are confident that, if it can be made to work satisfactorily at short baselines, it can be extended to long baselines by the development of an automatic fringe-tracking system. The cost of this instrument will surely be significantly less than that of an intensity interferometer of comparable performance.

Our pilot model of this new interferometer is almost complete and it will be tested during 1984. If these tests are successful we intend to build a full-scale instrument to carry on and extend the work which was started at Narrabri.

References

Davis, J.1979, in *IAU Coll.* 50*: High Angular Resolution Stellar Interferometry*, Eds J. Davis and W. J. Tango, Chatterton Astronomy Department, University of Sydney, p. 14.1.

Evans, D. S. 1957, *Astr. J.*, **62,** 83.

Hanbury Brown, R. 1974, *The Intensity Interferometer*, Taylor & Francis, London.
Labeyrie, A. 1978, *A. Rev. Astr. Astrophys*., **16,** 77.
Tango, W. J. 1979, in *IAU Coll.* 50*: High Angular Resolution Stellar Interferometry*, Eds J. Davis and W. J. Tango, Chatterton Astronomy Department, University of Sydney, p. 13.1.
Tango, W. J., Twiss, R. Q. 1980, in *Progress in Optics XVII*, Ed. E. Wolf, North-Holland, Amsterdam, p. 240.

J. Astrophys. Astr. (1984) **5,** 31–41

Distribution of Quasars on the Sky

Halton Arp *Mount Wilson and Las Campanas Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 9110-1292, U.S.A.*

Abstract. It is shown that high-redshift quasars of bright apparent magnitude are concentrated in the direction of the centre of the Local Group of galaxies. A number of them are distributed along a line originating from the Local Group companion galaxy, M 33. A similar, but shorter and fainter line of quasars is seen emanating from the spiral galaxy NGC 300 in the next nearest, Sculptor Group of galaxies.

The concentration of bright quasars in the Local Group direction is supported by bright radio sources catalogued in high-frequency surveys. One of the consequences of this large-scale inhomogeneity is to explain the different gradient of radio source counts in the direction of the Local Supercluster, a result discovered in 1978 but never investigated further.

Previously reported homogeneity and isotropy of radio-source counts over the sky would seem to be an effect of integrating nearby, large-scale groupings with more distant, smaller-scale groupings over different directions in the sky. More careful analyses as a function of flux strength and spectral index on various scales over the sky are now required. Previous conclusions about radio source and quasar luminosity and number evolution drawn from log N–log S counts would then need to be re-evaluated.

Key words: quasars, alignment—galaxies, Local Group, Sculptor Group—radio sources, counts

1. Introduction

Evidence that quasars are not at the great distances which conventional interpretations of their redshifts would require was first put forward eighteen years ago (Arp 1966). The evidence consisted of quasars falling closer to low-redshift galaxies than expected by chance, alignments with active galaxies, and associations with radio sources which are aligned with galaxies (Arp 1967, 1970). One consequence of associating quasars with nearby galaxies (galaxies with redshifts of the order of $v \simeq 1000$ km s^{-1}) was that many quasars were seen between galaxies; in areas where they seemed unassociated. This required that quasars must be distributed over a large angular extent of the sky around each galaxy of association. There needed to be field regions in which adjoining associations overlapped. Of course it was clear that very nearby galaxies would then have associated quasars which would appear projected over very large regions of the sky. Since there were far fewer very nearby galaxies, however, it was anticipated that they would only contribute small numbers of brighter quasars to the observed population.

The clue that these very nearby quasars were more important than originally perceived came from quasar surveys which showed groupings of quasars which favoured certain redshifts. Surveys which searched for ultraviolet excess candidates to faint apparent magnitudes over moderately large areas of the sky (Arp, Sulentic & di Tullio 1979; Arp & Hazard 1980; Surdej *et al.* 1983) showed some small groupings of $z \simeq 1$ quasars (Arp 1983b). The implication was that these quasars were relatively distant and that therefore the $z \simeq 1$ quasars were intrinsically more luminous than quasars of other redshifts. This conclusion was supported when tested on a group of quasars which had been identified as belonging to the Local Supercluster (Arp 1970). A plot of these quasars in the redshift-apparent-magnitude diagram revealed a preponderance of quasars near $z \simeq 1$, with quasars of higher and lower redshift having systematically fainter apparent magnitude (for analysis see Arp 1983b). If any quasars of redshift $z \sim 2$ were present in the Local Supercluster, they were apparently too faint to be seen. The crucial question then presented itself in the following form: If the quasars of redshift $z = 2$ are the least luminous, then those of the brightest apparent magnitude should be, of all quasars, the ones closest to us in space. Where were they located in the sky? The answer to that question proved to be quite stunning. First of all, it appeared that the quasars of $z \sim 2$ were concentrated toward one area of the sky. Secondly, that area was the direction of the centre of the Local Group of galaxies.

Since then, investigations of quasar groupings from this new perspective has led to some interesting results which are described in the present paper. In turn these results suggest new lines of analysis which may be able to promote further understanding of the spatial distribution of different kinds of quasars and radio sources.

2. The concentration of quasars in the Local Group

A striking example of an inhomogeneous distribution of a particular kind of quasar can be seen by noting that the distribution of radio quasars in the general direction of the Local Supercluster (9^h < R.A. < 15^h) shows a marked sparsity of quasars with $z \sim 2$. In contrast there is a strong concentration (by about a factor 3.5) of these kinds of quasars in the direction of the centre of the Local Group of galaxies (21^h < R.A. < 3^h). The plots are shown in Fig. 20 of Arp (1983b). This can hardly be a selection effect because all these quasars are from complete radio surveys of strong sources such as 3C and Parkes which are uniform around the sky in the declination zones involved. (Additional discussion of this point will be made in Section 4).

But if we look within the region of greatest concentration in the direction of the Local Group, we see a distinct line of high-redshift quasars extending from the region of the Local Group companion galaxy, M33. This line was first shown in Figs 21 and 22 of Arp (1983b), then in Arp (1984a, b). In Fig. 1 of the present paper, however, the line is shown at its most conspicuous because only those high-redshift quasars with radio fluxes between $0.3 \leqslant S_{11} \leqslant 1.0$ are plotted.

The enhancement of this line is an important feature because if there were a physically associated group of quasars we would expect these quasars to be distinguished from others in the area by characteristic values of parameters such as radio strength. On the other hand, if they were not physically associated there would be no particular parameters which should artificially create a line.

The statistical significance of the line needs to be tested taking into account the

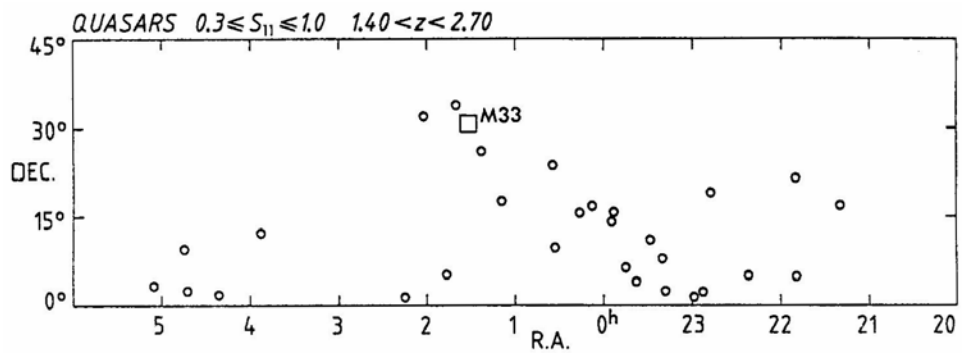


Figure 1. All radio quasars between the indicated limits of flux strength and redshift are plotted in the direction of the Local Cluster of galaxies. Data is from quasar catalogue of Veron-Cetty & Veron (1984).

effective boundaries of the region and the curvature of straight lines projected onto the sky in these coordinates. Such tests are being carried out by J. V. Narlikar and his associates (personal communication). Pending those results I will take the standpoint that the line is *prima facie* significant. Some questions that naturally arise about the line, I will try to comment on at this point.

2.1 *Comments on the Line*

Why a line and why M33? From the earliest associations of quasars with low-redshift galaxies the conclusion was that the quasars, like some other radio sources, were ejected out on either side of active galaxies. Evidence throughout the years has tended to empirically support the association of quasars in lines and pairs across the central galaxy (Arp 1980, a). Furthermore, by about 1975 (Arp, Baldwin & Wampler 1975), it began to become apparent that while quasars could be associated with a range of morphological types of galaxies, there was a distinct preference for them to be associated with galaxies which were companions in groups (for latest summary see Arp 1983a). M33, of course, is the most conspicuous companion to M31, the dominant member of the Local Group of galaxies. Finding quasar associations with this nearby companion galaxy fulfilled the predictions of the earlier work. Moreover, since our own Milky Way galaxy is also a companion in the Local Group, we would predict that the nearest quasars of all would belong to our own galaxy (see Arp 1984b).

Another aspect of the line as shown in Fig. 1 and perhaps even better in Fig. 2, is that toward the southwest end the line becomes broader and the quasars tend to become brighter and more radio-strong. This implies that the line of quasars to the SW of M33 is oriented at least somewhat toward us, reaches an appreciable portion of the distance toward us, and this to some extent accounts for the broadening of the line toward the SW end. Of course, the line of quasars does appear to extend on the other side of M33 to the NE. But it is difficult to be certain of the nature of the line in this direction for two reasons: (1) The galactic latitude N of M33 is becoming quite low and it is unclear how completely even strong radio quasars are known at very low galactic latitudes. (2) The Perseus cluster of galaxies is NE of M33 in the general direction of the line, and just on the basis of radio sources, may become entangled.

Finally, we can comment that some quasars fall off the line to the east of the line.

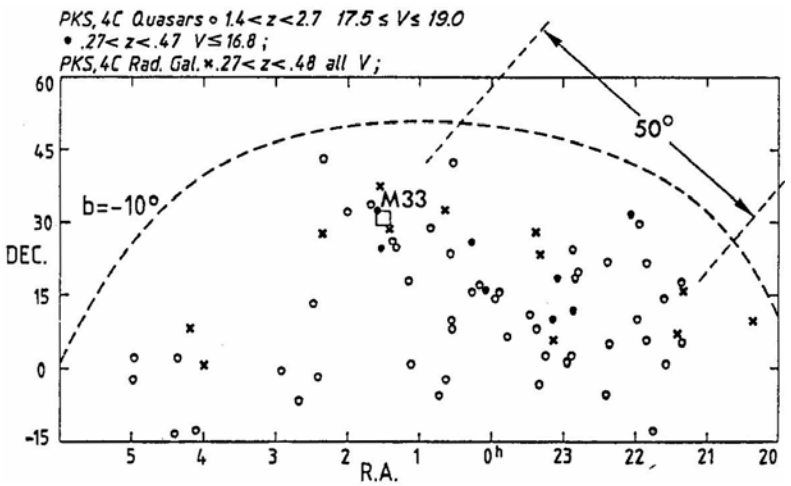


Figure 2. Radio quasars between the indicated limits of redshift and apparent magnitude. All quasars from the catalogue of Hewitt & Burbidge (1980). Open circles are high-redshift quasars as in Fig. 1. Filled circles are quasars of $0.27 < z < 0.47$ and crosses are radio galaxies in the same redshift range.

From a study of quasars and radio sources of all types which fall in this region it is my opinion that there are two subconcentrations of quasars; one associated with NGC 520 ($1^h22^m + 3°32'$) and a line to the SW (see Arp 1983b), and the other with 3C120 ($4^h + 5°14'$). This identification would probably require that both NGC 520 and 3C120 be peculiar, anomalously high-redshifted members of the Local Group.

2.2 *Concentrations of Other Kinds of Quasars toward the Local Group Centre*

It was noted in the Introduction that quasars of redshifts lower than the $1.4 < x < 2.7$ ones we have been considering turn out to have generally higher intrinsic luminosities. If we include, therefore, brighter quasars of lower redshift we should obtain a sample containing the remaining quasars in the Local Group. As Fig. 2 shows, these quasars, which are like 3C48, fall in a line extending SW from M33. This line may be rotated slightly anticlockwise from the line of the high-redshift quasars but it is close enough to give very good confirmation of that line which appeared in Fig. 1. Similarly, the radio galaxies which have the same redshifts as the low-redshift quasars (crosses in Fig. 2), corroborate the low-redshift quasar line.

One very strong conclusion we can come to from Fig. 2 is that even if the quasars were not associated with M33, this strong concentration of quasars of widely different redshifts in one area of the sky would by itself rule out the cosmological interpretation of the quasar redshifts. This is because if the redshifts were to be interpreted as distance indicators, we would have a long tube of quasars pointing at the observer.

3. High-redshift quasars in the region of the Sculptor Group of galaxies

Objective prism searches are particularly effective at discovering quasars with $z > 1.8$ because the Lyman-alpha emission line, the strongest in the quasar spectrum, comes

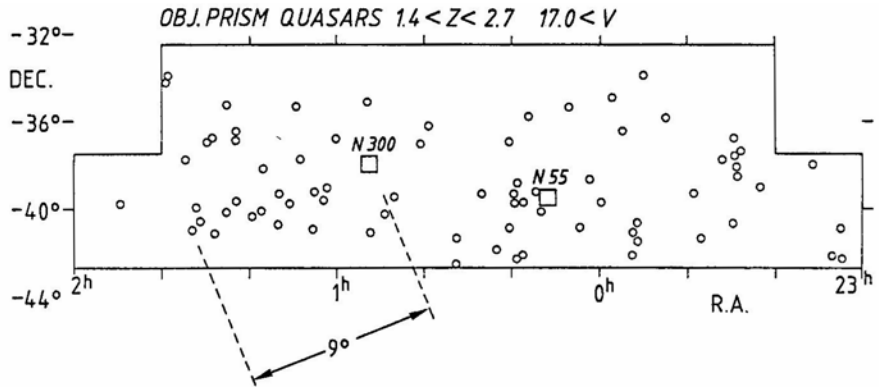


Figure 3. All quasars discovered by objective prism search techniques in and around the region of the Sculptor Group of galaxies. Quasars tabulated in Osmer & Smith (1980) and in Arp (1984c). The Sculptor Group galaxies, NGC 300 and NGC 55 are marked by open boxes

into the recorded spectral window at this redshift. Objective prism searches for quasars were carried out in the Dec. = –40° strip (Osmer & Smith 1980) and in a smaller supplementary region in the Dec. = –35° zone by Arp (1984c). This region, as outlined in Fig. 3, passes over the southern part of the Sculptor Group of galaxies. NGC 253, the largest angular extent galaxy in the Group is outside the frame at Dec. = –25°. But two of the larger Sculptor Group galaxies, NGC 300 and NGC 55 are shown by open boxes in Fig. 3.

Since the Sculptor Group is only about a factor of two more distant than M33, it is interesting to test whether in this Sculptor Group of high-redshift quasars there are any similar phenomena as those observed around M33. There are! As the earliest results in this strip showed (Osmer 1981; Arp 1980b) there is a strong concentration of quasars some degrees away from NGC 300. This line SE of NGC 300 shows conspicuously in Fig. 3 where it has been estimated to have about 9° projected length on the sky. If the line of quasars from M33, estimated at ~ 50° on the sky from Fig. 2, were moved to twice the distance, in the Sculptor Group, it would appear ~ 25° long. Considering the arbitrary angle of orientation which these lines can have toward the observer, the agreement between the projected length of the M33 and NGC 300 line of quasars is quite satisfactory.

It is of interest then to compare the apparent magnitudes of the quasars in these two lines. Fig. 4 does this. It is seen that the quasars in the M33 line are about $1\frac{1}{2}$ magnitudes brighter than those in the NGC 300 line. This is just the amount expected from the relative distances of the two galaxies.

Several comments could be made about Fig. 4. One is that the M33 quasars are radio quasars that come from larger areas of the sky than the objective prism quasars. This is a selection in the direction of brighter quasars. Also, the radio quasars are given in broad-band magnitudes—not in continuum magnitudes which are somewhat fainter for an individual quasar (Arp 1983a). On the other hand, the NGC 300 quasars seem to be still increasing in number toward the plate limit which is usually about 19.5 mag for these objective prism plates. So, the average apparent magnitude could be fainter than indicated in Fig. 4. In summary it seems fair to say about Fig. 4 that although there are

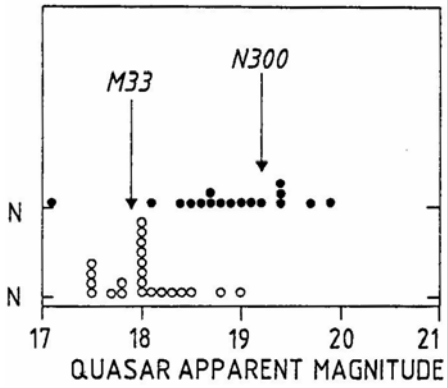


Figure 4. Distribution of apparent magnitudes in the M33 and NGC 300 lines of quasars. M33 magnitudes from Hewitt & Burbidge (1980), NGC 300 magnitudes from Osmer & Smith (1980).

these factors of uncertainty, if the two sets of quasars differed in distance by a factor of two then the average magnitude difference would be close to the one observed.

An opportunity to confirm the physical reality of the line of quasars coming SE from NGC 300 comes from H I observations of NGC 300. Matthewson, Cleary & Murray (1975) discuss an apparent extension of hydrogen from NGC 300 in a direction which coincides with the line of quasars (see discussion in Arp 1980b, p. 467). A similar, extended cloud of H I (Wright 1974) lies SW of M33 in the approximate direction of the line of quasars discussed in this paper.

It is also of considerable importance to note in Fig. 3 that there is a quite plausible line of quasars extending to the NW of NGC 300, marking a possible extension of the quasar line to the other side of NGC 300. Also of importance in Fig. 3 is the obvious grouping of high-redshift quasars around NGC 55. This grouping could also be part of line extending away to the SW and on the other side in a direction which would take it just north of NGC 300. We must also remember that NGC 253 is a large, active galaxy (Ulrich 1978) just north of the frame in Fig. 3 and that there are other companion galaxies in the Sculptor Group. The point is that if companion galaxies generally have lines of quasars coming from them they will, in many groups, appear to intersect and become confused. Only the cases of strong concentrations and favourable orientation may, if we are fortunate, be clearly identifiable.

4. The overall asymmetry between the Local Group and Local Supercluster direction

In the Introduction it was mentioned that there is a strong factor of 3 to 4 times as many high-redshift quasars in the Local Group direction than there is in the Local Supercluster direction. This difference was noted as long ago as 1966 by Strittmatter, Faulkner & Walmsley in the sense that they reported that the high-redshift quasars were distributed on the sky differently from the low-redshift quasars*. In an analysis by Arp (1984a, Fig. 3) it was shown that the kinds of quasar redshifts present in the line and region SW of M 33 were completely different from the kinds of redshifts present over

* After the writing of this paper the article by Shastri & Gopal-Krishna (1983) appeared. They independently report an inhomogeneous distribution over the sky of quasars with $z > 2$.

the R.A. = 12^h region. This result is so devasting for the cosmological interpretation of quasar redshifts that I expect that it will be attacked intensely. Therefore, I would like to discuss here some of the possible criticisms of this result.

First of all, the quasars which establish this asymmetry are all radio quasars from 3C and Parkes radio surveys. The 4C survey quasars also support the result. At any given declination zone these surveys are complete in right ascension. Therefore there should be no reason why radio quasar candidates in the 0^h right ascension region should be different from radio quasar candidates in the 12^h region. It might be argued that high-redshift quasars are preferentially flat spectral-index quasars (Kraus & Gearhart 1975) and that the Greenbank high-frequency survey (Pauliny-Toth *et al.* 1972) and the Ohio State Survey (Dixon & Kraus 1968; Fitch, Dixon & Kraus 1969) detected 3C and Parkes sources preferentially in the $21^h < \text{R.A.} < 4^h$ region which were then measured optically. The counterarguments to this scenario are that the Ohio State surveys observed about the same total area in the 12^h region as the 0^h region. Moreover, the University of Texas Deep Radio Survey from which the complete quasar samples of Wills & Wills (1979) were drawn, were from $03^h30^m < \text{R.A.} < 23^h30^m$, specifically excluding the 0^h region. The 3C radio sources have, of course, been exhaustively observed all over the sky (Kristian, Sandage & Katem 1974). Therefore, though there may be minor sampling inhomogeneities, the conclusion would seem to be that—on the average—3C and Parkes quasar, candidates were observed about equally around the sky.

But to make the argument completely rigorous it would be helpful to find a high-frequency radio survey which identified radio quasar candidates all around the sky. Then we could check the frequency of the quasar candidates *before* they had been measured for redshift. Fortunately, the Parkes 2700 MHz survey fulfils these conditions ideally. In a zone $4° < \text{Dec.} < 25°$ which passes through the concentration toward the Local Group centre which we have discussed in Figs 1 and 2. Shimmins, Bolton & Wall (1975) have identified all blue stellar candidates which coincide with their measured radio source positions. Fig. 5 shows a plot of these candidates all around the sky in R.A. The clear-cut result which emerges from this plot is that the 2700 MHz quasar *candidates* are just about 3 times more numerous in the previously named direction of the Local Group than they are in a comparable section in the 12^h region. This high-frequency survey therefore demonstrates a strong excess of quasars in the Local Group direction and the argument cannot be made that the two regions have equal numbers of high-frequency quasar candidates from which 0^h region candidates were favoured for optical measurement.

There is a concentration of quasar candidates between $16^h30^m > \text{R.A.} > 14^h30^m$ in Fig. 5. That region of the sky encompasses the Hercules Supercluster (Tarenghi *et al.* 1980). The Hercules cluster itself is well-known for containing numerous disrupted and active galaxies. Presumably it is more distant than the groups we have been discussing so far and we would therefore expect quasars nearer $z \simeq 1$ in redshift. A survey of flat spectrum radio quasars in the Dec. $\pm 4°$ zone shows six quasars in the range $14^h < \text{R.A.} < 16^h$ with $1.3 < z < 1.6$ (Wampler, personal communication). That there are concentrations of quasars over the sky on different scales and at different distances can hardly be doubted. What is important now is to measure systematically their redshifts and magnitudes and to try to identify where they are located in distance. Systematic measurement of all 2700 MHz quasars pictured in Fig. 5 would be a very interesting start on this problem.

Table 1. 2.7 GHz radio sources, 4° < Dec. < 25°.

R.A.	6^h	5^h	4^h	3^h	2^h	1^h	0^h	23^h	22^h	21^h	20^h
No.	7	14	4	14	11	13	11	11	6	7	
R.A.	18^h	17^h	16^h	15^h	14^h	13^h	12^h	11^h	10^h	9^h	8^h
No.	13	12	8	9	2	10	6	6	2	3	

$N(0^h) = 66$
$N(12^h) = 35$

Table 2. Bologna radio sources, 1 ≤ peak flux ≤ 14.

R.A.	6^h	5^h	4^h	3^h	2^h	1^h	0^h	23^h	22^h	21^h	20^h	
Integ. flux*	23	26	27	23	41	31	27	25	27	23		34°02′ > Dec. > 29°18′
Integ. flux*	47	25	30	34	49	26	37	25	25	30		29°30′ > Dec. > −24°02′
Total	70	51	57	57	90	57	64	50	52	53		

* Sums of integers of fluxes only

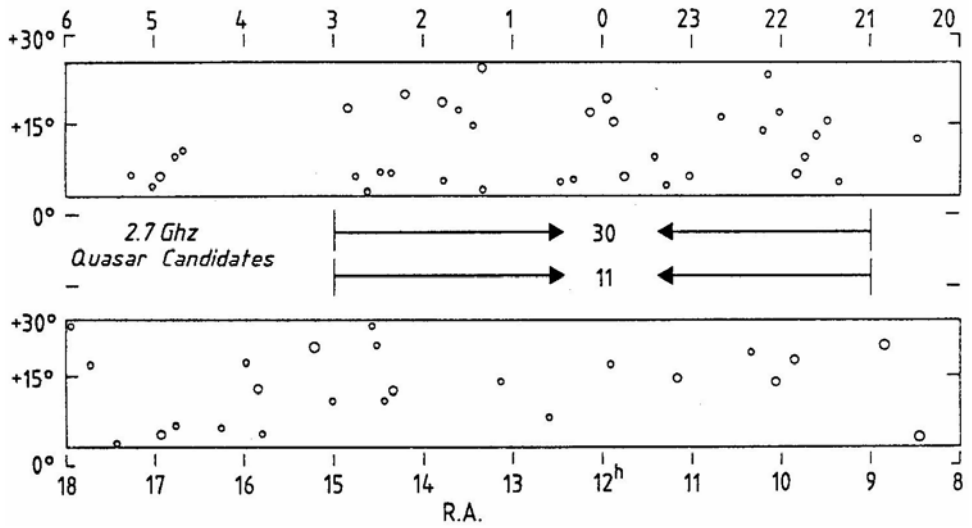


Figure 5. Plot of all quasar candidates (blue stellar objects at the position of radio source) from Parkes 2.7 GHz survey between 4° < Dec. < 25° (Shimmins, Bolton & Wall 1975). Total numbers are shown for a 6^h stretch of R.A. centred on 0^h and 12^h respectively.

5. Distribution of radio sources on the sky

The high-frequency quasar candidates pictured in Fig. 5 lead us to some interesting considerations about the distribution of radio sources. The question that is posed is whether the asymmetries and groupings shown by the radio quasars are reflected in distributions of radio sources in general.

Table 1 shows that the concentration of 2.7 GHz radio sources in the 0^h region relative to the 12^h region is present in all the radio sources taken together as well as in just the radio quasar candidates. The radio $N\ (0^h)/N\ (12^h)$ is greater for just the quasar candidates alone but because of the larger numbers involved in the total radio source count the imbalance is as significant or perhaps even more significant for the latter. In Table 2 we give the approximate integrated flux of Bologna radio sources (Colla *et al.* 1970, 1972) across the 0^h region. For all sources including the faint ones, the Bologna counts are rather level from R.A. = 20^h to 6^h. But for brighter sources, f.u. > 1, the summed flux rises significantly going across the centre of the 0^h region. Particularly across the position of M33, the total sum of bright-source flux reaches about 90, far in excess of base values on either side of the position of M 33. This is true even though we have omitted the strong radio source 3C48 at 37 Jy which is so close to M33. Previously it had been shown that the high-frequency radio sources from Galt & Kennedy (1968) (see Arp 1984b for analysis) peaked at the position of M33. All the radio surveys seem to show this peaking around M33.

Additionally, we can remark that the 5 GHz source counts between 70° > Dec. > 35° peak in 23^h < R.A. < 1^h region (Pauliny-Toth *et al.* 1978). This last point is a little difficult to interpret because so much of this region is at low galactic latitude. Taken together, however, all these sample cuts indicate a general increase in raw radio source counts as one crosses a position toward the centre of the Local Group.

Such a result confirms in a little more detail the significant difference found for the gradient of radio source counts off and on the position of the Local Supercluster (Pauliny-Toth *et al.* 1978). Careful inspection of that result shows that it was the presence of strong radio sources *away* from the Virgo region which gave the steeper gradient of log *N*-log *S* counts for the Local Supercluster. This would be interpreted, in terms of the discussion here, as due to the relatively bright radio sources contributed by the large area of the 0^h region which encompasses the Local Group direction.

The conclusion from this discussion would seem to be that there are significant groupings of radio sources in various regions of the sky. The claim that the radio sources are uniformly distributed over the sky (Webster 1977; Fanti, Lari & Olori 1978) must have come from having integrated over regions of different characteristics in different directions. The consequences of this are quite serious because the log N –log S curves which were supposed to be uniformly applicable have been used to derive conclusions about strong evolution as a function of look-back time in the universe. It would not have been very satisfactory to obtain different evolution rates in different directions in the universe.

In summary, it has been the supposed homogeneous distribution of quasars and radio sources which were used to support the interpretation of distant quasars and distant radio sources. These distributions, when looked at closely as we have started to do here, are in fact not homogeneous at all, but instead show groupings and concentrations which support specific local concentrations of radio sources and quasars. It would seem necessary now to make a re-analysis of radio-source distributions *de novo* paying close attention to distributions as a function of flux strength and testing for associations on a variety of angular scales.

References

Arp, H. 1966, *Science*, **151,** 1214.
Arp, H. 1967, *Astrophys. J.*, **148,** 321.
Arp, H. 1970, *Astr. J.*, **75,** 1.
Arp, H. 1980a, in 9*th Texas Symp. Relativistic Astrophys: Ann. N.Y. Acad. Sci.*, **336,** 94.
Arp, H. 1980b, *Astrophys. J.*, **239,** 463.
Arp, H. 1983a, *Astrophys. J.*, **271,** 479.
Arp, H. 1983b, in *Groups*, *Concentrations*, *and Associations of Quasars,* Liège Conference, June 1983, Inst. d'Astrophys, University of Liège.
Arp, H. 1984a, *Astrophys. J. Letters* (in press).
Arp, H. 1984b, *Publ. astr. Soc. Pacific* (in press).
Arp, H. 1984c, *Astrophys. J.* (submitted).
Arp, H., Baldwin, J. A., Wampler, E. J. 1975, *Astrophys. J.*, **198,** L3.
Arp, H., Sulentic, J. W., di Tullio, G. 1979, *Astrophys. J.*, **229,** 489.
Arp, H., Hazard, C. 1980, *Astrophys. J.*, **240,** 726.
Colla, G., Fanti, C., Fanti, R., Ficarra, A., Formiggini, L., Gandolfi, E., Grueff, G., Lari, C., Padrielli, L., Roffi, G., Tomasi, P., Vigotti, M. 1970, *Astr. Astrophys. Suppl. Ser.*, **1,** 281.
Colla, G., Fanti, C., Fanti, R., Ficarra, A., Formiggini, L., Gandolfi, E., Lari, C., Marano, B., Padrielli, L., Tomasi, P. 1972, *Astr. Astrophys. Suppl. Ser.*, **7,** 1.
Dixon, R. S., Kraus, J. D. 1968, *Astr. J.*, **73,** 381.
Fanti, C., Lari, C., Olori, M. C. 1978, *Astr. Astrophys*., **67,** 175.
Fitch, L. T., Dixon, R. S., Kraus, J. D. 1969, *Astr. J.*, **74,** 612.
Gait, J. A., Kennedy, J. E. D. 1968, *Astr. J.*, **73,** 135.
Hewitt, A., Burbidge, G. R. 1980, *Astrophys. J. Suppl Ser.*, **43,** 57.
Kraus, J. D., Gearhart, M. R. 1975, *Astr. J.*, **80,** 1.

Kristian, J., Sandage, A., Katem, B. 1974, *Astrophys. J.*, **191,** 43.
Mathewson, D. S., Cleary, M. N., Murray, J. D. 1975, *Astrophys. J.*, **195,** L97.
Osmer, P. 1981, *Astrophys. J.*, **247,** 762.
Osmer, P., Smith, M. G. 1980, *Astrophys. J. Suppl. Ser.*, **42,** 333.
Pauliny-Toth, I. I. K., Kellerman, K. I., Davis, M. M., Fomalont, E. B., Shaffer, D. B. 1972, *Astr. J.*, **77,** 272.
Pauliny-Toth, I. I. K., Witzel, A., Preuss, E., Kühr, H., Kellerman, K. I., Fomalont, E. B., Davis, M. M. 1978, *Astr. J.*, **83,** 451.
Shastri, P., Gopal-Krishna, 1983, *J. Astrophys. Astr.*, **4,** 109.
Shimmins, A. J., Bolton, J. G., Wall, J. V. 1975, *Aust. J. Physics, Astrophys. Suppl.*, **34,** 63.
Surdej, J., Swings, J.P. Henry, A., Josset, E.1983, in *Proc.* 24*th Liège Astrophys. Coll.* (inpress).
Strittmatter, P., Faulkner, J., Walmsley, M. 1966, *Nature*, **212,** 1441.
Tarenghi, M., Chincarini, G., Rood, H. J., Thompson, L. A. 1980, *Astrophys. J.*, **235,** 724.
Ulrich, M.-H. 1978, *Astrophys. J.*, **219,** 424.
Véron-Cetty, M.-P., Véron, P. 1984, *A Catalog of Quasars and Active Nuclei*, ESO.
Webster, A. 1977, *Mon. Not. R. astr. Soc.*, **179,** 511.
Wills, B. J., Wills, D. 1979, *Astrophys. J. Suppl. Ser.*, **41,** 689.
Wright, M. C. H. 1974, *Astr. Astrophys.*, 31, 317.

J. Astrophys. Astr. (1984) **5,** 43–53

The Radio Spectra of Galactic Centre Features

B. Y. Mills & M. J. Drinkwater *School of Physics, University of Sydney, N.S.W. 2006, Australia*

Abstract. The radio source Sgr A and neighbouring features have been mapped at a frequency of 843 MHz with a beamwidth of 43 × 87 arcsec. Comparisons have been made with published maps of comparable resolution at different frequencies in order to differentiate thermal and nonthermal regions. The arc feature to the north of Sgr A appears to consist of low-temperature ionized hydrogen and to extend partly over Sgr A itself causing patchy absorption at low frequencies; there is some evidence that the hydrogen in the arc has been expelled from the galactic nucleus. Previous suggestions that Sgr A East is a supernova remnant have been examined and the interpretation is found to be quite likely, but not compelling. The diffuse component of Sgr A West appears to be due entirely to ionized hydrogen surrounding the nucleus.

Key words: radio continuum—Sgr A—galactic centre

1. Introduction

Modelling of the emission sources associated with the galactic centre is fraught with difficulties. The distribution is complex at all scales of size and both thermal and nonthermal regions are almost inextricably mixed. Qualitatively a number of basic features can be recognized in the observed distribution. The integrated emission through the galactic disc forms a largely nonthermal ridge which stretches along the plane and has a half-brightness width of about 3 deg near the galactic centre. A disc-like nonthermal source, believed to be centred on the nuclear region, extends for two or three degrees along the plane on either side of the nucleus with a width of about half a degree; the emission appears to be strongly concentrated towards the nucleus. A similar but smaller distribution of ionized gas is also centred on the nucleus and extends for about $1\frac{1}{2}$ deg along the plane on either side; this is observed in absorption at very low frequencies and in emission at high frequencies. At high frequencies, also, several giant H II regions are observed in emission close to the nucleus. The dominant central source, Sgr A, embraces the accepted position of the nucleus and is the principal subject of the present investigation.

The brightness distribution of Sgr A is also complex and recognizable features depend strongly on the resolution and frequency of the observing telescope. At a resolution of ~ 1 arcmin we see the picture in Fig. 1, a bright centrally concentrated radio source with a faint arc-like feature extending to the north. At much higher resolutions, $\lesssim$ 5 arcsec, the arc disappears because of inadequate sensitivity but

structures in the central source become distinguishable (*e.g.* Ekers *et al.* 1983). A compact non-thermal source, coincident with the infrared (IR) source, is believed to represent the actual nucleus. Apparently associated with this is a bright spiral-like structure of ionized gas superimposed on a more extended diffuse feature with a flat but possibly non-thermal spectrum. This complex of features is called Sgr A West. A bright shell-like non-thermal source of diameter ~ 3 arcmin is centred about $1\frac{1}{2}$ arcmin to the east. It is designated Sgr A East and has been recognized as a possible supernova remnant. One side of this source coincides with the position of the Sgr A West maximum.

The present investigation makes use of observations with the Molonglo Observatory Synthesis Telescope (MOST) at a frequency of 843 MHz. We obtain some information about the unresolved components of Sgr A West but are principally concerned with spectral studies of well-resolved features, using for comparison purposes the 10.7 GHz map of Pauls *et al.* (1976) which has comparable solid angle resolution. Both maps are tied to a local zero of brightness but absolute values are needed for some purposes. Where necessary, these have been obtained making use of the lower resolution maps of Little (1974) at 408 MHz and Altenhoff *et al.* (1978) at 4.87 GHz. The 408 MHz data have also been very useful in helping to define the properties of the arc-like feature.

2. Map synthesis

The MOST has been described briefly by Mills (1981) and Durdin, Little & Large (1984). It is an east-west rotational synthesis array which uses a comb of fan beams to perform a synthesis in real time. The beams are formed using two co-linear parabolic reflectors each 780 m long by 11.6 m wide; they are directed in a north-south plane by rotating the reflectors about their long axis and in an east-west plane by phasing of the individual circularly polarized antenna elements at the line foci. To synthesize a map centred on a declination of – 29 deg, the maximum east-west beam swing is 61 deg which is the practical limit of the phasing system before marked deterioration sets in; there are already some problems of dynamic range evident in the synthesised map.

The region was observed on 1983 February 24 using a field size of 46 × 46 cosec δ arcmin. The synthesised beamwidth was 43 × 87 arcsec to half power with a negative sidelobe of –8 per cent. The effects of this and other sidelobes (< 1 per cent) were removed from the map using a standard cleaning process (Crawford 1984). The map was cleaned to 1.5 per cent of the peak and restored using the positive lobe of the actual beam. The resulting map is shown in Fig. 1.

The map zero is not well defined: it is depressed by an amount proportional to the 'uncleaned' emission received by the 2 deg fan beams and—near the galactic centre—this emission is substantial. The comparison map at 10.7 GHz has similar but less severe problems (Pauls *et al.* 1976). It has been tied to a local zero and there is also a weak positive 'spill-over' from the strong Sgr A response. For comparison we have chosen an effective zero for the MOST map at – 1.5 per cent of the Sgr A positive peak or at –0.37 Jy/beam. The zero was chosen to match as nearly as possible the zero of the 10.7 GHz map. The contour unit on the map is 1 per cent of the positive peak deflection and corresponds to a brightness temperature of T_B = 133 K.

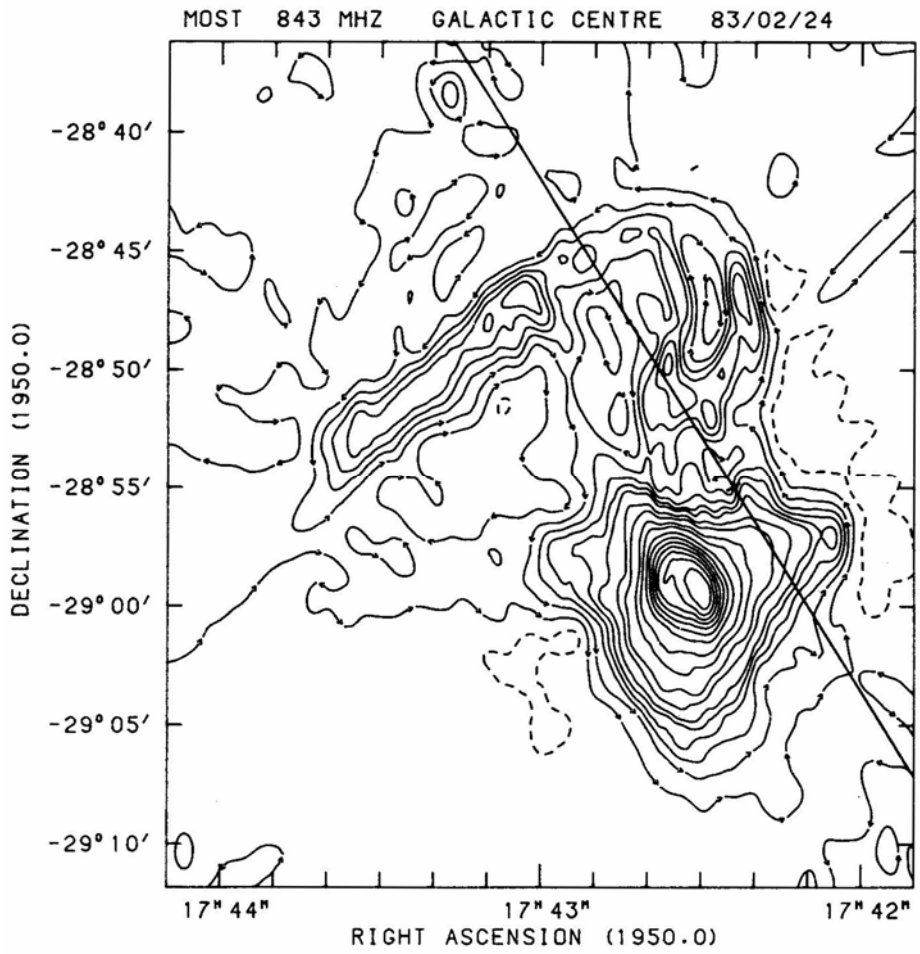


Figure 1. Cleaned MOST map of the emission near Sgr A at 843 MHz. The contours are at – 1 (dashed), 0, 1, 2, 3, 4, 5, 7.5, 10, 15, 20, 30 . . . 90 percent of the peak brightness of 24.9 Jy/beam (13300 K). The half-power ellipse of the MOST beam is given in the bottom left corner. The galactic equator ($b = 0$) is also shown.

3. The arc region

The most striking feature in Fig. 1 is the arc-like structure which extends northwards from the main peak of Sgr A and then bends sharply to cross the galactic plane in a south-easterly direction. The northerly extension breaks into two clearly defined ridges of emission which are also evident in the 10.7 GHz map of Pauls *et al.* (1976). A third very weak ridge, not evident at 10.7 GHz, is probably unassociated with the arc, possibly even an artefact.

Because the brightness of the arc feature is very low compared with Sgr A itself, there are some problems associated with determining its spectrum. Three well-defined maxima in the arc emission have been chosen, G 0.16 – 0.15, G 0.18 – 0.04 and G 0.10 + 0.18. From cross-sections of the arc constructed from the contour diagrams at these

Table 1. Spectra of three well-defined maxima in the arc emission.

	ΔT(K)			α	
	408 MHz	843 MHz	10.7 GHz	10.7–0.843	10.7–0.408
G 0.16−0.15	2900	690	4.3	−0.01	+0.01
G 0.18−0.04	2400	760	5.0	+0.02	+0.11
G 0.10+0.08	2000	640	3.5	−0.05	+0.06

positions, local zeros and the temperature excess due to the arc have been estimated. The excess temperatures, ΔT, are given in Table 1, together with the effective spectral index, α between 10.7 GHz and 843 MHz, defined by $\Delta T \propto \nu^{\alpha-2}$. Further information about low-frequency spectra has been obtained from the 408 MHz map of Little (1974). Applying a similar procedure the corresponding results are also listed in Table 1.

It is evident that the spectra are consistent with thermal emission from ionized hydrogen. The optical depths of all regions appear to be small but significant, particularly at 408 MHz.

In principle the electron temperature and emission measure of the H II regions may be obtained from these results although high accuracy cannot be expected. For a uniform H II region of electron temperature T_e and optical depth τ observed in front of a region of uniform brightness temperature T_B we have

$$\Delta T = (T_e - T_B)\,(1 - e^{-\tau}).$$

We adopt T_B = 1700 K at 408 MHz based on the map of Little (1974) on the assumption that slightly more than half the extended emission in the direction of the arc originates behind it. Although appreciably lower at 843 MHz, T_B turns out to be significant at this frequency also. We adopt a spectral index, α = – 0.3, based on comparison with the appropriate map of Altenhoff *et al.* (1978) at 4.87 GHz; at 843 MHz this gives T_B = 320 K. Using these estimates of T_B and the relation $\tau \propto \nu^{-2.09}$ (based on the variation of Gaunt factor over this frequency range) we have solved for T_e and τ at 408 MHz and 843 MHz in conjunction with the 10.7 GHz measurements. These results are shown in Table 2, together with the emission measures calculated from the estimated 10.7 GHz temperatures on assuming an electron temperature of 6000 K.

There is considerable scatter in the derived electron temperatures but this is understandable in view of the uncertain estimates of ΔT and T_B; the results are consistent with a mean temperature of about 6000 K. The temperatures of 6000–12000 K deduced from radio recombination-line measurements (Pauls *et al.* 1976; Pauls & Mezger 1980) are higher and no line could be detected at the position of

Table 2. Derived properties for three maxima in the arc emission.

	T_e(K)		τ		EM
	408 MHz	843 MHz	408 MHz	843 MHz	× 10^4
G 0.16−0.15	12000	3800	0.32	0.22	13
G 0.18−0.04	6500	3000	0.70	0.33	15
G 0.10+0.08	7300	7500	0.45	0.09	11

G 0.16–0.15, leading to the suggestion that the emission here is nonthermal. An underestimate of the underlying nonthermal emission may contribute to an overestimate of temperatures using radio recombination lines and—in view of all the uncertainties—we do not believe that the discrepancies are significant. To account for the failure to detect line radiation in G 0.16–0.15, a high electron temperature and/or a large velocity dispersion may be needed in addition. There can be little doubt that the latter source is largely thermal. A possible reason for an underestimate of the nonthermal component of brightness is a ridge of emission which can be recognized extending for about two degrees across and approximately normal to the plane close to the longitude of the arc feature. Comparison of the data of Little (1974) and Altenhoff *et al.* (1978) indicates that the ridge has a non-thermal spectrum with $\alpha \sim -0.6$, compatible with the general galactic radiation. This ridge is an unusual feature and as it is almost coincident with another unusual feature, the arc, the possibility of a physical association might be considered. However, we can suggest no plausible physical connection.

The mass of gas in the whole arc feature may be estimated roughly by adopting a mean emission measure of 6×10^4 and a model comprising a long cylinder lying in a plane perpendicular to the line of sight. An inclination of the plane increases the derived mass and an allowance for a filling factor decreases it. The model yields a mass M ~5000 $M_\odot$.

What then is the nature of the arc? The entire feature appears to be a region of ionized hydrogen with little or no associated nonthermal emission. Interpretation as a simple spiral feature inclined to the galactic plane seems impossible, even allowing for gas motion along the spiral, because of the distribution of radial velocities determined from radio recombination lines (Pauls *et al.* 1976; Pauls & Mezger 1980). These show a velocity $\lesssim -40$ km s^{-1} just north of Sgr A increasing to zero at the bend in the arc and rising further to + 40 km s^{-1} at the position of G 0.18 – 0.04. Such a pattern could be generated by an expanding ring or, perhaps more plausibly, by a precessing jet of ionized hydrogen arising in the nuclear region. The latter would involve much lower expulsion velocities than usually associated with jets (~ 100 km s^{-1}) but the shape of the feature, particularly the sharp bend followed by the straight south-westerly extension, can be readily modelled. The estimated age of such a feature at its extremity would be ~ 10^6 yr, roughly the same as the precession period, and the rate of expulsion of matter, $M \sim 5 \times 10^{-3}\ M_\odot$ yr^{-1}.

A precessing jet model on a much smaller scale has been proposed by Brown (1982) to account for the spiral-like structure of the thermal component of Sgr A West, but his derived parameters are incompatible with the arc feature; the precession period is shorter by two or three orders of magnitude and the precession is in the opposite sense. It seems that there is no simple model tying together these features but nevertheless an association of the arc feature with the nucleus appears likely.

4. Sagittarius A

For the purposes of analysis we define Sgr A as the emission region south of declination – 28° 54′. The sources conventionally named Sgr A East and Sgr A West are not separately resolved in Fig. 1 although the beginnings of the ring structure of Sgr A East can be recognised; these bright central sources are surrounded by an extensive low-

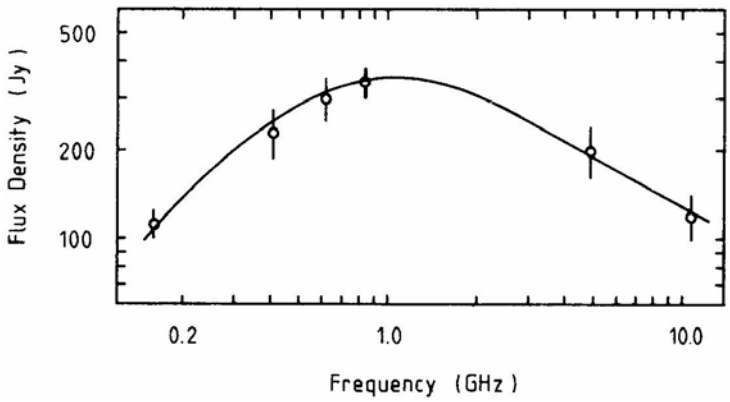


Figure 2. Radio spectrum of Sgr A, based on observations with resolution better than 3 arcmin. (Data sources are, 160 MHz: Dulk & Slee 1974; 408 MHz: Little 1974; 610 MHz: Downes *et al.* 1978; 843 MHz: this work; 4.875 GHz: Downes *et al.* 1978; 10.7 GHz: Pauls *et al.* 1976.)

brightness region. The 843 MHz appearance is qualitatively similar to the 10.7 GHz map of Pauls *et al.* (1976), although a detailed comparison reveals great variations in spectral index. At 843 MHz, the integrated flux density is 340 Jy. The spectrum of Sgr A between 10 GHz and 160 MHz is given in Fig. 2; the high frequency spectral index is $\alpha \simeq -0.57$. The low frequency turnover suggests absorption from intervening ionized hydrogen with a patchy distribution and a mean emission measure of $\sim 7 \times 10^4$, similar to the mean emission measure of the arc feature.

4.1 *Spectral Index Distribution*

The distribution of spectral index across Sgr A has been derived from a comparison of the present map with the 10.7 GHz map of Pauls *et al.* (1976). A zero level of – 1.5 per cent has been adopted for the 843 MHz map, as discussed in Section 2, and comparisons have been made only when the derived brightness is more than 4 per cent of the peak. This largely avoids problems associated with uncertainty of the mean zero level and its possible variation over the map. The derived contours of α are shown in Fig. 3.

Although the accuracy is poor, there is clearly a north-south gradient in α, evidently indicating a gradient in the proportion of thermal emission. On the northern border the emission is dominated by the ionized hydrogen which continues northwards to form the arc. The central region of low α is associated with the bright Sgr A East source which contributes most of the emission. Also there appears to be a real lack of hydrogen just north of this source, whereas the adjacent Sgr A West shows clearly a pocket of high α, undoubtedly associated with a concentration of ionized hydrogen around the nucleus. It appears that the nonthermal index of the fainter extended emission is $\alpha \simeq -0.7$ but, because of the zero level problem, the uncertainty is large. This emission appears to be associated with the nucleus and may represent the central peak of the extended non thermal source which stretches for some 5 deg along the galactic plane.

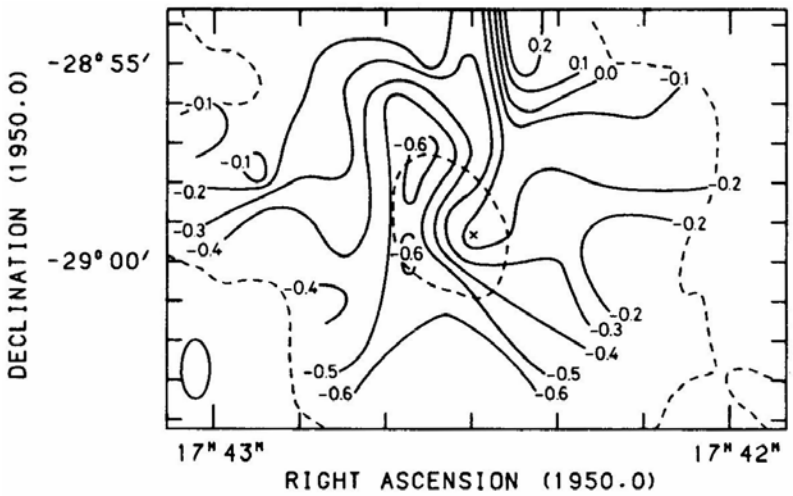


Figure 3. Distribution of spectral index near Sgr A, measured between 843 MHz and 10.7 GHz. The zero and 50 per cent contours of Fig. 1 are shown dashed and the position of Sgr A West is indicated.

4.2 *Sgr A East*

The central Sgr A concentration is an elliptical source with the major axis parallel to the galactic plane. The two components are not clearly resolved in the map of Fig. 1 but comparisons of cross-sections through the source at 843 MHz and 10.7 GHz shown in Fig. 4 indicate the two regions of different spectra. These agree closely with the detailed VLA maps presented by Ekers *et al.* (1983). The 10.7 GHz–843 MHz spectral indices at the centres of the East and West sources are α (East) = – 0.56 and α (West) = – 0.17.

The eastern source has been known for some time to have a ring-like structure and the possibility that it is a supernova remnant (SNR) has been discussed by several authors (Jones 1974; Gopal-Krishna & Swarup 1976; Goss *et al*., 1983). The strongest case has been made out by Goss *et al*, who used VLA observations which demonstrate a morphology similar to that of many SNRs; the spectral index is also compatible. If the source is a SNR situated on a line of sight that passes close to the nucleus, the question of its distance arises. Is it a chance alignment or is it actually in the central region? Goss *et al.* (1983) conclude that it is most probably within the nuclear bulge, at least. Using the results of Mills *et al.* (1984) we find that a typical remnant having the properties of Sgr A East would have an even chance of being located within about 2–3 kpc of the nucleus, on either side. As the occurrence of a supernova is a chance event proportional to the local stellar density, a location in the central region is very much more probable than any other single location along the line of sight, but the probability is still low.

X-ray observations by Watson *et al.* (1981) provide further clues. Several X-ray sources, including Sgr A West, were detected close to the nucleus but no source was found at the position of Sgr A East. If this were a typical small-diameter SNR (< 10 pc), the data of Long, Helfand & Grabelsky (1981) on the LMC supernova remnants would suggest a high luminosity, $\sim 10^{37}$ erg s^{-1} in the energy range 0.15–4.5 keV. However, the absorption is also high; Watson *et al.* estimate that the hydrogen column density to the nucleus lies between 2×10^{22} and 10^{23} cm^{-2}. Detection of a SNR near the nucleus

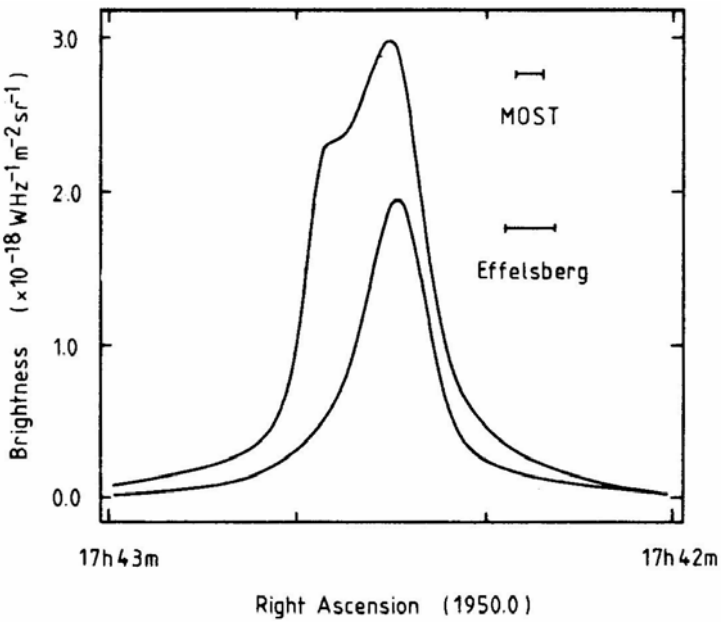


Figure 4. Gross-sections in R.A. of Sgr A at 843 MHz (upper curve) and 10.7 GHz (lower curve). The respective half-power beamwidths are shown.

with the above luminosity might be expected at the lower extreme of density, but not at the higher. Thus no positive conclusion is possible although a location much in front of the nucleus would seem to be unlikely. For a typical SNR the results of Mills *et al.* (1984) imply an age of ~ 300 yr and a possible expansion rate of ~ 0.1 arcsec yr^{-1}; proper motion studies may be profitable before long. However, if the supernova were located in the dense environment close to the galactic centre it is most unlikely that these results would be applicable and an old slowly expanding remnant might be expected.

Finally, we must consider the possibility that the source is not a supernova remnant. The positional coincidence of the nucleus and the brightest side of the source, together with the symmetry about the minor axis, does suggest a possible physical relationship to the nucleus, but the form this relationship might take is unclear. Further speculation is pointless at present; eventually polarization data may contribute to an understanding of the source.

4.3 *Sgr A West*

Ekers *et al.* (1983) describe three features associated with Sgr A West, a compact non-thermal source coincident with an infrared source and believed to represent the nucleus, a spiral feature of bright ionized hydrogen and a diffuse source of 'non-thermal' emission of spectral index $\alpha \simeq -0.3$ centred on the nucleus. The compact source is weak and irrelevant to the present observations. The diffuse 'non-thermal' source is located in Fig. 3 in the region of flattish spectral index coincident with the nucleus (peak $\alpha = -0.17$). We prefer to interpret it as a thermal source superimposed on the non-thermal Sgr A East; that is, we attribute the excess emission to diffuse ionized hydrogen

surrounding the nucleus. There appears to be no need to introduce another class of non-thermal sources with a different spectral index, either here or to interpret the VLA results. With this interpretation it is possible to model the emission and obtain information about the possible locations of Sgr A East with respect to the nucleus.

Three models have been examined in which (1) Sgr A East is located in front of the nucleus, (2) Sgr A East is behind the nucleus, and (3) the western part of Sgr A East is located at the nucleus and surrounds the thermal source. We assume that the thermal source has dimensions given by the region of flattest spectral index in the distribution of Ekers *et al.* (1983), that is a projected area of 0.33 arcmin^2, which is approximately 2/5 of the Molonglo beam size. The depth is taken as equal to the average width of the distribution $\simeq 0.7$ pc.

(1) The brightness temperature at 843 MHz is given by

$$T'_{.843} = \beta_M T_e (1 - e^{-\tau}) + T_{.843},$$

where $\beta_M = 0.4$ is the fraction of the Molonglo 843 MHz beam occupied by the thermal source, T_e is the electron temperature of the thermal source and $T_{.843}$ is the brightness temperature of the adjacent region of Sgr A East. A similar equation may be applied to the 10.7 GHz map of Pauls *et al.* (1976), when we have

$$T'_{10\,7} = \beta_E T_e\ \tau + T_{10\,7},$$

where $\beta_E = 0.26$ is the fraction of the Effelsberg 10.7 GHz beam occupied by the thermal source. Assuming $\tau \propto \nu^{-2\,1}$ and $T_\nu \propto \nu^{-2.56}$, these equations may be solved to give $T_e = (7600 \pm 2000)$ K and $\tau_{.843} = 4 \pm 2$. The derived mass for a uniform model is $M = (70 \pm 30)\ M_\odot$.

(2) The expression for the 843 MHz brightness temperature now becomes

$$T'_{.843} = \beta_M [T_e(1 - e^{-\tau}) + T_{.843}\ e^{-\tau}] + (1 - \beta_M)T_{.843}.$$

Solving as before, we find $T_e \sim 20000$ K and $\tau_{.843} \sim 1.5$ with considerable uncertainty. The derived mass is $M \sim 100\ M_\odot$.

(3) With the thermal source surrounded by the nonthermal Sgr A East we have

$$T'_{.843} = \beta_M [T_e (1 - e^{-\tau}) + T_{.843} (1 + e^{-\tau}) (l - s)/2l] + (1 - \beta_M)T_{.843},$$

where s is the depth of the thermal source and l is the depth of the non-thermal source. From the maps of Ekers *et al.* (1983) we estimate $s/l \simeq 1/7$. Solving, we then find $T_e = (13000 \pm 3000)$ K and $\tau_{.843} = 2.4 \pm 0.7$. The mass of the ionized hydrogen is $M = (80 \pm 30)\ M_\odot$.

The derived electron temperature is very sensitive to the location of Sgr A East. In view of the uncertainties, none of the models can be rejected, but a location behind the nucleus is less plausible because of the necessary high electron temperature. On the other hand the derived mass of the H II region is insensitive to location and appears to be $\sim 80\ M_\odot$, in reasonable agreement with the high frequency VLA results of Brown & Johnston (1983).

Although the spiral feature is too small for a recognizable contribution to the present results, the data of Ekers *et al.* (1983) suggest to us a model which apparently has not been considered. Qualitatively, the shape of the feature and the pattern of velocities measured in the Ne II line at 12.8 μm (Lacy *et al.*, 1980) could be explained by a model comprising a precessing northwards flowing jet, combined with a rapidly rotating accretion disc into which gas flows from the galactic plane along distorted paths, of

which we observe the regions of greatest optical depth. The rate and direction of the precession required to account for the velocity distribution along the jet is, however, not compatible with the jet model of the arc discussed in Section 3. Further velocity data would be needed to explore the model quantitatively.

5. Conclusions

Observations of the galactic centre with a resolution of ~ 1 arcmin over a wide frequency range provides information supplementary to the high-resolution VLA observations and, in particular, it removes some uncertainties of interpretation resulting from the absence of low spatial frequencies. Our conclusions may be summarized as follows:

(1) The northern arc feature is an elongated distribution of ionized hydrogen, probably of rather low temperature. Modelling of the velocity field is difficult but it does fit a picture of a low-velocity 'precessing' jet arising in the nuclear region; it does not fit a simple spiral structure.

(2) The integrated spectrum of Sgr A shows evidence for patchy absorption of ionized hydrogen. The greatest concentration of hydrogen is located at the position of the galactic nucleus, but otherwise, it increases northwards in the direction of the arc.

(3) Sgr A East remains a problem. Interpretation as a supernova remnant within, or possibly in front of, the nuclear region seems most likely but a possible interpretation as a source deriving from the nucleus cannot be excluded.

(4) Sgr A West is most simply interpreted as a thermal source plus the very compact non-thermal component identified with the nucleus itself. We are inclined to believe that the spiral feature evident in VLA maps may combine both accretion and ejection but cannot directly associate it with the northern arc structure, with which it does have some features in common.

Acknowledgements

Operation of the Molonglo Observatory Synthesis Telescope is supported by the Australian Research Grants Scheme and by the University of Sydney.

References

Altenhoff, W. J., Downes, D., Pauls, T., Schraml, J. 1978, *Astr. Astrophys. Suppl. Ser.*, **35,** 23.

Brown, J. L. 1982, *Astrophys. J.*, **262,** 110.

Brown, R. L., Johnston, K. J. 1983, *Astrophys. J.*, **268,** L85.

Crawford, D. F. 1984, *URSI/IAU Symp. on Measurement and Processing for Indirect Imaging* (in press).

Downes, D., Goss, W. M., Schwarz, U. J., Wouterlout, J. G. A. 1978, *Astr. Astrophys. Suppl. Ser.*, **35,** 1.

Dulk, G. A., Slee, O. B. 1974, *Nature*, **248,** 33.

Durdin, J. M., Little, A. G., Large, M. I. 1984, *URSI/IAU Symp. on Measurement and Processing for Indirect Imaging* (in press).

Ekers, R. D., van Gorkom, J. H., Schwarz, U. J., Goss, W. M. 1983, *Astr. Astrophys.*, **122,** 143.

Gopal-Krishna, Swarup, G. 1976, *Astrophys. Lett.*, **17,** 45.
Goss, W. M., Schwarz, U. J., Ekers, R. D., van Gorkom, J. H. 1983, in *IAU Symp.* 101: *Supernova Remnants and their X-ray Emission*, Eds J. Danziger and P. Gorenstein, D. Reidel, Dordrecht, p. 65.
Jones, T. W. 1974, *Astr. Astrophys*., **30,** 37.
Lacy, J. H., Townes, C. H., Geballe, T. R., Hollenbach, D. J. 1980, *Astrophys. J*., **241,** 132.
Little, A. G. 1974, in *IAU Symp.* 60: *Galactic Radio Astronomy,* Eds F. J. Kerr and S. C. Simonson III, p. 491.
Long, K. S., Helfand, D. J., Grabelsky, D. A. 1981, *Astrophys. J*., **248,** 925.
Mills, B. Y. 1981, *Proc. astr. Soc. Aust*., **4,** 156.
Mills, B. Y., Turtle, A. J., Little, A. G., Durdin, J. M. 1984, *Aust. J. Phys.* (submitted).
Pauls, T., Downes, D., Mezger, P., G., Churchwell, E. 1976, *Astr. Astrophys.*, **47,** 407.
Pauls, T., Mezger, P. G. 1980, *Astr. Astrophys.*, **85,** 26.
Watson, M. G., Willingale, R., Grindlay, J. E., Hertz, P. 1981, *Astrophys. J.*, **250,** 142.

J. Astrophys. Astr. (1984) **5,** 55–65

Ionospheric Refraction in Radio Source Observations at Long Radio Wavelengths

W. C. Erickson* *Clark Lake Radio Observatory, Astronomy Program, University of Maryland, College Park, MD 20742, USA*

Abstract. Ionospheric refraction effects encountered in radio source observations in the 30 to 75 MHz range with the Clark Lake TPT telescope are discussed. It is found that simple calibration procedures are sufficient to provide positions of unknown sources with an accuracy of approximately one arcmin. Observations made near sunrise, or during disturbed ionospheric conditions must be discarded. If no corrections are applied, RMS errors of a few arcmin are to be expected.

1. Introduction

The Clark Lake TPT (Erickson, Mahoney & Erb 1982) is a high resolution radio-telescope which operates in the 15 to 125 MHz frequency range. Its location is (116°17′E, 33°20′N). The best sensitivity of the system (about 1 Jy) is in the 25 to 75 MHz range and its beamwidth varies from 13.8 to 4.6 arcmin over this frequency range. Thirty-two signal outputs from the 3 km East-West arm are digitally correlated with sixteen outputs from the 1.8 km North-South arm of the 'T'. The resulting 512 correlator outputs are averaged for a few minutes and then Fourier transformed to produce a map of the area of sky under observation. Successive maps are stacked for 30 to 80 minutes; a final map is then produced and cleaned.

Every few days a strong source such as Cyg A is observed and these data are used to adjust the phases and gains of the individual receiver channels. This adjustment compensates for ionospheric refraction existing at the time of this observation. Therefore, no fundamental measurements are made; all measurements are relative to the apparent position of the source used for instrumental calibration.

Further corrections for ionospheric refraction are often possible. The field of view of the system is large, 2 to 5 deg, and many fields contain at least one source of accurately known coordinates which can be used to calibrate the positions of unknown sources in the same field.

We also observe strong ($\gtrsim$ 50 Jy) isolated sources several times a day as a check on the operational status of the system and to determine whether or not data should be discarded because of severe ionospheric scintillation. It will be shown below that the displacements of the apparent position of these sources from the field centre, presumably caused by ionospheric refraction, are generally consistent from source to source observed over intervals of many hours. These data can be used to correct the

* On leave at Radiosterrenwacht Dwingeloo, The Netherlands.

positions of unknown sources when there are no sources of known coordinates simultaneously in the field.

One would be foolish to attempt astrometric observations in this frequency range. Normally it is unnecessary because even steep spectrum sources observed below 100 MHz can be detected by sensitive instruments at gigahertz frequencies. Occasionally, however, the identification of the low-frequency source with the proper

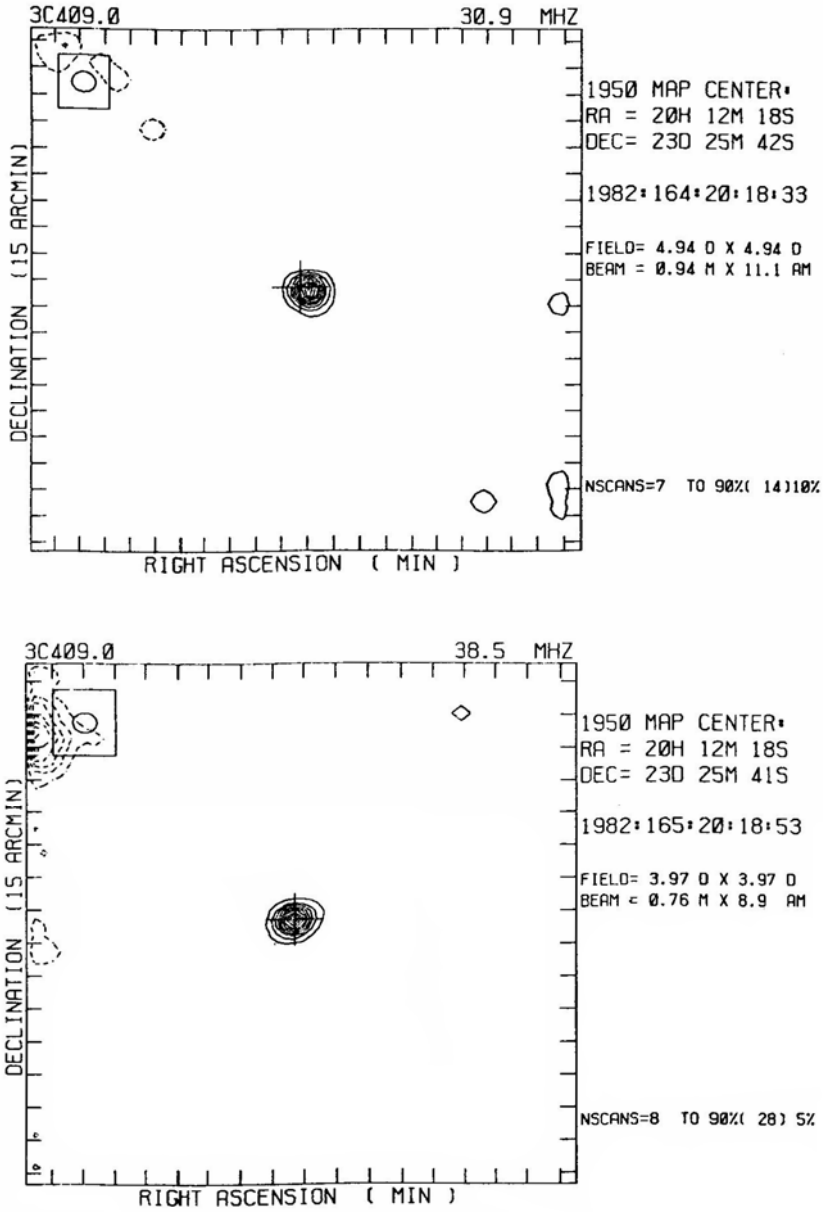


Figure 1. Examples of the data used for this analysis. Contours are at ± 5, ± 15, . . . ± 95 per cent of the peak on the map. The beamsize is shown in the upper left corner.

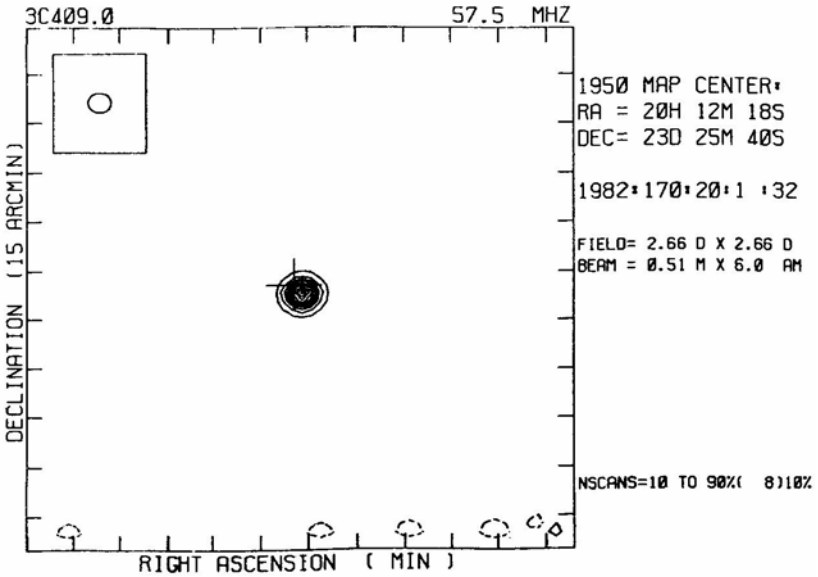


Figure 1. continued

high-frequency object is ambiguous and an accurate low-frequency position is needed. A recent example of this was the case of the first millisecond pulsar. The position determined at Clark Lake for the steep-spectrum, low-frequency source was 2–4 arcmin south of the extended source (4C21.53W) and was, in fact, the pulsar (Erickson 1983).

Another reason for studying refraction effects at long wavelengths is that new long-wavelength instruments are being considered at VLA, in India, and elsewhere. Practical information concerning ionospheric stability is important for the design of such instruments.

2. Observations

The observations included in this study are from data that I obtained during four observing trips to Clark Lake, each about two weeks long, in 1981 November-December and in 1982 February, March and June. Most of the data were obtained at night because terrestrial interference is less common at night and, also, the telescope is usually occupied by solar programmes in the daytime. The data are thus typical of those that can be obtained at night during most parts of the year, but they certainly do not represent a statistically complete sample. About 20 per cent of the original observations were recognized as being obviously bad because of interference, severe ionospheric scintillations, or severe wedge refraction that occurs near sunrise. Such data were discarded in the early stages of reduction. No further attempt has been made to select particularly good data; all observations of strong, isolated sources made during these observing sessions are included here.

It would be interesting to study the short-period ($\simeq 5$ s) fluctuations in the apparent radio source positions. This can be done but it would require special software and

reprocessing of the original data tapes. Only the long-period ($\simeq$ 1 hour average) refraction effects are studied in this paper.

A total of 110 source observations at frequencies of 30.9, 38.5, 57.5 and 73.8 MHz are included in this study. Fig. 1 gives examples of the data. For each observation the displacement of the source was measured relative to the map centre (given by an accurate, high-frequency position). All observations were made fairly close to transit, so displacements in right ascension represent E-W refraction while declination offsets represent N-S refraction.

3. Results

Table 1 summarizes the RMS displacements observed at the various frequencies. Although the displacements increase with wavelength, they do not show a clear λ^2 dependence. However, as shown in Fig. 2 where a line of slope – 2.0 is fitted through the 30–75 MHz data, the magnitude of the low-frequency refraction effects is consistent with that found by Spoelstra (1983) using the WSRT at 608.5 MHz.

Fig. 2 shows that the RMS displacements are not significantly different in the E-W and N-S directions. They may be crudely estimated by

$$\text{RMS error} \sim (80/F)^2$$

where the RMS error is in arcmin and F is the observing frequency in MHz. This is the accuracy that can be anticipated if no corrections for ionospheric refraction are made.

Expressions for the refraction caused by both a spherically symmetrical component and a wedge component of the ionosphere were derived by Komesaroff (1960) and by Lowen (1962). These expressions have been extended by Spoelstra (1983) who shows that quite a good correction can be made if one has sufficient ionospheric data to determine both the vertical profile and the horizontal gradients of electron density. These data are not easily available and the process would be rather difficult to implement in our case, so I have adopted the simple procedure of observing calibration sources before and after the observation of an unknown source. It is useful to determine to what accuracy corrections can be made using such data and whether or not any systematic trends can be found in these data. In order to search for such trends, I have plotted the displacements against many parameters. In particular, one might expect that the N-S displacements would vary systematically with declination, or that the E-W displacements might depend upon local solar time. As is shown in Fig. 3, any such trends are below the noise level.

Table 1. RMS displacements observed at different frequencies.

Frequency MHz	RMS (E-W) $\cos\delta\,\Delta\alpha$ arcmin	RMS (N-S) $\Delta\delta$ arcmin	Number of observations
30.9	4.20	3.89	21
38.5	3.32	2.72	15
57.5	3.58	3.00	67
73.8	1.13	1.12	7
(608.5)	~0.0185	~0.0185	...

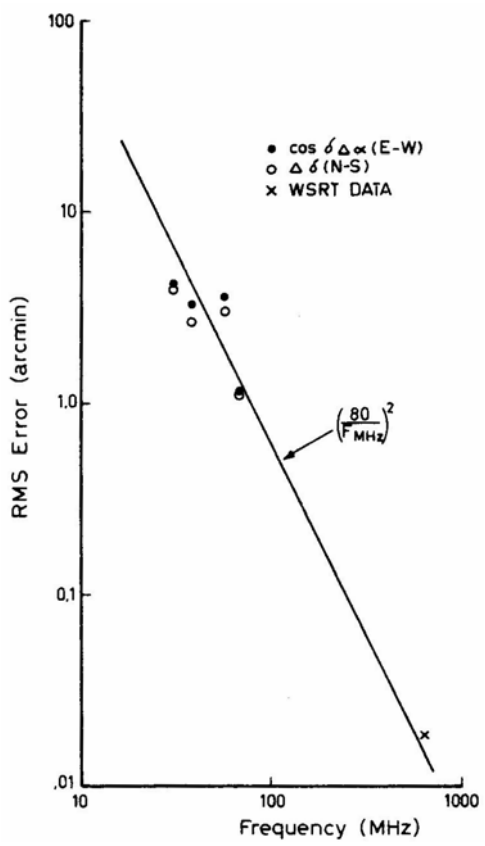


Figure 2. The variations of RMS displacements as a function of frequency.

I can find only one effect that appears to be significant. The scatter of the displacements is a factor of two smaller for sources that lie between 6 and 16 h right ascension than for sources that lie outside this range. For sources within this range of right ascension both the E-W and the N-S displacements have a Standard deviation of 1.4 arcmin while for sources outside this range the corresponding Standard deviations are both 2.8 arcmin. One possible explanation for this effect involves the system noise levels. The system noise is dominated by the galactic background which is the lowest during 6 to 16 h interval of sidereal time. However, if system noise were contributing to the errors I would expect that the displacements would be smaller for the stronger sources. A plot of RMS displacements versus source flux shows no such tendency.

Another possible explanation involves solar or seasonal variations in ionospheric activity. Since the observations were made near transit and mostly at night, I tended to observe different ranges of right ascension during the four different observing sessions. If the ionosphere happened to be much more stable during one or two of the sessions, the quality of the data and the scatter of the displacements in the observed range of right ascension might be lower. To check on this possibility I scaled all of the displacements to 57.5 MHz and averaged them for each observing session. The average displacements and the standard deviations from these averages are shown in Table 2. No significant

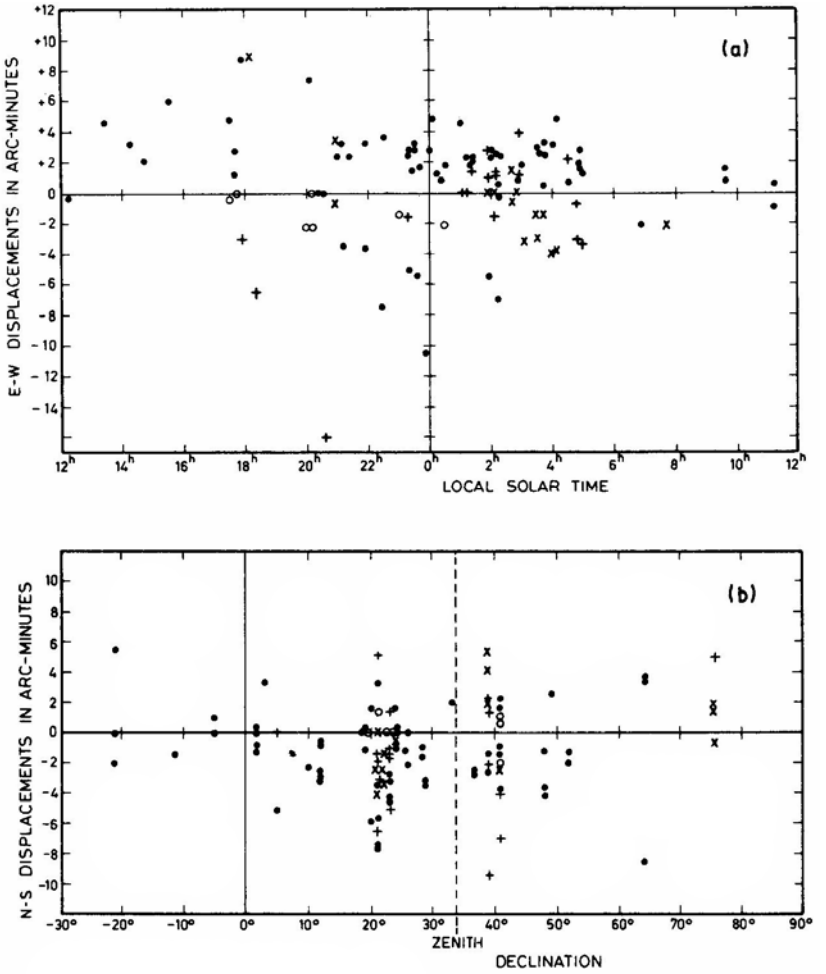


Figure 3. (a) E-W displacements as a function of local solar time, (b) N-S displacements as a function of declination. The displacements do not significantly depend upon these or any other parameters for which similar plots were made. Different symbols are used for measurements at different frequencies: 30.9 MHz (+), 38.5 MHz (×), 57.5 MHz (●), 73.8 MHz (O).

Table 2. Average displacements observed during different sessions scaled to 57.5 MHz.

Observing session yr d	Average E-W displacement arcmin	Average N-S displacement arcmin	Number of observations
1981 314–351	−3.3±3,7	+1.0±2.1	14
1982 029–041	0.0±2.0	−0.1+1.6	15
1982 067–075	+2.5±2.0	−1.4±2.3	43
1982 160–177	+0.4±1.5	−1.7±2.8	38

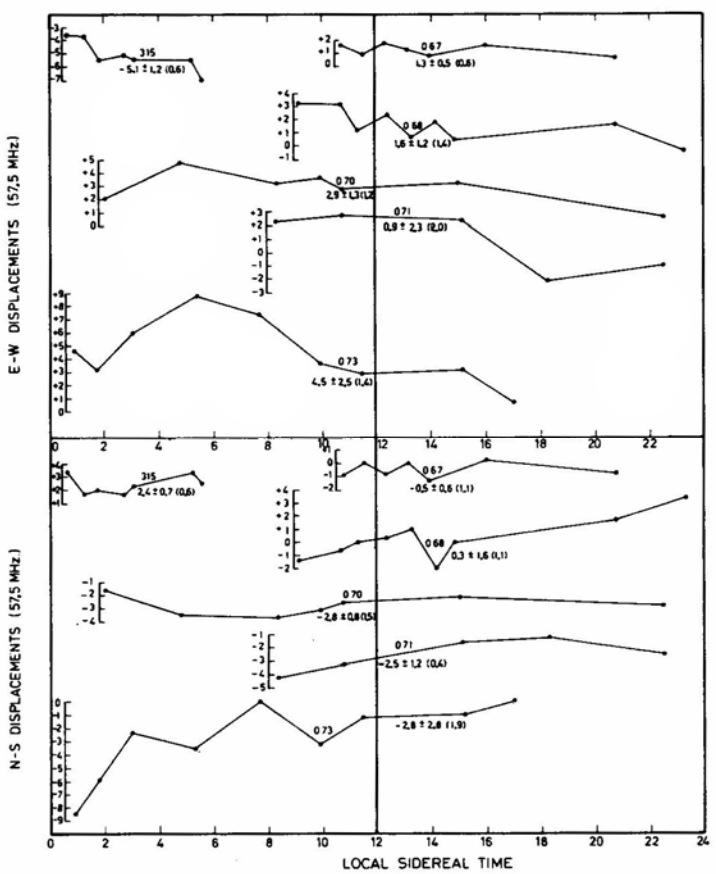


Figure 4. The variation of refractive displacements as a function of time. Sources at widely varying declinations were observed near transit. The scale to the left of each plot gives the displacement in arcmin. The date of observation is given above each plot; the average displacement and its standard deviation is given below each plot. Also, the average errors between the measured displacements and those predicted by linear interpolation between adjacent observations are shown in parentheses.

dependence of the quality of the data upon observing session is apparent so the explanation of the effect remains a mystery.

There may also exist a weak correlation between observed displacements in source position and observed errors in source flux, but the correlation is not strong enough to be of any practical use in estimating the error of either quantity.

At 57.5 MHz there exist six days during which I observed 5 to 9 calibration sources successively over periods of 5 to 20 h. These periods can be used to estimate the stability of the displacements over periods of several hours. As shown in Fig. 4, the displacements are quite stable. If I simply calculate a mean displacement for each day and determine the scatter of the observations about that mean, the average errors are 1.5 arcmin (E-W) and 1.4 arcmin (N-S). On the other hand, if I predict source displacements by linear interpolation between observations taken before and after any given one, the average differences between the predictions and the observations are 1.2

Table 3. Comparison of Texas Survey source positions with Clark Lake observations.

	Texas survey			Clark Lake observation				
Source Name	R.A. (1950) h m s	Dec. (1950) ° ′	S (365 MHz) mJy	R.A. (1950) h m s	Dec. (1950) ° ′	ΔR.A. s	ΔDec. arcmin	S (38.5 MHz) Jy
1008+193	10 08 44	19 22.1	333	...	...	...	...	...
1009+190	10 09 14	19 04.7	329	10 09 12	19 05.3	−02	0.6	1.4
1009+207	10 09 15	20 43.7	401	10 09 18	20 46.8	(−03)	(3.1)	(0.9)
1009+211	10 09 26	21 11.3	453	10 09 31	21 11.5	(−05)	(0.2)	(0.7)
1010+197	10 10 22	19 45.0	204	...	...	...	...	...
1010+220	10 10 37	22 00.9	337	...	...	...	...	...
1010+209	10 10 57	20 56.7	410	10 10 54	20 56.9	−03	0.2	1.9
1012+203	10 12 08	20 23.4	173	...	...	...	...	...
1012+184	10 12 26	18 24.8	578	...	...	...	...	...
1012+196	10 12 57	19 39.0	251	10 13 04	19 37.4	(07)	(−1.6)	(0.7)
1013+208	10 13 59	20 52.8	847	10 13 58	20 51.3	−01	−1.5	4.4
1014+217A	10 14 09	21 47.1	356	...	...	...	...	...
1014+217B	10 14 53	21 46.2	427	...	...	...	...	...
1015+187	10 15 03	18 47.5	1296	10 15 04	18 47.3	01	−0.2	2.1
1015+201	10 15 32	20 06.6	193	10 15 27	20 06.4	−05	−0.2	3.1
1015+203	10 15 56	20 21.1	228	10 15 50	20 22.7	−06	1.6	1.5
1016+188	10 16 29	18 51.1	929	10 16 31	18 51.4	02	0.3	4.2
1017+201	10 17 01	20 06.5	867	10 17 05	20 06.1	04	−0.4	5.8
1019+211	10 19 08	21 08.9	397	...	...	...	...	...
1019+199*	10 19 26	19 58.0	1303	10 19 22	19 57.6	−04	−0.4	4.3

1020+191	10 20 12	19 08.8	647	...	...	...	...	...
1020+197	10 20 28	19 46.9	247	...	...	...	...	...
1020+215	10 20 45	21 32.2	386	...	...	...	...	...
1020+205	10 20 59	20 33.4	520	10 21 01	20 32.6	02	−0.8	3.1
1021+207	10 21 23	20 47.6	305	...	...	...	...	...
1021+182	10 21 50	18 16.8	174	...	...	...	...	...
1022+194*	10 22 02	19 27.6	1482	10 22 02	19 28.4	00	0.8	10.6
1022+182	10 22 28	18 13.7	285	...	...	...	...	...
1022+204†	10 22 37	20 25.6	4368	10 22 36	20 27.2	−01	1.6	24.3
1023+190	10 23 29	19 00.3	250	...	...	...	...	...
1024+201	10 24 02	20 10.4	467	10 24 03	20 08.7	01	−1.5	2.1
1024+217	10 24 31	21 46.0	232	...	...	...	...	...

Average errors: R.A.. -0.9 ± 3.0 s or -0.2 ± 0.7 arcmin; Dec.: 0.0 ± 1.0 arcmin
* 4C 19.33 S(178 MHz) = 2.0 Jy
# 4C 19.44 S(178 MHz) = 2.4 Jy
† 4C 20.22 S(178 MHz) = 6.8 Jy

arcmin (E-W) and 1.0 arcmin (N-S). The reasons why the latter method is not significantly better than the former one are that measurement errors are appreciable at the one arcmin level and also a significant fraction of the displacements are caused by ionospheric gravity waves that have periods much shorter than the interval between observations. Nevertheless, these methods permit position determinations of about one arcmin accuracy.

Finally, I studied the accuracy that can be attained if position measurements are made relative to known sources in the field of view. For this purpose I needed a field which contained many sources with accurately known coordinates, and the best source of such data is the 365 MHz Texas Survey (Douglas *et al.* 1980). I chose a field in Leo that was recently observed at 38.5 MHz by R. J. Hanisch (1983, personal communication) to look for emission from the flare star, AD Leo. This field lies in the declination strip covered by the Preliminary Texas Survey and contains 32 Texas sources. Three of these sources have fluxes greater than 2 Jy at 178 MHz and appear in the 4C catalogue. As shown in Fig. 5, sixteen of the Texas sources appear on the Clark Lake map. Three of them are below 1 Jy at 38.5 MHz and I consider their identification too unrealiable for inclusion in the statistics. The Texas and the Clark Lake positions are in excellent agreement as is shown by Table 3. Again, I find RMS position errors of about one arcmin in each coordinate. A large part of the errors is caused by inaccuracies in measuring the maps and, since the sources are relatively weak, part of the errors may be caused by noise fluctuations.

Using either method of calibration the errors appear to be random. Since errors in the individual measurements are about one arcmin, the average of a number of independent measurements should provide positions accurate to less than an arcmin.

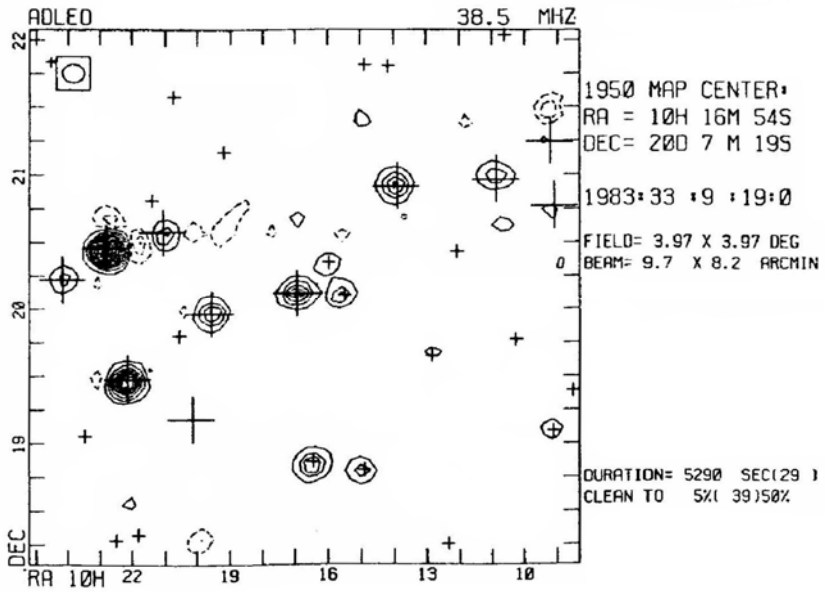


Figure 5. The comparison between Clark Lake and Texas Survey positions for a field in Leo. The large '+' indicates source with 365 MHz fluxes greater than 400 mJy; the small '+' denotes sources between 100 and 400 mJy in flux.

4. Conclusions

Under normal, night-time conditions the ionosphere is sufficiently stable to permit radio-source position determinations of about one arcmin accuracy in the 30 to 75 MHz range with single observations of about one hour duration. The average refraction is often stable for many hours and is not strongly dependent upon source position. Multiple, independent observations can be expected to yield average positions accurate to a small fraction of an arcmin. Different calibration procedures have been tested and all yield similar accuracies.

Acknowledgements

This analysis was performed while I was a guest of the Netherlands Foundation for Radio-astronomy at the Radiosterrenwacht Dwingeloo. I wish to thank T. A. Th. Spoelstra for useful discussions and comments concerning the work. The Clark Lake Radio Observatory is supported by the U.S. National Science Foundation under Grant AST-82 15463.

References

Douglas, J. N., Bash, F. N., Torrence, G. W., Wolfe, C. 1980, *Univ. Texas Publ, Astr*., No.17.
Erickson, W. C. 1983, *Astrophys. J*., **264,** L13.
Erickson, W. C., Mahoney, M. J., Erb, K. 1982, *Astrophys. J. Supp. Ser.*, **50,** 403.
Komesaroff, M. M. 1960, *Aust. J. Phys*., **13,** 153.
Lowen, R. W. 1962, *J. geophys. Res.*, **67,** 2339.
Spoelstra, T. A. T.1983, *Astr. Astrophys.*, **120,** 313.

J. Astrophys. Astr. (1984) **5,** 67–78

Nonconservation of Baryons in Cosmology—Revisited

J. V. Narlikar *Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Bombay 400005.*

Abstract. The concept of the steady-state universe discussed by Hoyle & Narlikar two decades ago is revived in the light of the present discussions of the phase transition in the early big-bang universe. It is shown that with suitable scaling the bubble universe solution bears a striking similarity to the inflationary scenarios being discussed today. The currently discussed idea of cosmic baldness was also anticipated in the *C*-field cosmology of the steady-state universe.

Key words: Cosmology—steady-state universe—inflationary universe

1. Introduction

The de Sitter model of the universe was the second cosmological model to come out of general relativity. Although first proposed in 1917, it has been found to be of relevance in different cosmological scenarios. Thus it featured as the line element of the steady-state universe in 1948 (Bondi & Gold 1948; Hoyle 1948) and more recently, it has been invoked to describe the inflationary phase of the early big-bang universe (Guth 1981).

The physical motivation in each case has been different. The original de Sitter universe was supposed to be empty but had the feature of expansion based on trajectories of test particles ('motion without matter'). The steady-state theory arrived at this space-time either from the perfect cosmological principle or from a dynamical field theory while in inflationary scenarios a phase transition generates this solution. The purpose of this paper is to highlight the extraordinary similarity of ideas in the *C*-field theory of steady-state cosmology and the main features of the presently popular inflationary models.

In the mid-1960s Hoyle & Narlikar published a series of three papers (1966a, b, c; hereafter Papers 1, 2 and 3 respectively) on cosmology and cosmogony, based on the *C*-field theory of matter creation (Hoyle & Narlikar 1963). Paper 1 dealt with the concept that the strong gravitational fields of collapsed objects (black holes were still to gain currency in those days) would facilitate the creation of baryons in their vicinity. The de Sitter line element

$$ds^2 = c^2dt^2 - e^{2Ht}\,[dr^2 + r^2\,(d\theta^2 + \sin^2\theta\, d\phi^2)] \tag{1}$$

was thus seen as describing the large-scale space-time of the steady-state universe in which, on an average, matter creation in a large region keeps pace with its expansion. The overall Hubble constant H was seen to be related to the baryon mass m and the constant f, coupling the C-field to the newly created matter:

$$H^2 = \frac{4\pi}{3} Gfm^2. \tag{2}$$

Here G is the gravitational constant.

In Paper 2 Hoyle & Narlikar considered the possibility of departures for the steady state, when Equation (2) does not hold. We showed that if baryon creation was 'switched off' in a given region of space-time, that region would expand essentially as the standard Friedmann model. This steadily rarefying region would therefore appear as a 'bubble' in a denser medium, and it was argued that we live in one such bubble. In a radical departure from the steady-state assumption of Paper 2 we then argued that the coupling constant f was considerably higher (by $\sim 10^{20}$) than that given by Equation (2); that is, the Hubble constant of the denser medium outside the bubble was higher (by $\sim 10^{10}$) than that estimated at present.

The same phenomenon on a smaller scale led us to the formation of elliptical galaxies around dense massive nuclei. This was discussed in Paper 3 where it was argued that because of the observed absence of rotation in ellipticals it was hard to imagine their formation through a condensation process.

It is interesting that the ideas outlined in these three papers are now finding currency. The difficulty of low angular momentum in ellipticals is being realized as a major difficulty of the theory which seeks to form them by condensation of a gas cloud (Efstathiou &Jones 1979).The discovery of a massive collapsed object in M87 (Sargent *et al.* 1978; Young *et al.* 1978) has emphasized the possible dynamical importance of massive galactic nuclei. Recently Carr & Rees (1984) have argued that supermassive pregalactic objects might nucleate galaxies around them. It would appear that the objections of 18 years ago to the ideas of Paper 3 seem to have disappeared in the meantime.

However, it is the ideas in the first two papers that I wish to discuss here. Although the C-field cosmology worked within the framework of general relativity and thus ensured the conservation of energy and momentum, its notion of baryon nonconservation was anathema to theoretical physicists of the 1960s. Not so now! Under the grand unification programme the creation or annihilation of baryons is considered not only possible but also probable. Further, the inflationary phase in the universe makes use of the de Sitter expansion (1), coupled with the idea that our observable universe is a tiny bubble in the cosmological substratum (Guth 1981). Although the basic motivation may be different in the two cases, the striking similarity of the two cosmological models warrants taking a second look at the C-field cosmology.

In the following section the basic formalism of the C-field theory is described. In Section 3 the bubble solution is discussed with new boundary conditions relevant to the present calculations of the early universe. In the final section we compare the bubble universe with the inflationary universe and highlight the features of the latter which were anticipated by the former.

2. The *C*-field cosmology

2.1 *The Basic Formalism*

Although in his first and subsequent papers on the discussion of continuous creation of matter Hoyle (1948, 1949) had used scalar field theories, a simple and elegant formulation was given in 1960 by the late M. H. L. Pryce (personal communication).

Following the principle of Occam's razor, Pryce assumed the field to be scalar, with zero mass and zero charge, and derived its properties from an action principle. In our discussions of matter creation Hoyle & Narlikar adopted the Pryce formulation.

In the following discussion we will use the Hilbert action principle and assume that the space-time contains a set of particles a, b, . . . with masses m_a, m_b, . . . which do not interact except via gravity and the scalar C-field of Pryce. Accordingly, the action is given by

$$\mathscr{A} = \frac{1}{16\pi G}\int R\sqrt{-g}\,\mathrm{d}^4x - \sum_a \int m_a \mathrm{d}s_a$$
$$-\frac{1}{2}f\int C_i C^i \sqrt{-g}\,\mathrm{d}^4x + \sum_a \int C_i \mathrm{d}x_a^i, \tag{3}$$

where we have taken the speed of light = 1. In Equation (3) C_i stands for the derivative $\partial/\partial x^i \equiv C_{.i}$, x_a^i and s_a are the coordinates (i = 0, 1, 2, 3) and proper time along the world line of particle a, while f is a coupling constant.

The apparently simple form (3) conceals the non-trivial aspect of matter creation which becomes clear when we examine the last term in $\mathscr{A}$. If there were no matter creation, this term would be path-independent and make no contribution to the action. If, however, the world line of a has end points at A_- (annihilation) and A_+ (creation) then it contributes to $\mathscr{A}$ through the last term, an amount

$$\int_{A_-}^{A_+} C_i \mathrm{d}x_a^i = C(A_-) - C(A_+). \tag{4}$$

In other words, the C-field does not interact with matter except when it is created or annihilated.

Thus the variation $\mathscr{A} \to \mathscr{A} + \delta\mathscr{A}$ which is caused by varying the world line of particle a gives for $\delta\mathscr{A} = 0$ the geodetic equation

$$\frac{\mathrm{d}^2 x_a^i}{\mathrm{d}s_a} + \Gamma_{kl}^i \frac{\mathrm{d}x_a^k}{\mathrm{d}s_a}\frac{\mathrm{d}x_a^l}{\mathrm{d}s_a} = 0 \tag{5}$$

together with the end-point conditions

$$m_a \frac{\mathrm{d}x_a^i}{\mathrm{d}s_a} = C^i, \qquad C^i C_i = m_a^2. \tag{6}$$

The variation of C gives, for $\delta\mathscr{A} = 0$,

$$\Box C \equiv C^k{}_{;k} = \frac{1}{f}n \tag{7}$$

where n = net number of creation events in unit proper 4-volume. Each point of A_+ type contributes + 1 to n while each point of A_- type contributes – 1.

The variation of the metric gives the modified Einstein field equations

$$R^{ik} - \frac{1}{2}g^{ik}R = -8\pi G\left\{{}^{T^{ik}}_{(\mathrm{m})} + {}^{T^{ik}}_{(\mathrm{c})}\right\} \tag{8}$$

where $T^{ik}_{(\mathrm{m})}$ is the matter energy tensor for the system of particles a, b, . . . and $T^{ik}_{(\mathrm{c})}$ is

given by

$$\underset{(c)}{T^{ik}} = -f\left\{C^iC^k - \frac{1}{2}g^{ik}C^lC_l\right\}. \tag{9}$$

Although $T^{ik}_{(c)}$ has the familiar form for a scalar field it is different in one important aspect: it has a minus sign in front which (for $f > 0$) implies that the C-field has negative energy density. Under normal circumstances this would be a cause for concern from the quantization point of view. However, here the situation is somewhat different. As part of Einstein's equations the C-field is coupled to gravity and any quantum cascading down the negative energy states would result in a rapid expansion of space which acts as a control on the cascading process. In the 'steady state' the expansion of the universe just balances the cascading tendency so that $T^{ik}_{(c)}$ is finitely negative.

The divergence of Equation (8) gives

$$\underset{(m)}{T^{ik}}{}_{;k} = fC^iC^k{}_{;k}. \tag{10}$$

This is the modified conservation law of energy. If there is net creation of matter then the left hand side of Equation (10) is nonzero. From Equation (7) we see that the right-hand side is also nonzero. On the other, hand we can also get solutions with *no* net creation (or annihilation) for which

$$\Box C = 0. \tag{11}$$

As we shall see in Section 2.2 below, these solutions are analogous to the Friedmann models.

2.2 *Cosmological Solutions*

We now consider applications of this formalism to cosmology and will first discuss the steady-state solution and the bubble universe. Accordingly we take the space-time to be given by the Robertson-Walker line element

$$\mathrm{d}s^2 = \mathrm{d}t^2 - S^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right] \tag{12}$$

where $k = 0, +1$ or -1. The field equations in the case of a dust universe with density ρ become

$$\frac{\dot{S}^2 + k}{S^2} = \frac{8\pi G}{3}(\rho - \frac{1}{2}f\dot{C}^2) \tag{13}$$

$$2\frac{\dot{S}}{S} + \frac{\dot{S}^2 + k}{S^2} = 4\pi G f\dot{C}^2. \tag{14}$$

Here, in the homogeneous isotropic case C depends on t only.

Equation (7) takes the form

$$\frac{f}{S^3}\frac{\mathrm{d}}{\mathrm{d}t}(\dot{C}S^3) = n(t), \tag{15}$$

where $n(t)$ is the rate of creation of particles of mass m per unit proper 3-volume. Thus

we get from 7

$$\dot{\rho} + \frac{3\dot{S}}{S}\rho = nm. \tag{16}$$

Assuming that particles of mass m are created at rest in the cosmological substratum we get from (6)

$$\dot{C} = m, \tag{17}$$

If, however, *no* particles are created then from (15) we get

$$\dot{C} = S^{-3}. \tag{18}$$

Equations (17) and (18) represent the two different classes of solutions possible in the *C*-field cosmology.

The steady-state solution given by the de Sitter line element

$$ds^2 = dt^2 - e^{2Ht}[dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)] \tag{19}$$

belongs to the first of the two classes with

$$\rho = fm^2 = \frac{3H^2}{4\pi G}\ (= \text{constant}). \tag{20}$$

Notice that the characteristic parameter of the de Sitter space-time—the Hubble constant H—is related to the *C*-field coupling constant f and the mass m of the particle created.

If m is the typical baryonic mass (= mass of the proton, say) then we can express f in terms of the present value of Hubble constant, with the help of Equation (20):

$$f \simeq 1.6 \times 10^{19}\, h_0^2\, g^{-1}\, \text{cm}^{-3}. \tag{21}$$

Although we obtained the value of f above using the observed value $H = 100\ h_0$ km s^{-1} Mpc^{-1}, the actual cosmological reasoning is the reverse: it is the value of f that determines how fast the steady-state universe should expand.

It is also worth pointing out the difference of interpretation of the energy tensor in this solution and in the de Sitter model as obtained by de Sitter, and in the inflationary models. In de Sitter's version the space-time was considered empty but the Einstein equations contained the λ-term. In the inflationary scenario the phase transition gives rise to a λ-term. However, it appears on the right-hand side of the equations

$$R_{ik} - \frac{1}{2}g_{ik}R = -\lambda g_{ik}. \tag{22}$$

This right hand side describes the energy tensor of a cosmic fluid with density $\rho = \lambda/8\pi G$ and negative pressure $p = -\rho$. It is interesting to recall that W. H. McCrea. (1951) gave a similar interpretation to cosmic fluid in the steady-state universe.

In the second class of solutions with no creation we get the following equation for the scale factor *S:*

$$\dot{S}^2 = -k + \frac{A}{S} - \frac{B^2}{S^4} \tag{23}$$

where A and B are constants related to ρ and $\dot{C}$ by

$$S^3\rho = \frac{3A}{8\pi G}, \qquad S^3\dot{C} = \frac{B}{(4\pi Gf)^{1/2}}. \tag{24}$$

Notice that Equation (23) describes a non-singular universe. For $k = 0$, it has the explicit solution

$$S = (\alpha t^2 + \beta)^{1/3}, \qquad \alpha = \frac{9}{4}A, \qquad \beta = \frac{B^2}{A}. \tag{25}$$

At large t this behaves like the Einstein-de Sitter solution.

In an earlier paper, Narlikar (1974) had discussed this solution as arising from explosive creation at a single epoch t_0 so that $n(t) \propto \delta\,(t - t_0)$. This model thus provided a nonsingular discussion of the big-bang event.

In Paper 2, however, Equation (23) was supposed to arise when creation was spontaneously 'switched off' in a given space-time region. The switch-off would occur if the creation condition (6) failed to be satisfied in a finite region due to local fluctuations. In that event the region would expand according to Equation (23) in a de Sitter background given by Equation (19), somewhat like a low-density air-bubble in a denser liquid. For a reason outlined below, it was suggested, however, that the outer steady state background corresponded to a Hubble constant several orders of magnitude smaller than the currently estimated value 100 h_0 kms^{-1} Mpc^{-1}.

In Paper 1 it was argued that for the creation of particles of mass m_a the condition (6) must be satisfied. In a universe containing a uniform distribution of massive objects the possibility emerges that in the vicinity ofa massive body the magnitude of $C^i\,C_i$ is raised above the average cosmological value. Hence, if the average value of $C_i\,C^i$ is below the required threshold m_a^2, but rises above it near a typical massive body then creation of particles would take place only near the body. Thus according to Paper 1, the steady state is maintained by creation of matter around existing masses.

However, the value of f given by Equation (21) was found to be too small to explain explosive outpouring of particles near active galactic nuclei. In order to explain such events as the origin of high-energy cosmic rays it was necessary to raise f by a factor $\sim 10^{20}$ above the value given by Equation (21). The relation (20) then implied that the Hubble constant must be larger than its presently observed value by a factor $\sim 10^{10}$. In other words, the characteristic cosmological timescale of the steady-state model turned out to be ~ 1 yr rather than $\sim 10^{10}$ yr.

It is in such a universe that the bubble is formed by the spontaneous cut-off of the creation process. The present Hubble constant of 100 h_0 km s^{-1} Mpc^{-1} corresponds not to the steady-state solution but to the Friedmann-like solution (25). The steady state solution only provides the initial conditions for the formation of the bubble to which our direct observations of the universe have so far been confined.

3. The early universe

3.1 *The Creation of Relativistic Particles*

We now consider certain modifications in the above picture to take into account the radiation-dominated early universe. We will consider epochs at which baryons as well

as leptons obey the relativistic approximation (Narlikar 1983). Following the cosmological principle we assume, as before that C depends on t only. We will, however, modify the single-particle creation scenario which led to conditions (6) and assume that two particles of equal and opposite 3-momenta $\pm P$ are created at each point A_+. Then the action principle gives

$$\dot{C} = 2\sqrt{P^2 + m^2} \equiv 2E. \tag{26}$$

Here m is the rest-mass of each particle created and E its energy. In the relativistic approximation

$$E \simeq P \gg m. \tag{27}$$

In the formalism to be described below, condition (27) is assumed to hold although it is not difficult to develop a similar theory for the nonrelativistic case.

If CP is conserved in the creation process, the two created products will form a particle antiparticle pair. If CP is broken both particles could be of the same type. Equation (26) ensures, however, that the total energy and momentum are conserved between the C-field and the two created particles. Since C is a scalar field the spin is conserved by ensuring that the created particles carry opposite spins.

In the two-particle creation at a given place the symmetry of isotropy is spontaneously broken. However, since the directions of motion of the created pair are random, the symmetry is hidden. It is therefore correct to assume that on a macroscopic scale C still depends on t only.

It is convenient to write $\dot{C}$ as a function of the scale factor S, Thus we will write Equation (26) in the form

$$\dot{C} \equiv g(S) = 2P \tag{28}$$

so that at the epoch of scale factor S the created particles have momentum $g\,(S)/\,2$. After creation, the momentum decreases according to the law

$$P \propto 1/S. \tag{29}$$

Let $N\,(P,\,S)\,\mathrm{d}P$ denote the number density of particles at epoch of scale factor S with momenta in the range P and $P + \mathrm{d}P$. The pressure $p(S)$ and energy density $\varepsilon(S)$ of the cosmological material are then given by

$$p(S) = \frac{1}{3}\varepsilon(S) = \frac{1}{3}\int_0^\infty PN(P, S)\mathrm{d}P. \tag{30}$$

Because of the relation (29), the function $N\,(P,\,S)$ satisfies the equation

$$\frac{\partial N}{\partial P}\cdot\frac{P}{S} = \frac{\partial N}{\partial S} + \frac{2N}{S} \tag{31}$$

which integrates to

$$N(P, S) = \frac{1}{S^2}F\,(PS). \tag{32}$$

The arbitrary function $F\,(PS)$ is related to how the particles are created. In general we expect it to have a step-function type discontinuity at $P = g(S)/2$ to take into account the injection of new particles according to Equation (28). Because of the relation (29), if

in an expanding universe

$$\frac{\mathrm{d}}{\mathrm{d}S}[Sg(S)] < 0, \tag{33}$$

then the particles which are already in existence at epoch S will have momenta greater than $g(S)/2$. Hence

$$N(P, S) = \frac{1}{S^2} F(PS)\theta\left[P - \frac{1}{2}g(S)\right]. \tag{34}$$

Here θ is the Heaviside function.

Likewise, if

$$\frac{\mathrm{d}}{\mathrm{d}S}[Sg(S)] > 0, \tag{35}$$

then

$$N(P, S) = \frac{1}{S^2} F(PS)\theta\left[\frac{1}{2}g(S) - P\right]. \tag{36}$$

We will refer to the two cases leading to Equations (34) and (36) as Cases 1 and 2 respectively.

Consider now the relation (10). This becomes in the present case

$$\frac{\mathrm{d}}{\mathrm{d}S}(\varepsilon S^3) + 3pS^2 = f\dot{C}\frac{\mathrm{d}}{\mathrm{d}S}(\dot{C}S^3). \tag{37}$$

Using Equations (30) and (34) we get for Case 1

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}S}(\varepsilon S^3) + 3pS^2 &= \frac{\mathrm{d}}{\mathrm{d}S}\int_{\frac{1}{2}g(S)}^{\infty} PSF(PS)\,\mathrm{d}P + \int_{\frac{1}{2}g(S)}^{\infty} PF(PS)\,\mathrm{d}P \\ &= -\frac{1}{2}g(S)F\left\{\frac{1}{2}Sg(S)\right\}\frac{\mathrm{d}}{\mathrm{d}S}\left[\frac{1}{2}Sg(S)\right].\end{aligned}$$

The right hand side of Equation (37) is simplified by using Equation (28). Case 2 can be similarly handled and we get in the two cases the final result

$$F\left\{\frac{1}{2}Sg(S)\right\} = \pm 4fS^2\frac{3g(S) + Sg'(S)}{g(S) + Sg'(S)}. \tag{38}$$

This minus sign on the right-hand side corresponds to Case 1 and the plus sign to Case 2. In either case, the particle distribution function is determined if $g(S)$ is specified, or vice versa. Once $g(S)$ is determined the function $S(t)$ is fixed by the Equations (13) and (14).

3.2 *The Steady-State Solution*

Consider the simple example where $F(x) \propto x^2$, say,

$$F(x) = \lambda x^2, \qquad \lambda = \text{constant} > 0. \tag{39}$$

Here, F is proportional to the geometrical volume of the momentum space. Then

Equation (38) gives (taking the positive sign) after simple integration,

$$g(S) = \text{constant} = \sqrt{48f/\lambda} = 2P_0 \text{ (say)}. \tag{40}$$

This leads us to the steady-state line element with

$$\frac{1}{3}\varepsilon = p = \frac{12f^2}{\lambda} = fP_0^2, \qquad N = 4fP_0 \tag{41}$$

$$H^2 = \frac{8\pi Gf}{3}P_0^2, \tag{42}$$

where P_0 is the momentum of the particle created. At any given epoch the momentum distribution of the created particles follows the distribution function

$$N(P, S) = \lambda P^2, \qquad \lambda P \leqslant P_0. \tag{43}$$

This distribution function presupposes that particles are created at random at relativistic speeds, but once they are created they do not collide and alter their momenta. Thus the relation (29) denotes the way in which each particle loses its momentum with expansion.

In the actual situation prevalent in a high-density universe, the no-collision condition will be satisfied provided the collision rate Γ_C of various particles is less than the rate of expansion of the universe, *viz.*, H. We will assume that $\Gamma_C \ll H$.

What should be the value of H? From Equation (20) H is determined by f and m. However, rather than specify f and m first, we will proceed in an empirical manner, and use Occam's razor.

First we note that the only constants at the disposal of a gravity theory like general relativity are G and c. We may also add $\hbar$ to the list if we wish to include the effects of quantum theory. From G, $\hbar$ and c a time-scale emerges which is given by

$$\tau_P = \sqrt{G\hbar/c^5}. \tag{44}$$

For $\tau \lesssim \tau_p$ the discussion of various phenomena must proceed via quantum rather than classical gravity.

Work by several authors (see for example Atkatz & Pagels 1982; Brout *et al.* 1980, Vilenkin 1982; Padmanabhan 1983) has shown that empty flat space-time is unstable to quantum fluctuations and that dynamical discussions of such fluctuations lead inevitably to matter creation and C-field like (negative energy) terms in the T^{ik}. Therefore we could argue that if our steady-state solution evolved this way, the resulting H would be comparable to τ_p^{-1}. Accordingly we set

$$H = \beta\tau_p^{-1}, \qquad \beta \lesssim 1. \tag{45}$$

The condition $\beta < 1$ is necessary to ensure that our classical description has some validity.

Next we will conjecture about the created mass m, again in a heuristic way. The present ideas in grand unification theories (GUTs) suggest that the massive X-boson plays a crucial role in baryon-nonconservation. We therefore identify its mass m_x with m and write Equation (20) as

$$H^2 = \frac{4\pi G}{3}fm_x^2. \tag{46}$$

From Equations (45) and (46) we are able to determine f:

$$f = \frac{3c^5\beta^2}{4\pi m_x^2 G^2 \hbar} = \frac{3}{4\pi}\frac{c^4\beta^2}{G\hbar^2}\frac{m_p^2}{m_x^2}, \tag{47}$$

where

$$m_P = \sqrt{\frac{c\hbar}{G}} \tag{48}$$

is called the Planck mass. We will consider the numerical values of f, m_x, m_p *etc.* later.

3.3 *The Growth of a Bubble*

In this highly dense steady-state universe we next consider the idea that creation is switched off in a finite region which subsequently expands as a bubble. To estimate the physical size of such a bubble we proceed as follows.

How long does an X-boson survive after creation? Its lifetime may be estimated on dimensional arguments to be $\tau_x = \Gamma_x^{-1}$ where

$$\Gamma_X = \gamma\frac{m_X c^2}{\hbar} \tag{49}$$

and γ is a dimensionless constant. To estimate γ we suppose that there are altogether g effective degrees of freedom in the cosmological mixture of particles. This quantity g is determined in the usual way by

$$g = g_b + \frac{7}{8}g_f \tag{50}$$

where g_b = total number of boson spin states and g_f = total number of fermion spin states. Then we expect that

$$\gamma = \alpha g \tag{51}$$

where α is a constant estimated by some GUTs in the range 10^{-2} to 10^{-5}.

The 'switching off' of creation may be linked to the disappearance of X-bosons. Thus, during the lifetime τ_x of the created X-boson a characteristic cosmological 3-volume of linear size c/H will expand to

$$L = \frac{c}{H}\exp(\tau_X H) = \frac{c}{H}\exp\left(\frac{\beta\tau_X}{\tau_P}\right). \tag{52}$$

This is the size of the bubble at the onset of its expansion as a Friedmann universe. To estimate its present size we use the fact that during expansion the scale-factor increases inversely as temperature. The radiation temperature at the Planck epoch was

$$T_P = \frac{m_P c^2}{k} \tag{53}$$

where k = Boltzmann's constant. If the present temperature is given by T_0, the present size is given by

$$L_0 \simeq \frac{c}{H}\exp\left(\frac{\beta\tau_X}{\tau_P}\right)\frac{T_P}{T_0} = \frac{c}{H}\cdot\frac{T_P}{T_0}\cdot\exp\Sigma, \tag{54}$$

say.

In Equation (54) the quantities T_p, τ_p, c, H are determined in terms of elementary constants c, G, $\hbar$, *k etc.*, while $T_0 \simeq 3K$ is given by observations. Using Equations (49) and (51) we write

$$\Sigma = \frac{\beta \tau_X}{\tau_P} = \frac{\beta}{\gamma} \frac{m_P}{m_X} \simeq \frac{\beta}{\alpha g} \frac{m_P}{m_X}. \tag{55}$$

We will first estimate Σ from Equation (54) by using $L_0 = 10^{28}\ h_0^{-1}$ cm· Then we have

$$\Sigma = \ln \left[\frac{L_0 H}{c} \frac{T_0}{T_P} \right] \simeq 67 - \ln h_0 \tag{56}$$

where the current uncertainty of the value of Hubble constant suggests that | In h_0 | < 1. We will therefore ignore it.

In Equation (55) set $\beta \simeq 1$ and $g \simeq 200$ as the approximate numbers of degrees of freedom of all particle species in the early universe. Then we get

$$m_X \simeq 7.5 \times 10^{-5}\, \alpha^{-1}\, m_P \simeq 7.5 \times 10^{14}\, \alpha^{-1}\ \text{GeV}. \tag{57}$$

Note that this limit is consistent with the present lower bounds on the proton lifetime.

If all the GUT parameters were fully determinable, we could have had more reliable estimates of m_x, g, α *etc.* Also, the relation (51) could then be stated more accurately. The current work suggests that since $\alpha < 1$ in Equation (57) the mass of the X-boson is expected to be higher than 7.5×10^{14} GeV. Also, since we expect $m_p > m_x$, α should not be lower than $\sim 10^{-4}$. This requirement comes from the consistency of the overall cosmological scheme presented here and could be compared with the values of α given by various GUTs.

4. A comparison with the inflationary scenarios

The exponential term exp Σ in Equation (54) is analogous to the inflationary term in the big-bang cosmology. That the value of Σ is the same (within small calculational uncertainties) in the two pictures may come as a surprise; but on closer examination this is to be expected. The reason is as follows.

In our bubble picture as in the standard Friedmann cosmology the rate of expansion $\sim t^{2/3}$ is comparatively slow. As a result, the present observable universe of linear dimension $\sim 10^{28}$ cm has to come out of a relatively large region of the early universe. In the inflationary scenarios this largeness is achieved by a temporary de Sitter like phase which is associated with phase transition. In the present model the background universe is always in de Sitter (steady-state) form but the growth of a bubble is associated with the switching off of the creation process. The timescale for switch-off is linked with the disappearance of X-bosons in a given volume. A volume of cosmological dimension c/H inflates during this time to a linear size c/H exp Σ. At this stage the bubble formation is complete and the bubble expands as the Einstein-de Sitter model would.

The picture presented here is still phenomenological since it does not discuss the dynamical aspects of how the creation is switched off. The constant Σ is in principle calculable if a fully developed grand unified theory and *C*-field theory is available. The numerical estimates given in Section 3.3 suggest that a self-consistent detailed theory may be possible.

The above weakness apart, the present scheme offers certain advantages over the standard inflationary scenario. The background universe is singularity free and the bubble itself starts from a well-defined initial state. The background de Sitter space-time is free from particle horizons and there is thus no impediment towards its achieving a highly homogeneous state. In fact, as discussed within the old *C*-field theory (Hoyle & Narlikar 1963), the newly created matter serves to homogenize the universe and to wipe out any earlier 'memories' of inhomogeneities. This idea has been suggested anew recently by Barrow & Stein Schabes (1983) under the concept of "cosmic no-hair conjecture".

Since the de Sitter space-time is flat in the spatial sense ($k = 0$), the emerging bubble is also spatially flat. Thus the density parameter

$$\Omega = \frac{8\pi G\rho S^2}{3\dot{S}^2} \tag{58}$$

will be very close to unity. The departure from unity is given by the last term of Equation (23) for the case $k = 0$. This term carries the rapidly diminishing negative *C*-field energy and is negligible by the present epoch. Thus this model would predict $\Omega = 1$ to a high degree of accuracy.

Finally, because of its nonsingular beginning this model holds out hopes of relating the behaviour of the background steady model to investigations of quantum cosmology.

Acknowledgement

The author thanks Dr T. Padmanabhan for critical discussions during the preparation of this manuscript.

References

Atkatz, D., Pagels, H. 1982, *Phys. Rev.*, **D25,** 2065.
Barrow, J. D., Stein-Schabes, J. 1983, preprint.
Bondi, H. Gold, T. 1948, *Mon. Not. R. astr. Soc.*, **108,** 252
Brout, R., Englert, F., Frere, J.-M., Gunzig, E., Nardone, P., Trunin, C. 1980, *Nucl. Phys*., **B170,** 228.
Carr, B., Rees, M. J. 1984, *Mon. Not. R. astr. Soc.*, **206,** 801.
Efstathiou, G., Jones, B. J. T. 1979, *Mon. Not. R. astr. Soc.,* **186,** 133.
Guth, A. 1981, *Phys. Rev*., **D23,** 347.
Hoyle, F. 1948, *Mon. Not. R. astr. Soc.*, **108,** 372.
Hoyle, F. 1949, *Mon. Not. R. astr. Soc.*, **109,** 365.
Hoyle, F., Narlikar, J. V. 1963, *Proc. R. Soc.*, **A 273,** 1.
Hoyle, F., Narlikar, J. V. 1966a, *Proc. R. Soc.*, **A 290,** 143 (Paper 1).
Hoyle, F., Narlikar, J. V. 1966b, *Proc. R. Soc.*, **A 290,** 162 (Paper 2).
Hoyle, F., Narlikar, J. V. 1966c, *Proc. R. Soc.*, **A 290,** 177 (Paper 3).
Lindley, L. 1981, *Nature*, **291**, 391.
McCrea, W. H. 1951, *Proc. R. Soc.*, **A206,** 562.
Narlikar, J. V. 1974, *Pramana*, **2,** 158.
Narlikar, J.V.1983, *Introduction to Cosmology*, Jones and Bartlett, Boston.
Padmanabhan, T. P. 1983, *Phys. Lett.*, **93A,** 116.
Sargent, W. L. W., Young, P. J, Boksenberg, A., Shortridge, K., Lynds, C. R., Harwick, F. D. A. 1978, *Astrophys. J.*, **221,** 731.
Vilenkin, A. 1982, *Phys. Lett*., **117B,** 25.
Young, P.J.,Westphal, J. A., Kristian, J.,Wilson, C. P., Landauer, F. P.1978, *Astrophys. J.*, **221,** 721.

J. Astrophys. Astr. (1984) **5,** 79–98

Cosmology: Myth or Science?

For the Golden Jubilee of the Indian Academy of Sciences, representing a culture which has investigated cosmology for four millennia

Hannes Alfvén *Royal Institute of Technology, Stockholm, and University of California, San Diego*

1. Pre-Galilean cosmologies

1.1 *Ancient Cosmological Myths*

Cosmology began when man began to ask: What is beyond the horizon and what happened before the earliest event I can remember? The method of finding out was to ask those who had travelled very far; they reported what they had seen, and also what people they had met far away had told them about still more remote regions. Similarly, grandfather told about his young days and what his grandfather had told him and so on. But the information was always increasingly uncertain the more remote the regions and the times.

The increasing demand for knowledge about very remote regions and very early times was met by people who claimed they could give accurate information about the most distant regions and the earliest times. When asked how they could know all this they often answered that they had direct contact with the gods, and got revelations about the structure of the whole universe and how it was created. And some of these prophets were believed by large groups of people. Myths about the creation and structure of the universe were incorporated as essential parts of religious traditions.

In different cultures, the mythologies became drastically different, depending on the way the philosophical thinking developed and on the personalities of great prophets.In several of the world religions, both the universe and the gods were believed to the eternal; in others, the gods or one God created the universe. In some religions, there is no conflict between these views; initially the universe was identical with a god and the different members of his body developed into the different parts of the universe. In India during the Vedic period, the god Purusa was initially identical in the whole world, and part of his body became the Earth, other parts the Heaven; the Sun formed from his eyes and the Moon from his soul. In other philosophical-mythological schools, both Heaven and Earth are regarded as gods and as parents of gods. Sometimes one god—in India, Agni or Soma or Rudra—and sometimes all gods together are said to have generated or created the whole universe.

In Rigveda, there is a remarkable poem telling that originally

> "There was neither Aught nor Naught, no air nor sky beyond".

There was only

> "A self-supporting mass beneath, and energy above.
> Who knows, who ever told, from whence this vast creation rose?
> No gods had yet been born – who then can e'er the truth disclose?"

During the more than three millennia which have passed since the Vedic period, Indian mythology has developed a jungle of co-existing creeds, in part absorbed from neighbours, and in part from earlier cultures which had collapsed. The sophisticated mythological philosophy is, perhaps, somewhat less chaotic. There is a general tendency to consider the evolution of the universe as well as the human society to be periodic. Indeed, there is a hierarchy of periods. A golden age, followed by a silver, a bronze and the present iron age (Kaliyuga) forms a Mahayuga of 54,000 years. A number of Mahayugas forms a larger period, and so on in steps up to the Kalpa or the day of Brahma, which is 4×10^9 years. This is only half an order of magnitude smaller than what according to the Big-Bang hypothesis should be the 'age of the universe'. However, there are 365 Brahma days in one Brahma year, and Brahma lives for 100 years, so the ancient Indians used time units which were four orders of magnitude longer than in the Big Bang. (Of course, when Brahma dies after his 100 years, he is immediately reborn). Indian estimates of the size of the world were not so fantastic. Sometimes the figure10,000 yojanas is given, which means less than half the distance to the Moon.

The Mediterranean-Middle East thinking was initially as closely related to the Indian mythologies as Greek, Latin, and Persian are related to Sanskrit. The way of life of the people speaking these languages was also similar. The battle of Kuruksetra and the battle before the walls of Ilion took place at about the same epoch and were fought in a similar way. The heroes spent day after day fighting, and at dusk they went back to their camps, drinking and bragging. Their gods took a decisive part in the fight. (By the way, in Scandinavian mythology, the Vikings who fell in battle came immediately to Valhalla, where they enjoyed the same type of daily life).

In the same way, the Mediterranean mythology was initially similar with a golden, silver, bronze and iron age in sequence. However, the Greek cosmological philosophy which took the lead at the Greek cultural explosion around 500 B.C. did not develop like the Indian. First of all, the world remained very limited in time. Indeed, the guesses of the age of the world considered periods of some thousand years, which is only one micro-kalpa. On the other hand, the estimates of the size of the universe were not so different.

Not all the early cosmologies were so intimately connected with religion. The sages of China had no preconceived theories, and seem to have based their cosmological thinking more on phenomena which they observed. But the observations they could make did not suffice for any certain conclusions, and any more elaborate scenarios were no less speculative than those which originated from divine revelation to prophets.

1.2 *Buddhist Cosmology*

Buddhism developed views on cosmology which were drastically different from the other Indian cosmologies. As Buddhism is basically an agnostic religion, it does not deny the possible existence of gods, but it does not claim that there are any. The existence of gods is irrelevant to the aim of Buddhism, which is to find the right way to salvation, to the annihilation of desire, to the state of Nirvana.

As a logical consequence of this, when the Buddha was asked whether the universe was eternal or created he is reported to have answered in his characteristic style:

It is wrong to say that it is eternal.
It is wrong to say that it is created.
It is wrong to say that it is both eternal and created.
It is wrong to say that it is neither eternal nor created.

Perhaps this is an echo from the quoted Rigveda poem which probably derives from one millenium earlier: As man got his knowledge about the early states of the universe from prophets who got their knowledge directly from the gods, then no information could be gained about the epoch when the gods had not yet been born. Similarly, as the Buddha did not believe in gods—or in any case, did not care much about them—there was no possibility to get information about early cosmology.

Perhaps one could also find an echo two millenia later, when Descartes proclaimed: *De omnibus est dubitandum* (We should question everything). However, this is not altogether correct because Descartes had also inherited the Galilean scientific tradition according to which controversial issues should be settled by reference to experiment and observation. But there does not seem to be any basic logical conflict between Descartes and the agnosticism of Rigveda and the Buddha.

1.3 *Rise of Mathematics*

1.3.1 *The Pythagoreans*

A new element in the cosmological discussion was introduced by the rise of science and natural philosophy in Greece as a part of the cultural explosion around 500 B.C. The Greeks had absorbed astronomical knowledge both from the Mesopotamian and Egyptian cultures, and, as we have mentioned already, their mythology was genetically related to the Indian.

The new element consisted of the rise of geometry, which to a large extent derived from Egypt, where it was of practical importance for land surveying. The Greeks developed this to the still unsurpassed masterpiece of logically stringent structure which we know as Euclidean geometry. It is questionable whether the beauty of the theorem of the regular polyhedrons will ever be surpassed. By a simple discussion which anyone can understand in a few minutes, the *a priori* surprising conclusion is reached that there are five and only five such bodies.

Strongly connected with this, a much wider breakthrough of new thinking was achieved by the Pythagoreans. They demonstrated that the basis of musical harmony was simple ratios of integers. It is quite understandable that this led to a philosophical optimism. The Pythagoreans tried to incorporate astronomy and cosmology as well into their philosophy. They claimed that astronomy should be to the eye what musical harmony was to the ear.

This was indeed a revolutionary idea. It was the first attempt to construct a comprehensive mathematical scheme of cosmology and to work out a synoptic view of the universe as a whole.

One may say that its basic principle is that because the world was created by the gods, there must be a sublime order in its basic structure – even if many regrettable local disorders were obvious. According to the Pythagoreans, the most ‘perfect’ geometrical figure is the circle, and the most ‘perfect’ of all solid bodies is the sphere. *Ergo* the Earth must be a circular disk or a sphere, surrounded by a number of crystal spheres, on which

the planets and the stars were located. Further, the most perfect motion was uniform motion. *Ergo* the crystal spheres must rotate with uniform velocity. This was necessary for the 'harmony of the spheres.'

1.4 *Relation between Theory and Observation*

Neither the Pythagoreans nor Plato cared very much for a comparison with observations. The Pythagoreans formed a secret society with no real contact with the rest of Greek society. Indeed, traitors were severely punished. The rules of Plato's Academy included: "Let none who has not learnt geometry enter here," and he advised all scholars to "concentrate on the theoretical side of their subject and not spend endless trouble over physical measurements to the neglect of theoretical problems."

This was in conformity with the general attitude of the intellectual aristocracy in Greece. The belief was that technology,including technological innovation, ought to be largely relegated to the lower classes, especially to slaves. It was degrading for a philosopher to get his hands dirty.

It has been suggested that this cleft between sophisticated theoretical thinking and practical work, including experiments, was the basic reason why the highly advanced science in ancient Greece never led to the scientific breakthrough which took place in Europe two millenia later.

1.4.1 *The Ptolemaic System*

When, in spite of Plato, observations began to attract interest, the Pythagorean cosmology seemed to be confirmed by observations in one respect: the outermost crystal sphere, the one on which the stars were fixed, did apparently move with a constant speed. This was just what could be expected because this sphere was the outermost one, closest to where the gods lived, and hence most divine. Unfortunately, the theory did not agree so well with observational results when applied to the planets, including the Sun and Moon. The Sun and the Moon sometimes moved more to the north, sometimes to the south, and a planet like Jupiter sometimes reversed its motion in relation to the stars.

It was obvious that something was wrong. But the basic principles—uniform motion and perfect geometrical figures—were sacrosanct and could not be given up even if they were in conflict with observations. Instead, very ingenious auxiliary ideas were forwarded. Planets are not directly fixed on the crystal spheres, but each is fixed on a small circle, an epicycle, which moved with a constant velocity with its centre fixed on the crystal sphere. For a time such theories looked promising, but better observations demonstrated that they were not accurate. The reaction of the scientists was to try to patch up an old fiction instead of asking themselves whether, after all, its basis was laid in truth. They tinkered instead of recreating. Hence, increasingly complicated additions to the system were made.

The result of this was the Ptolemaic system, which was worked out in the third century A.D. No less than 54 epicycles, eccentrics, *etc*., had been introduced. But at the same time, as it became more complicated, it became more sacrosanct. When an avalanche of religious fanaticism put the classical culture into a deep freeze for more than a millenium, it did not develop very much, and age made it still more sacrosanct. Criticism was dangerous, and it was a rare exception when the famous astronomer,

King Alphonse X of Castile, complained about its degree of complexity: "Had I been present at the creation, I could have rendered profound advice."

1.4.2 *Astronomy*, *Astrology*, *and Myth*

This mathematically based cosmology did not come into serious conflict with the ancient myths. They became to a certain extent incorporated, and a jungle grew up of mathematics, astronomy, astrology, and myths from many earlier cultures. Gods and spirits of all kinds began to settle on the crystal spheres, soon causing a population explosion. For example, one group of constellations depicts how Perseus saved Andromeda from Medusa, whose terrible head is represented by a variable star. Still more dramatic is the giant hunter Orion, who, followed by the Big Dog and the Small Dog, lifts his club against the red-eyed Bull.

The early motion of the Sun along the ecliptic was illustrated by a number of sun-myths. For example, when Heracles fought a bull and later a lion, this is thought to represent the Sun's entry—on its walk along the zodiac—into the constellations Taurus and Leo. Another sun myth, in which Delilah cuts Samson's hair from which his strength derives, tells us that in the fall, when the Sun enters the constellation Virgo, its rays lose their heating power and he becomes a captive for half a year, until spring, when he has regained his force.

This chaotic conglomeration of mathematics, astronomy (including cosmology), and myths from many religions has turned out to be a permanent ingredient in our culture. Today, after more than 2000 years, it has as much vitality as ever. Newspapers and periodicals usually have astrological columns; every jeweller sells pendants and pins with signs of the zodiac. From the point of view of our commercialized society, there are many more dollars in astrology than in astronomy.

1.5 *Creation Ex Nihilo Versus Ungenerated Universe*

The rise of the monotheistic religions meant that one of the gods became more important than the others; He became the Pharaoh, the dictator of the Heavens, God with capital 'G'. He also became more important than the material world. He alone was eternal. He was not a product of the evolution of the universe, as in Rigveda. On the contrary, the whole world was a secondary structure created by Him. In the Bible the creation takes six days. It still has the character of bringing order into a pre-existing chaos. It was not until the first few centuries A.D. that creation was thought of as the production of the world *ex nihilo* (but this is never taught in the Bible). God had now become powerful enough to create the whole world by just pronouncing some magic words, or by his will-power.

Monotheistic religions have often a tendency to become fanatic. Certainly Christendom did so, at least during some periods. Tertullian said *Credo quia absurdum* (I believe because it is absurd). Hence there should be no serious attempt to reconcile religion and science.

In the Aristotelian philosophy the material world was 'ungenerated and indestructible', a view which is not in conflict with some of the Rigvedic views. It was not until medieval times that Aristotle's views were accommodated to the idea of creation *ex nihilo* essentially by Saint Thomas, who remodelled the Aristotelian philosophy in accordance with the requirements of ecclesiastical doctrine.

It is of interest to remember that even Saint Thomas confessed that *reason* could only be satisfied with the assumption that the world had no.beginning. "The doctrine of a beginning or the non-eternity of the world is to be received *sola fide*, as an act of pure faith in deference to authority."

Not even the monotheistic religions were fatal to the old myths. The 'pagan' gods changed their names—some became devils, others became saints. In Italy, one pays homage to saints in the same places in the woods where earlier a nymph or a dryad used to live. They have only acquired more modern dress. Midwinter solstice was in 'pagan' times the festival of the Sun-god, and a fertility Moon-goddess was worshipped at the first full moon after the vernal equinox. These nice old traditions remain, even today, although with a modified meaning.

The ancient belief that the wandering stars governed the life of men was conserved and developed further. Astrology, mythology and religion formed an increasingly complicated, fascinating structure. The basic conflict between an omnipotent God and the old belief that our destiny is governed by the stars was patched over by the formula:

Astra regunt hominem, *sed regit astra Deus*
(Stars rule men, but God rules the stars)

The scientific basis of the Ptolemaic system, *viz*., that the stars move according to certain mathematical laws, was forgotten.

1.6 *Myth Versus Science; Mathematical Myths*

The Ptolemaic system was initially a quite attractive theory but, during the centuries, it developed into a sacred and rigid structure increasingly impotent in incorporating new discoveries. The reason for this was that fundamentally the approach was not scientific but mythological.* The basic ideas were the perfect geometrical figures and uniform motion. The idea of building a world system on such general principles represented great progress, because earlier it had been generally believed that events in the world were governed by the will or the whimsies of gods. The Ptolemaic system did not necessarily question that the celestial system was created by the gods, but it claimed that they must have acted according to certain philosophical or mathematical principles which it was possible to analyze and understand. A sufficiently sophisticated mathematician might find out what the divine mathematic principles were.

The Ptolemaic system originated from what we may call a *mathematical myth.*

The Pythagorean philosophy had a logical beauty which could well be called 'divine.' By pure abstract thinking the theoreticians claimed to have discovered the principles according to which the gods acted when they created the world. And when these principles were found, it was held that the world must be structured according to them. In a way, the demiurges had no choice; it was not even necessary that they existed. But not even observations of physical reality were necessary. The system was based on divine inspiration or logical-mathematical necessity. If Galileo claimed that in his telescope he saw celestial bodies or sunspots which *a priori* do not exist, it was his telescope and not the theoretical system which was wrong.

* It is a semantic question whether a model initially deriving from 'divine inspiration' should be called a myth even if it includes philosophical and mathematical elements. Some would no doubt prefer to call it, for example, '*a priori* metaphysics'.

But long before Galileo, new ideas had appeared in Islamic culture, which took the lead in science less than 100 years after the Hegira. In the twelfth century, Avaroës from Cordova claimed that the world is eternal—not created, but in a state of evolution (Singer 1959), a view which is similar to the hierarchical cosmology of today. In his impressive treatise Mugadema Ibn Khaldun (around 1400 AD) dared to oppose Plato's view that the world could be explored by logical thinking alone. Indeed, he said that "logic is not a safe way of thinking, because of its tendency towards abstraction and its remoteness from the tangible world" (Baali & Ward 1981). This is similar to Bertrand Russell's warning half a millenium later against 'unaided reason'. Ibn Khaldun claimed explicitly that *cosmology must be based on observations.*

1.7 *The Copernican System*

The Ibn Khaldun idea had to hibernate for two hundred years until it reappeared in Europe, where it led to the well-known crisis which resulted in the victory of the Copernican heliocentric system (but after some time the latter had to abdicate in favour of a 'galactocentric' system).

1.8 *The Tycho-Brahe Compromise*

During the fight between the geocentric and the heliocentric cosmologies, an ingenious compromise was proposed by Tycho Brahe. His cosmology accepted that all the planets moved around the Sun, but the Sun (together with all the planets) moved around the Earth. (The Moon also moved around the Earth.) In this way he satisfied the observations which indicated that the planets moved around the Sun, but he conserved the sacrosanct geocentric cosmology. The Tycho-Brahe cosmology agreed with observations about as well as the Copernican cosmology. But it soon turned out that the basic issue was another. It was the survival or defeat of a sacrosanct myth. The myth had been sterile. It had not been able to predict a single new phenomenon which later was confirmed by observation.

2. The introduction of the telescope

2.1 *Empirical Approach; Newton*

The real importance of the Copernican revolution was not that a geocentric cosmology was replaced by a heliocentric one, but that the new approach to cosmology was based on observations, not on mathematical-philosophical principles. The Ptolemaeans had never clearly understood that—as Bertrand Russell puts it—"mathematics is the science in which you never know what you are talking about, if what you are saying is true". Indeed, "it deals with hypothetical entities and it is only concerned with their relationships to each other, being indifferent to whether anything in the real world corresponds". This means that mathematics is suitable to give prestige to any idea, but if the idea is a myth, mathematics can turn it into a 'mathematical myth', but not guarantee that it has anything to do with reality.

The observational approach was essential because Galileo's introduction of the telescope led to a rapidly increasing avalanche of observational facts. Galileo, Kepler and Newton established new laws of nature which accounted for the observational facts with a surprising accuracy. From them it was possible to predict several phenomena which later were observed. At the same time they had a mathematical 'beauty' which perhaps even surpassed that of the old laws. But it was clearly understood that they were not sacrosanct. Newton said: '*Hypotheses non fingo*' (I do not make any hypothesis.) However, they remained unchallenged until the beginning of this century. The transition from a heliocentric cosmology to a galacto-centric cosmology and later to cosmologies with the centre in our cluster of galaxies, *etc*., did not lead to any crises. Indeed this transition was predicted by the Newtonian theory.

An important result of the new approach to cosmology was the abolishment of the old division of physics into 'mundane physics' and 'celestial physics'. According to Aristotle, all phenomena '*sub luna*' (below the Moon) were ruled by the former, whereas the latter ruled events at or above the lunar orbit. The only one who earlier had questioned this was Giordano Bruno, but it was proved to be true by Galileo and Newton. It was the falling apple in Newton's garden which smashed the sphere which separated the two disciplines of physics.

Let us now return to the difference between myth and science. This is the difference between divine inspiration or 'unaided reason' (as Bertrand Russell put it) on the one hand and theories in intimate contact with observation on the other. The enormous inflow of observational material caused by the introduction of the telescope could not be accommodated within the crystal spheres. They were blown up by the injection of so many new observational facts. It is fair to say that *the 'Copernican revolution' was caused more by Galileo's introduction of the telescope than by the Copernican theory*. In fact, Aristarchos had proposed the same theory 2000 years earlier, but because there were no telescopes it could not be proved.

2.2 *Limitations of Newtonian Theory*

At the beginning of this century the Newtonian formalism was challenged in four different respects:

1. It was obvious that it was not applicable to atoms, where it had to be replaced by quantum mechanics.
2. Motions with velocities which were not negligible in comparison to the velocity of light must be treated by the special theory of relativity.
3. The general theory of relativity required that the three-dimensional Euclidean space of Newton be replaced by a four-dimensional curved space.
4. It became obvious that electromagnetic phenomena were of decisive importance for the motion of ionized diffuse media. It was necessary to introduce magneto-hydrodynamics and plasma physics into cosmic physics.

While the consequences of (1) and (2) are non-controversial, we shall discuss (3) and (4) later in Sections 3 and 4.

2.3 *Science and Old Myths*

How did the scientific breakthrough affect the old myths? To several of the pioneers it seems not to have been a real conflict. Tycho Brahe and Kepler, for example, were not

only prominent scientists but also prominent astrologers. (Even in present times, when a day has been extremely unlucky, Scandinavians often exclaim: "Today is a real Tycho-Brahe day!" The reason for this is that he published a calender of those days when the constellations were very unfavourable.) With regard to some of the pioneers, this lack of conflict may have been because it was dangerous to oppose the existing beliefs, or because some earned their living as royal astrologers. But it seems obvious that such an explanation does not suffice. In his letter to Bishop Bently, Newton himself wrote that his celestial mechanics proved the existence of God, and he spent his old age in calculating how many angels there were according to the Apocalypse.

2.4 *Science and New Myths*

The victory of science over myth in the field of celestial mechanics spread slowly to other fields. It took more than two centuries before it seriously invaded biology. In our century the scientific approach has embraced other areas which earlier were alien to it, such as the origin of life and the functioning of the human brain.

However, this does not mean a complete and definite victory of common sense and science over myth. In reality we witness today an antiscientific attitude and a revival of myth. This tendency has at least two causes. The popular creationism in the South in the United States derives from religious fanaticism. But in a way, the most interesting and also most dangerous threat comes from science itself. In a true dialectic sense it is the triumph of science which has released the forces which now once again seem to make myths more powerful than science and causes a 'scientific creationism' inside academia itself.

2.5 *Special Relativity*

One of the most beautiful results of science was the *special* theory of relativity. It was essentially based on the Michelson-Morley experiment and on Maxwell's theory of electromagnetism, which in an elegant way described all the results of the study of electric, magnetic, and optical phenomena. Already when expressed in an ordinary three-dimensional Cartesian coordinate system, the special theory of relativity is a beautiful theory, but its mathematical beauty is definitely increased somewhat if it is expressed in four-dimensional space.

This fact was given an enormous importance. It was claimed that "Einstein has discovered that space is four-dimensional", a statement which is incorrect. In fact, H. G. Wells (1894) has based his ingenious novel, *The Time Machine,* on the 'generally accepted idea' that space was four-dimensional, with time as the fourth coordinate. This novel was published when Einstein was fifteen years old.

However, the fourth coordinate which Einstein introduced was not time, but time multiplied by $\sqrt{-1}$. From a *mathematical* point of view this is elegant, because it meant that the Lorentz transformation can be depicted as a turning of a coordinate system in four-dimensional space. However, from a *physical* point of view it does not give any new information.

Many people probably felt relieved by being told that the true nature of the physical world could not be understood except by Einstein and a few other geniuses who were able to think in four dimensions. They had tried hard to understand science, but now it

was evident that science was something to *believe* in, not something which should be understood. Soon the bestsellers among the popular science books became those that presented scientific results as insults to common sense. The more abstruse the better! The readers liked to be shocked, and science writers had no difficulty in presenting science in a mystical and incomprehensible way. Contrary to Bertrand Russell, science became increasingly presented as the *negation of common sense.* One of the consequences was that the limit between science and pseudoscience tended to be erased. To most people it was increasingly difficult to find any difference between science and science fiction, except that science fiction was more fun.

But let us return to the theory of relativity and its direct impact on scientists. The four-dimensional presentation of the *special* theory of relativity was rather innocent. This theory is used every day in laboratories for calculating the behaviour of high-energy particles, *etc.* As experimental physicists have a strong feeling that their laboratories are three-dimensional, firmly located in a three-dimensional world, the four-dimensional formulation is taken for what it is: a nice little decoration comparable to a cartoon or a calendar pinup on the wall.

3. General relativity and the universe

3.1 *Revival of Pythagorean Philosophy*

On the other hand, in the *general* theory of relativity the four-dimensional formulation is more important. The theory is also more dangerous,because it came into the hands of mathematicians and cosmologists, who had very little contact with empirical reality. Furthermore, they applied it to regions which are very distant, and counting dimensions far away is not very easy. Many of these scientists had never visited a laboratory or looked through a telescope, and even if they had, it was below their dignity to get their hands dirty. They accepted Plato's advice to "concentrate on the theoretical side of their subject and not spend endless trouble over physical measurements". They looked down on observers and experimental physicists whose only job was to confirm their highbrow conclusions. Those who were not able to confirm them were thought to be incompetent. Observing astronomers came under heavy pressure from prestigious theoreticians.

The general theory of relativity opened an extremely fascinating possibility. Similar to the Earth's surface, which is without borders but is still finite, one can in a four-dimensional space have a hypersphere without any limits and still with a finite volume. This idea was certainly worthwhile investigating.

General relativity paved the way for a revival of Pythagorean thinking. Once again it was believed possible to explore the universe by pure mathematics. All the arguments against this, which had caused the downfall of Ptolemaean cosmology, were wiped away. The sign at the entrance to Plato's Academy, "Let none who has not learnt (Euclidean) geometry enter here", was modernized to "Let none who has not learnt Minskowskian geometry enter here". The cosmological discussion became monopolized by Big-Bang believers who had studied general relativity for years. No one else is allowed to have any views about cosmology. Textbooks on 'modern cosmology' start with general relativity and often, do not even mention the existence of heretical views.

Still more serious is the fact that only those observations which by any stretch of imagination could be interpreted as supporting the Big Bang are mentioned. The increasing number of observations which prove the Big-Bang hypothesis to be wrong are swept under the rug.

Also, the Pythagorean idea of correspondence between microcosmos and macrocosmos attracted new interest.

3.2 *Eddington's Cosmology*

One of the most interesting attempts to apply general relativity to cosmology was due to Eddington. With general relativity as background, he derived mathematical relations between the fine structure constant, the ratio between the gravitational and the electrical attraction, the age of the universe expressed in atomic time units, and the number of particles in the universe. The latter was found to be $2.36216 \ldots \times 10^{79}$. It was not really necessary to take the trouble of going out to count them all. He knew that at his writing desk he had counted every single one! Indeed, he followed Plato's advice to "concentrate on the theoretical side of the subject" and did "not spend endless trouble over physical measurements to the neglect of theoretical problems".

Eddington's cosmology was no doubt an intellectual masterpiece of the scientist whom Chandrasekhar calls " the most distinguished astronomer of his time". In a way it is a pity that it did not survive confrontation with fact. Eddington had good reason to say—like King Alphonse—"had I been present at the creation, I would have rendered profound advice".

3.3 *Big-Bang Hypothesis*

But the main stem of general relativity carries several other branches. If Eddington's cosmology is the most ingenious one, the most popular one is the Big-Bang cosmology. It is based on Friedman's solution of Einstein's equations. This solution has a *singular point.* To a mathematician a singular point is nothing very remarkable, but to a physicist it had *earlier* meant that something had gone wrong, a warning that the theory could not be applied to a real problem. However, without any serious discussion, this old tradition in physics was suddenly neglected. Instead, it was generally accepted that the singular point represented reality, and meant that at a certain time the whole universe consisted of one single point only. From this singular point the universe began to expand, so that all parts of it rush away from each other with velocities which are proportional to the distance between them.

These types of mathematical solutions seemed to be applicable to the 'expanding universe' which Hubble's famous empirical law describes. The way was now open for a grand new cosmology.

One of the originators of this was Abbé Lemaître, who called the universe when it was at the singular point '1' Atome Primitive'. Its great propagandist was Gamow. Neither Lemaitre nor Gamow went to the extreme in postulating that the whole universe ever was a mathematical *point.* The 'initial state' was supposed to be a concentration of 'all mass in the universe' in a very small sphere. This mass is heated to a temperature of several billion degrees. When this 'atomic bomb explodes', its parts are

thrown out with relative velocities which are sometimes close to the velocity of light. (As there is no pressure gradient, the analogy with an exploding bomb is misleading.)

This model, which at least from certain points of view was fascinating, was believed to explain the main evolution and the present structure of the universe. A number of consequences were claimed to derive from it: in less than half an hour after the explosion the elements we find now were formed by nuclear reactions in the very hot and very dense matter. At an early time a heat radiation was produced which, on further expansion, cooled down and should be now observed as a blackbody radiation with a temperature of 50 K. At a later stage the expanding matter condensed to form the galaxies we observe today. The average density in the universe must be at least 10^{-29} g cm $^{-3}$ in order to close it.

3.4 *Big Bang and Observations*

There is not a single one of these early agreements with observations which have not proved to be wrong. In fact, the Big Bang believers of today claim only two observational supports of their hypothesis.

One is the '3 K blackbody radiation' which obviously has a very high isotropy. Compared to the early prediction of a 50 K isotropic radiation, this represents a discrepancy of 10^4 in energy (because the energy is proportional to T^4), but with 'generally accepted' modifications of the scenario the claim that it supports the hypothesis must be taken seriously.

The other support is that the observed abundance of some light elements is too large to be explained by the nucleosynthesis in stars, which is accepted to explain the abundance of the other ~ 90 elements, (the Big-Bang believers claimed initially that they could account for the production of all elements, but now they admit that this is untenable). Because both the observational values of the cosmical abundances and the theory of nucleosynthesis in stars may very well be uncertain by a considerably larger factor, this is not a very strong support.

On the other hand, there are an increasing number of observational facts which are difficult to reconcile in the Big-Bang hypothesis. The Big-Bang establishment very seldom mentions these, and when non-believers try to draw attention to them, the powerful establishment refuses to discuss them in a fair way. A collection of objections has recently been published by Oldershaw (1983). Other critical arguments are summarized by Alfvén (1981).

The present situation is characterized by rather desperate attempts to reconcile observations with the hypothesis to 'save the phenomena'. One cannot avoid thinking of the state under the Ptolemaean epoch. An increasing number of *ad hoc* assumptions are made, which in a way correspond to the Ptolemaean introduction of more and more epicycles and eccentrics. Without caring very much for logical stringency, the agreement between these *ad hoc* assumptions with the Big-Bang hypothesis is often claimed to support the theory.

In reality, with the possible exception of the microwave background condition, there is not a single prediction which has been confirmed. The Big-Bang era has seen the discovery of quasars which have a fantastic release of energy. Unpredicted and explainable only by a precarious mechanism. X-ray astronomy and gamma-ray astronomy have introduced a new era with discoveries of incredibly rapid enormous

energy explosions (time constant of a fraction of a second!). Unpredicted again and even *post facto* difficult to reconcile in the Big-Bang cosmology.

The Big Bang is indeed a cosmology of the same character as the Ptolemaean: absolutely sterile. Will it have the same life-expectancy?

3.5 *Cosmic Black-Body Radiation*

It is increasingly evident that there is only *one* phenomenon which the Big-Bang believers seriously claim to prove their cosmology; the 3 K radiation. Schramm & Wagoner (1977) write "the primary reason for believing that our universe did emerge from a Big Bang remains the 3 K background radiation" and Weiss exclaims enthusiastically that the background radiation "satisfies almost beyond expectations the simple hypothesis that it is a remnant of a primeval explosion".

However, if we look at the background radiation without any preconceived ideas, how convincing is it? We measure an extremely cold radiation in a 'universe' which is 10^{10} light years or 10^{26} m, and conclude that this must derive from a state which was billions of degrees hot. Indeed, the expansion from, say, a millimeter-sized universe to the present 10^{10} light year size is by a factor 10^{29}. Is there any other field of science where such an extrapolation in one jump is accepted without very strong proof ? One seems never to have asked seriously whether at intermediate states there could not have been other mechanisms for isotropisation of the background radiation. As we have seen above, the Big-Bang universe contains so many phenomena which this cosmology cannot explain, so it would not be surprising if we discovered such a mechanism.

Indeed one such mechanism may already have been discovered. According to Wright (1981), it is quite reasonable that "needle-shaped conducting grains can provide sufficient capacity to produce the observed spectrum".

It will probably be objected that no one has observed the existence of such grains. However, long ago Spitzer had already shown that the existence of such grains were required in order to account for the interstellar polarization of light. If the choice is either to postulate the existence of such grains or accept the Big-Bang cosmology—which according to its believers has no other certain support—the needles may be preferred by all who are not fanatical believers.

3.6 *Creation ex Nihilo*

A very important conclusion from the Big-Bang cosmology, which is seldom drawn explicitly, is that the state at the singular point necessarily presupposes a divine creation.

To Abbé Lemaître this was very attractive, because it gave a justification to the creation *ex nihilo*, which Saint Thomas had helped establish as a credo. To many other scientists it was more of an embarrassment because God is very seldom mentioned in ordinary scientific literature. There seem to be rather few scientists (but among them Whittaker and Milne) who, like Jastrow (1978) in his book *God and the Astronomers*, explicitly draw what seems to be the logical conclusion of the Big-Bang cosmology, *viz*., that the universe was created *ex nihilo* by God. "When the scientist has scaled the mountains of ignorance, he is about to conquer the highest peak; as he pulls himself over the final rock, he is greeted by a band of theologians who have been sitting there for

centuries." However, most of the Big-Bang believers prefer to sweep creation under the rug. In fact, they fight *against* popular creationism, but at the same time they fight fanatically *for* their own creationism. Peratt (1983) suggests that the creationism *extra muros* is inspired by the Big-Bang creationism *intra muros.*

3.7 *Hierarchical Cosmology*

Are there any alternatives to the Big Bang? Indeed there are, although the Big-Bang believers very seldom mention this. Like in the good old days, even mentioning of the existence of a heresy is a crime. One of the mostinteresting alternatives is the hierachical cosmology,which envisages an *infinite universe with a hierarchical organization.* It is based on an approach which attracted considerable interest in the beginning of this century, long before the Big Bang, indeed even before the general theory of relativity.

Inspired by Fournier-d'Albe, Charlier demonstrated that in order to avoid the Olbers and Seeliger objections to a Euclidean infinite universe, it is necessary that the universe is 'clumpy', with a hierarchical matter distribution. This means that stars should be organized in galaxies G_1, a large number of these galaxies form a larger 'galaxy of type G_2'—we would today prefer to speak of a 'cluster'—, a large number of these a still larger structure G_3, and so on into infinity. Charlier showed that the mean density of a structure of size R must obey

$$\sigma \sim R^{-a} \tag{1}$$

with $a > 2$. This leads to an *infinite universe with infinite mass but with average density zero.* It satisfies both the Olbers and the Seeliger objection.

The Charlier school speculated whether our *metagalaxy* (a synonym for what is the Big Bang formalism is considered as the whole 'universe') may have sisters which together form a still larger structure (a 'teragalaxy'), thus continuing one step further in the hierarchy. (This is, of course, against the Big-Bang view).

With the arrival of the Big-Bang cosmology,the Charlier model was considered to be of historical interest only. However, in a classical paper, de Vaucouleurs (1970) revived that model by demonstrating that within wide limits, the maximum *observed density distribution* satisfies Equation (1), but with $a = 1.7$.

In his theoretical interpretation of the observations deVaucouleurs (1970) must take into account the Hubble expansion, which means that his hierarchical model is not identical with Charlier's.

Peebles and collaborators (*cf.* Peebles 1980) have treated the observational data with advanced statistical methods, and have essentially confirmed the de Vaucouleurs hierarchical model. (See survey article by Groth *et al.* 1977. However, they find a value of a which is somewhat higher: $a = 1.77$.)

3.8 *A Tycho-Brahe-Type Compromise*

The hierarchical structure does not necessarily come into conflict with the Big Bang. A number of scientists (including even de Vaucouleurs and Peebles) prefer a Tycho-Brahe compromise: Certainly observations demonstrate that the universe has not at all the homogeneity which it should have according to the Big Bang, but the inhomogeneities may be explained by secondary effects, *e.g.*, instabilities. In this way, an open conflict

with the sacrosanct Big Bang is avoided. However,even if a hierarchical structure could be derived from the Big Bang, *this does not prove that the Big Bang can be derived from the observed hierarchical structure*! Moreover, the real advantage with the hierarchical structure is that it saves us from the *singular point and creationism.* So this compromise seems to be just as superficial as Tycho Brahe's.

Neither Charlier nor anyone else has given any reason *why* matter has this structure and is distributed in this way. Only by implication do they claim that there must be *some law of physics* which produces a hierarchical structure.

In any case, it seems legitimate to look for alternatives to the Big Bang. However, it is beyond the scope of this paper to discuss these (see Alfvén 1981, Chapter VI; Alfvén 1982a).

4. Introduction of spacecraft

4.1 *Importance of Electromagnetic Forces*

Independent of the introduction of the General Theory of Relativity into the cosmological discussion, there was another drastic change in our approach to cosmical physics, namely, the realization of the importance of *electrodynamic effects* to the motion of dispersed media. Because the ratio of Coulomb attraction to Newtonian attraction between elementary particles is 10^{39}, electromagnetic effects are decisive to the dynamics in all cases when the number of positive charges are not almost exactly compensated by the same number of negative charges. This is the case for all massive celestial bodies down to grains of the size of the order of microns, but very seldom for the diffuse media in interplanetary, interstellar and intergalactic space. In fact, hydromagnetic and plasma phenomena dominate most of those regions which (by volume) constitute more than 99.999 . . . per cent of the universe.

In the following, we shall see that it is not only Newton and Einstein, but also Maxwell who are important to cosmology.

4.2 *Space Research and the Paradigm Transition in Cosmic Physics*

Scientific progress depends on the development of new instruments. The change from Ptolemaic to Copernican cosmology was to a large extent caused by the introduction of telescopes. Similarly, space research has changed our possibilities to explore our large-scale environment so drastically that a thorough revision of cosmic physics is now taking place.

First of all, *space observations* have made almost the whole electromagnetic spectrum available to observation. Earlier, less than one-third of the octaves (visual and a region of the radio frequencies) supplied us with information.

The new regions include X-ray and gamma-ray astronomy, and most of the new phenomena discovered in these regions are obviously due to plasma effects. This means that the decisive importance of hydromagnetics and plasma physics has now become increasingly obvious.

Moreove

the solar wind region ('solar magnetosphere') drastically changed our understanding of the properties of cosmic media. Further, we have learned how to generalize results from plasma investigations in one region to other regions. This means that laboratory investigations of plasmas of the size of, say, 10 cm can be used to achieve better understanding of cosmic plasmas of magnetospheric dimensions; say, 10^{10} cm. By another step of 10^9 we can transfer laboratory and magnetospheric results to galactic plasmas of, say, 10^{19}cm. A third jump of 10^9 brings us up to the Hubble distance 10^{28} cm and hence to cosmological problems.

All this has led or is leading to a revision of our concept of cosmic plasma, which in many respects is so drastic that it is appropriate to speak of a *change in paradigm.*

As our cosmic environment consists of plasma to more than 99.999 . . . per cent (by volume), this means a revision of a large part of cosmic physics.

A list of fourteen fields of astrophysics which must be revised has been given (Alfvén 1981, 1982b, 1983). Those of most interest in this connection are:

(a) Electric double layers, which did not attract very much interest until five or ten years ago. They are now known to accelerate charged particles to kilovolt energies in the terrestrial magnetosphere. Double layers may also exist elsewhere and accelerate particles to even higher energies. Carlqvist (1982) has treated relativistic double layers which may accelerate particles to cosmic ray energies. The breakthrough in the acceptance of electric double layers came with the Risø Symposium (Michelsen & Rasmussen 1982). In a cosmological connection, the rapid release of magnetically stored energy in exploding double layers is of considerable interest.

(b) Cosmic plasmas are often not homogeneous, but exhibit *filamentary structures*, which in accessible regions are known to be associated with currents parallel to the magnetic field. It is likely that filamentary structures in interstellar clouds as well as further out are also produced by filamentary currents.

(c) In the magnetospheres there are thin, rather stable *current layers* which separate regions of different magnetization, density, temperature, *etc*

(d) It is difficult to avoid the conclusion that similar phenomena exist also in more distant regions. This is bound to give space a general *cellular structure* (or more correctly, a cell-wall structure).

(e) The arguments for the non-existence of antimatter in the cosmos are not valid (Rogers & Thompson 1980). There are sound arguments for the existence of antimatter, which means that *annihilation* should be considered an important source of energy. In fact, annihilation seems to be the only reasonable energy source for those celestial objects which emit very large amounts of energy (*e.g*., quasars).

(f) Radio, X-ray, and gamma-ray emissions and cosmic-ray acceleration are largely due to plasma processes. Theories of, for example, double radio sources, the formation of stars and planetary systems from interstellar clouds, energy release in quasars and acceleration of cosmic radiation upto 10^{19} must be based on plasma physics.Hence the paradigm transition implies a revision of considerable parts of radio, X-ray and gamma-ray astronomy, the theory of cosmic rays, and also of cosmology. These sciences must ultimately be based on the observed properties of laboratory and magnetosphere plasmas.

Hence, in conclusion, there is not very much left of the observational support for Big Bang. Indeed, *the space age gives a picture of space as essentially three-dimensional and highly inhomogeneous, because of the dominance of hydrodynamics and plasma physics.* In contrast, *the Big-Bang scenario is a four-dimensional and basically homogeneous space.*

4.3 *Mundane and Celestial Mechanics*

At present, it is agreed at least in principle that up to a distance of some per cent of the Hubble distance, Newtonian mechanics and a Euclidean space should normally be used. Of course, special relativity must be applied for all motions and velocities comparable with the velocity of light. Further, hydromagnetics and plasma physics are often of decisive importance. Even if it is admitted that *in principle* general relativity is valid, the difference between this and Newtonian mechanics is negligible except in a few special cases. In fact, for calculations of planetary orbits, and also the orbits of spacecraft, Newtonian mechanics is used, because the relativity correction is only of the order of 10^{-8}. Similarly, for the large-scale dynamics, we expect the relativity correction for galaxies to be, say, 10^{-6} and for galactic clusters and superclusters, perhaps 10^{-5}. Exceptions are special cases like neutron stars and black holes (if there are any!). If, using the empirical formula of de Vaucouleurs, we extrapolate to a hierarchical order of the whole metagalaxy (to the Hubble distance), we get a relativity correction of about 10^{-4}. This is also negligible considering the accuracy which at present is useful in treating large-scale phenomena in regions so distant.

It goes without saying that in all these regions, hydromagnetics and plasma effects are, in general, much more important.

The exploration of increasingly distant regions have demonstrated that the strong inhomogeneity characteristic of a hierarchical structure is valid out to at least some per cent of the Hubble distance. Indeed, very large void regions of dimensions 10^{20}–10^{25} m have been discovered and also massive regions of different structures out to similar distances. If we use the Charlier-de Vaucouleurs relation between average density and size of the hierarchical structures we obtain an average density which is four orders of magnitude less than what is required for closure. Hence, those who want to close the 'universe' at the Hubble distance have to assume a drastic change in average density to take place within the last order of magnitude out to this limit.*

Hence, there is a reasonably well-defined limit between strongly inhomogeneous and essentially Euclidean space which is extrapolated from space research results and the homogeneous four-dimensional space which is postulated by Big-Bang believers. This limit is given by the present reach of reliable observations. The limit has been retreating with the advances in observational technique. But, of course, we cannot be absolutely sure that it will retreat still further.

This limit may be compared with the limit in the Aristotelian cosmology between mundane laws, valid below the lunar orbit, and celestial laws, valid above. For example, according to mundane mechanics heavy bodies fall down, but the Moon, the planets, the Sun and the stars do not fall down because they obey celestial laws. Similarly, out to several per cent of the Hubble distance, we are confident that the 'mundane' laws of laboratory and near space hold. But the Big-Bang believers claim that their 'celestial' laws hold outside the limit.

Allowing for the uncertainty which is inherent in all cosmologies, it seems that the present cosmological situation is similar to what it was at the time of Saint Thomas: "Reason can only be satisfied with the assumption that the world has no beginning. The

* What is said should not be interpreted as a questioning of the general theory of relativity. It is only an attempt to clarify to what extent it is applicable to cosmology. Einstein expressed himself in a much more careful way than many of his epigones.

doctrine of a beginning or the non-eternity of the world is to be received *sola fide* as an act of pure faith in deference to authority".

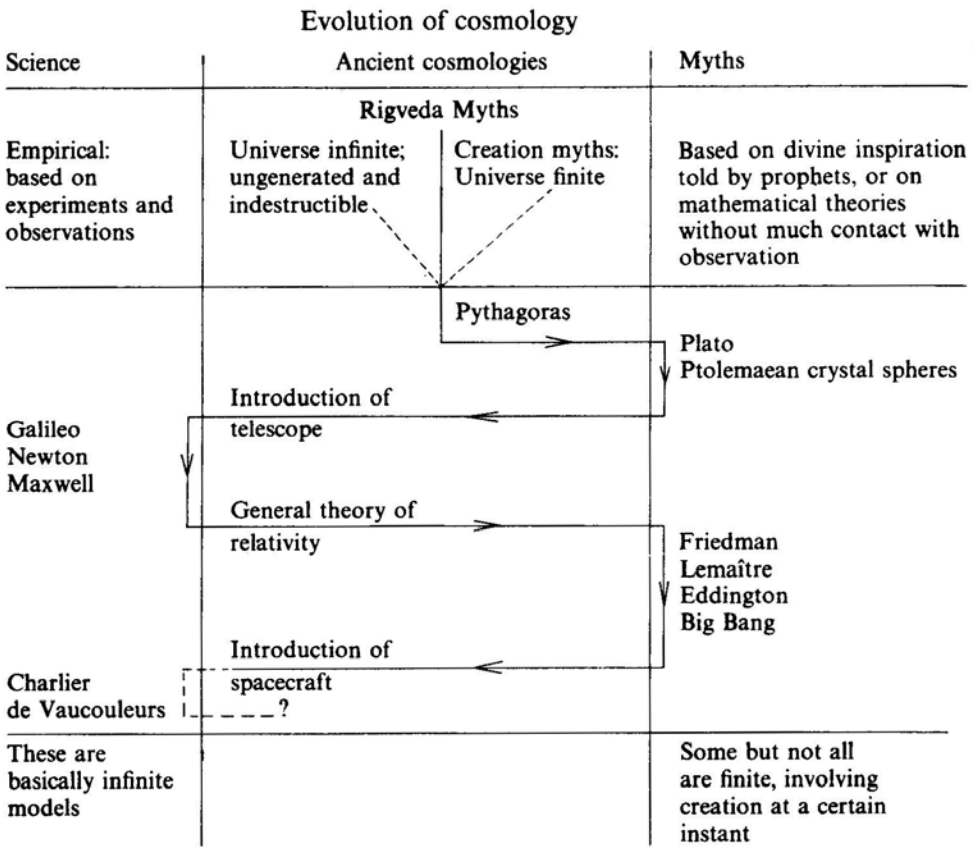
5. The cosmological pendulum

Three or four millenia of cosmological speculation has resulted in essentially three different types of approaches to cosmology:

1. The scientific approach. As science is basically *empirical*, this means that cosmology should be based on observations with experimental results (from laboratory or nowadays also space experiments) as a background.

The *Newtonian* theory was largely based on accurate observation of planetary motion. It turned out to be applicable—at least to a good first approximation—to motions of galaxies,'and clusters of galaxies.

Today, especially after *in situ* magnetospheric measurements and the birth of X-ray and gamma-ray astronomy, it must be fused with *Maxwellian* theory, which leads to hydro-magnetics and plasma physics as basic to the study of our cosmic environment. Maxwell's theory is a summary of the results of electromagnetic investigations, and—like Newton's theory—it turned out to be applicable to a number of problems in other fields.



2. *The agnostic attitude.* This is the Rigvedic and Buddhist approach: How can we know about or why should we care about problems so distant?
3. *The mythological approach.* If venerable prophets have told us that by divine inspiration they know that the universe was created, and how it was created, how can we doubt what they tell us? This approach is closely related to the *mathematical myths*: It is possible to explore the structure and evolutionary history of the universe by pure theoretical thinking without very much contact with observations. Typical examples are the Pythagoras-Plato-Ptolemaean cosmology, or in our day, the Eddington cosmology, but also the Big Bang.

There has been—and will perhaps always be—an oscillation between mythological and scientific approaches. This is summarized here in the diagram called the *Cosmological Pendulum* which is a summary of what has been said in this paper.

It is interesting to ask whether the pendulum could come to rest in an intermediate position. Eddington himself has given the answer: "In one sense, deductive theory is the enemy of experimental physics." Since the birth of science, there has never been a time when there could be a compromise between myth and empirical science.

Will there—in a still more distant future—again be a swing back to created cosmologies? Perhaps. However, we cannot expect such a future model to resemble the Big Bang any more than Big Bang does the crystal spheres. Its size and timescale will be much larger. In fact, the age of the Big Bang is just a few *Kalpas*, and the life expectancy of Brahmais ten thousand times this. Perhaps this future model will be of a scale which only the Vedic cosmologists dared to imagine.

Philosophizing over the swings of the cosmological pendulum we may remember the words of the Buddha:

> "It is wrong to say that the world is infinite and eternal."
> Yes, at least during some periods it has been.
> "It is wrong to say that the world is finite and created."
> Yes, at least during some periods it has been.

In this sense the Buddha was correct. But of course what he meant was something else, much deeper and more sophisticated.

Acknowledgements

I wish to thank Dr. W. B. Thompson for useful discussions, and Jane Mead Chamberlin for editing the manuscript. The first paragraphs of this paper are essentially based on Hastings (1961a,b), Basham (1959) and Dicks (1970).

Bibliography

Alfvén, H.1977, in *Cosmology*, *History and Theology*, Eds W. Yourgrau and A.D. Beck, Plenum Press, NewYork.
Alfvén, H. 1981, *CosmicPlasma*, D. Reidel, Dordrecht, Chapter VI.
Alfvén, H. 1982a, *Astrophys. Space Sci*., **89,** 313.
Alfvén, H. 1982b, *Phys. Scripta*, **2,** 10.
Alfvén, H. 1983, *Geophys.Res.* Lett., **10,** 487.
Basham, A . L . 1959, *The Wonder that Was India.* Evergreen Encyclopedia, Vol. I, New York.

Carlqvist, P. 1982, in *Symposium on Plasma Double Layers*, Eds P. Michelsen and J. J. Rasmussen, Rise National Laboratory, Roskilde.
de Vaucouleurs, G. 1970, *Science* **167,** 1203.
Dicks, D. R. 1970 *Early Greek Astronomy to Aristotle*, Cornell Univ. Press.
Eddington, A. 1924, *The Mathematical Theory of Relativity*, Cambridge Univ. Press.
Eddington, A. 1952, *The Expanding Universe,* Cambridge Univ. Press.
Groth, E. J., Peebles, P. J. E., Seldner, M., Soneira, R. M. 1977, *Scientific American*, **237,** Nov. 1977, p. 76.
Hastings, J. (Ed.), 1961a, *Encyclopedia of Religion and Ethics*, *Vol. I. Agnosticism* (*Buddhist*), New York.
Hastings, J. (Ed.), 1961b, *Encyclopedia of Religion and Ethics, Vol. IV. Cosmogony and Cosmology*, NewYork.
Ibn Khaldun 1379, Muqaddimah (English translation 1958 by Franz Rosenthal. See also Muhsin Mahdi: *Ibn Khaldun's Philosophy of History*, 1975, and Charles Issawi: *An Arab Philosophy of History*, 1950).
Jastrow, R. 1978, *God and the Astronomers* Readers Library.
Michelsen, P., Rasmussen, J. J. (Eds), 1982, *Symposium on Plasma Double Layers,* Risø National Laboratory, Roskilde.
Munitz, M. K. 1957, *Space*, *Time and Creation.*
Oldershaw, R. L., 1983, *Astrophys. Space Sci.* **92,** 347.
Peebles, P. J. E. 1980, *The Large-Scale Structure of the Universe*, Princeton Univ. Press.
Peratt, A. L. 1983, *Physics Today*, 36, 15.
Rogers, S., Thompson, W. B. 1980, *Astrophys. Space Sci*., **71,** 257.
Schramm, D. N., Wagoner, R. V. 1977, *Ann. Rev. Nucl. Sci*., **20,** 41.
Singer, C. 1959, *A Short History of Scientific Ideas,* Oxford Univ. Press.
Trattner, E. R. 1940, *The Great Theories of Mankind*. New York.
Walsh, L. G. M. 1910, *The Doctrine of Creation.*
Weiss, R. 1980, *A. Rev. Astr. Astrophys*., **20,** 489.
Wells, H. G. 1894, *The Time Machine.*
Wright E. I. 1981, *Astrophys. J*., **255,** 401.

J. Astrophys. Astr. (1984) **5,** 99–110

On Gamma Radiation from the Magellanic Clouds and Galactic Supernova Remnants —Possible Antiproton Sources in the Galaxy

V. L. Ginzburg *P. N. Lebedev Physical Institute of the USSR Academy of Sciences, Leninsky Prospect 53,117924 GSP, Moscow, B–333, USSR*

V. S. Ptuskin *Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation, USSR Academy of Sciences, Moscow Region 142092, USSR*

(*Invited article*)

Abstract. Radiation from the Magellanic Clouds is discussed from the point of view of near future possibilities in observational γ-ray astronomy. The γ-ray fluxes expected according to the metagalactic and galactic theories of the origin of cosmic rays are compared. It is shown that the strongest test of the metagalactic hypothesis will be provided by a determination of the ratio of γ-ray fluxes from SMC and LMC. The γ-ray luminosity of a typical young supernova remnant that can generate sufficient antiprotons is estimated. It is shown that such remnants must have a short phase during which they are very powerful γ-ray emitters.

Key words: cosmic rays—γ-ray astronomy—Magnellanic Clouds—anti-protons—supernova remnants

1. Introduction

The vast scope of γ-ray astronomy is being recognized only now, though its potential was already foreseen two decades ago from theoretical investigations, particularly of cosmic rays (*cf.* Ginzburg & Syrovatskii 1964: Sections 18 and 19). Unfortunately, many of the much needed observations require complex satellite-borne γ-ray telescopes such as are planned to be launched some time during the present decade. Hence it appears relevant to discuss here two problems of γ-ray astronomy-the γ-radiation from the Magellanic Clouds and from galactic supernova remnants. Our scientific interests lead us naturally to explore the close connection between γ-ray astronomy and the problem of the origin of cosmic rays.

2. Gamma-radiation from the Magellanic Clouds

One of the ‘eternal’ questions concerning the cosmic rays observed near the Earth is the choice between a galactic or metagalactic origin for them. The present authors have always advocated models with a galactic origin for the cosmic rays (Ginzburg & Syrovatskii 1964; Ginzburg & Ptuskin 1976), but they agree that a final resolution requires a more rigorous proof than just indirect arguments based mostly on energetics. As for the electron (and positron) component of cosmic rays in the Galaxy, such a proof

appeared after the discovery in 1965 of the relic thermal radio emission with a temperature $T \sim 3$ K. Due to inverse Compton scattering of relativistic electrons by the 3 K background photons, the electrons with an energy $E \gtrsim 10^{10}$ eV cannot reach the Earth even from the nearest radio galaxy—Centaurus A (*cf.* Ginzburg 1975). Thus a major part of the electron component of the cosmic rays in the Galaxy must have a galactic origin. Further, we have indirect evidence against a metagalactic origin for the proton-nuclear component of cosmic rays (Ginzburg 1975). However, for an unequivocal proof of the galactic origin of the main part of this component, we must show that outside the Galaxy (in the metagalactic space), the cosmic-ray energy density

$$w_{\mathrm{cr,mg}} \ll w_{\mathrm{cr,G}} \sim 10^{-12}\ \mathrm{erg\ cm^{-3}}, \tag{1}$$

where the subscripts mg and G refer to metagalactic space and the Galaxy, respectively. As proposed by Ginzburg (1972), in the context of the Magellanic Clouds, the energy density (or intensity) of the protonnuclear component far from the Earth can be measured by gray astronomical methods*. Specifically, we mean the detection of γ–photons from the decay of π° mesons produced by collisions of cosmic rays (relativistic protons and nuclei) with particles (say, in the interstellar gas).

The γ-ray flux (above an energy E_γ) emitted by a discrete source is given by

$$F_\gamma(>E_\gamma) = \int_\Omega I_\gamma(>E_\gamma)\,\mathrm{d}\Omega = \frac{\bar{q}_\gamma N(V)}{R^2} = \frac{5\times 10^{23}\bar{q}_{\gamma,0}M}{R^2}\ \mathrm{photons\ cm^{-2}\ s^{-1}} \tag{2}$$

where $I_\gamma(>E_\gamma)$ is the source intensity of γ-rays above an energy E_γ, Ω is the solid angle subtended by the source, R is the distance to it (in cm), $N(V) = nV$ is the number of particles in the gas inside the source (V is the source volume and η is the mean concentration of nuclei in it), and $M = 2 \times 10^{-24} N(V)$ is the gas mass (in g) in the source. Here we have assumed that the chemical composition of the gas corresponds to the mean elemental abundance, and accordingly the mass of a mean nucleus is taken equal to 2×10^{-24} g. Further, the average emissivity, *i.e.* the number of photons in unit solid angle is

$$\bar{q}_\gamma = \overline{\sigma I(>E_\gamma)} = \int_{E_\gamma}^{\infty}\int_{E=E_\gamma}^{\infty} \sigma(E_\gamma, E) I_{\mathrm{cr}}(E)\,\mathrm{d}E\,\mathrm{d}E_\gamma, \tag{3}$$

where $I_{\mathrm{cr}}(E)$ is the intensity of cosmic rays with energy E and σ is the corresponding effective cross-section. Using their known spectrum, we have for cosmic rays near the Earth

$$\bar{q}_{\gamma,0}(E_\gamma > 100\ \mathrm{MeV}) = 10^{-26}\ \mathrm{s^{-1}\ sr^{-1}} \tag{4}$$

(Ginzburg 1972). Hence

$$F_\gamma(E_\gamma > 100\ \mathrm{MeV}) = \frac{5\times 10^{-3} M\,(w_{\mathrm{cr,s}}/w_{\mathrm{cr,G}})}{R^2}\ \mathrm{photons\ cm^{-2}\ s^{-1}}, \tag{5}$$

where $w_{\mathrm{cr,s}}$ is the cosmic-ray energy density in the source. We have assumed here that the form of the cosmic-ray spectrum in the source as well as in the Galaxy is the same as observed near the Earth, so that we may write $w_{\mathrm{cr,s}}/w_{\mathrm{cr,G}} = I_{\mathrm{cr,s}}/I_{\mathrm{cr,G}}$. The following estimates (from Bok 1966) for the distance and H I content of the Magellanic Clouds

* We do not touch upon the region of very high energies $E \gtrsim 10^{12}$ eV, where one can, in principle, use neutrino detection techniques (see *e.g.,* Berezinsky & Ginzburg 1981).

were used in Ginzburg (1972), *i.e.*,

$$\begin{aligned} R(\mathrm{LMC}) &= 55\,\mathrm{kpc} \\ R(\mathrm{SMC}) &= 63\,\mathrm{kpc} \\ M_{\mathrm{HI}}(\mathrm{LMC}) &= 5.4\times 10^{8}\,M_{\odot} = 1.1\times 10^{42}\,\mathrm{g} \\ M_{\mathrm{HI}}(\mathrm{SMC}) &= 4\times 10^{8}\,M_{\odot} = 8\times 10^{41}\,\mathrm{g}. \end{aligned} \tag{6}$$

In the metagalactic models, cosmic rays in the Galaxy and in the Magellanic Clouds may be assumed to have the same intensity. Substituting accordingly in Equation (5) we get

$$w_{\mathrm{cr,s}} = w_{\mathrm{cr,mg}} = w_{\mathrm{cr,G.}} \tag{7}$$

This result is obtained under the assumption of isotropy and stationarity of cosmic rays. Then, using the values (6), we obtain

$$\begin{aligned} F_{\gamma,\,\mathrm{LMC}}(E_\gamma > 100\ \mathrm{MeV}) &\simeq 2\times 10^{-7}\ \mathrm{photons\ cm^{-2}\ s^{-1}}, \\ F_{\gamma,\,\mathrm{SMC}}(E_\gamma > 100\ \mathrm{MeV}) &\simeq 1\times 10^{-7}\ \mathrm{photons\ cm^{-2}\ s^{-1}}. \end{aligned} \tag{8}$$

Unfortunately, the fluxes (8) are not very accurate for the following reasons. The emissivity (4) takes into account only the contribution from the proton-nuclear component, but at $E_\gamma > 100$ MeV the bremsstrahlung radiation of electrons is also important. Including this effect for the cosmic rays in the Galaxy, Wolfendale (1981) derives a value of $\bar{q}_\gamma$, which is 2.2 times the one given by Equation (4). (Note that Wolfendale calculates the emissivity per H atom and uses the more common $q/4\pi$ for q.) He also shows that the values (8) should further be increased by a factor of two if the Magellanic Clouds contain the same amount of molecular hydrogen as the Galaxy. The uncertainty in determining the distance R, which is about 10 percent, is a further source of inaccuracy in the fluxes (8).

The situation is quite different, however, for the ratio Δ of fluxes from SMC and LMC. As the equality (7) or, more precisely, the equality $I_{\mathrm{cr,\,LMC}}(E) = I_{\mathrm{cr,\,SMC}}(E)$ holds in the metagalactic models, if the gas composition (including the fraction of H_2) is assumed similar for LMC and SMC, we get

$$\Delta = \frac{F_{\gamma,\,\mathrm{SMC}}(>E_\gamma)}{F_{\gamma,\,\mathrm{LMC}}(>E_\gamma)} = \frac{M_{\mathrm{HI}}(\mathrm{SMC})/R^2_{\mathrm{SMC}}}{M_{\mathrm{HI}}(\mathrm{LMC})/R^2_{\mathrm{LMC}}} = \frac{\phi_{21}(\mathrm{SMC})}{\phi_{21}(\mathrm{LMC})}. \tag{9}$$

It should be noted that the cross-section σ, the amount of molecular hydrogen, and even the distance R drop out since we use the ratio M_{HI}/R^2 which is proportional only to the frequency-integrated flux ϕ_{21} in the 21-cm hydrogen line.

As is well known, the role of bremsstrahlung radiation from the galactic electron component is still significant at $E_\gamma = 100$ MeV. (Furthermore, since the losses are small, electrons with an energy $E_\gamma \sim$ 100–300 MeV could well have originated in the metagalactic space confusing the issue even more.) So, in order to separate the contribution of the purely proton-nuclear component, one must consider only energies $E_\gamma >$ 100–300 MeV. Undoubtedly, the separation of the contribution of the electron component to the γ-ray emissivity is of great importance. We begin here with the suggestion that this can be done to obtain the contribution of the proton-nuclear component. Further, one should also take into account the contribution of discrete γ-ray sources. There is no reason to expect, nor any evidence to show, that discrete sources in the Magellanic Clouds play a more significant role than those in the Galaxy.

The contribution of discrete sources to the galactic γ-luminosity is about 20 per cent (Wolfendale 1981). If the same value is assumed for LMC and SMC, the estimate of fluxes from these Clouds changes little. In any case, the ratio (9) remains unchanged even accounting for the discrete sources if we assume that they contribute the same fraction to both the luminosities.

Thus, under the above conditions, the relation (9) must hold in purely metagalactic models for the origin of the proton-nuclear component. Using the estimates (6), we find that $\Delta = 0.56$ to a better degree of accuracy than that of Equation (8). Even so, the accuracy of this value is still dependent on that of the estimates of the hydrogen masses and distances of LMC and SMC. For example, replacing the estimate of M_{H1} (SMC) given in (6) by the improved value $4.8 \times 10^8\, M_\odot$ (Hindman 1967), but keeping the other numbers the same, changes the corresponding value of Δ to 0.68. It is thus clear that in order to achieve an accuracy higher than 10 per cent, one needs more precise data on the integrated fluxes ϕ_{21} (LMC) and ϕ_{21} (SMC) than we understand are presently available.

Since γ-ray fluxes from the Magellanic Clouds have not yet been measured, attention in recent years has naturally focused on a method somewhat different from, but related to, the one proposed above for testing the validity of galactic models of cosmic-ray origin. In these models, the intensity I_{cr} and the energy density w_{cr} of cosmic rays evidently decrease as one goes away from the galactic centre and especially as one goes over to the metagalactic region. This would affect the γ-emissivity in the galactic anticentre direction depending on the distance of the radiating region from the Sun (Dodds, Strong & Wolfendale 1975). This method, which is quite good in principle, seems to have yielded results in favour of galactic models (Wolfendale 1981; Dodds, Strong & Wolfendale 1975). Unfortunately, the available data are not yet sufficiently convincing. While Wolfendale has found evidence for a gradient in the cosmic-ray density away from the galactic centre, Bloemen *et al.* (1983) have not. We hope future measurements in this respect will make the picture clearer. However, a new generation of γ-ray telescopes which could measure γ-ray fluxes from the Magellanic Clouds are necessary for this purpose. The preceding discussion amply justifies the need for such measurements.

In this connection we repeat here the calculations of the γ-radiation from the Magellanic Clouds by Houston, Riley & Wolfendale (1983). The reasons are two-fold. Firstly, we disagree with their model and prefer a different distribution of the galactic cosmic rays; secondly, it will make our assumptions clearer in the process. We assume the following model of the galactic disk (Ginzburg & Ptuskin 1976): All the gas, with a total mass M_{g}, is concentrated in a disk of thickness $2h_{\mathrm{g}}$ and radius R_{g}, *i.e.* of volume $2\pi h_{\mathrm{g}} R_{\mathrm{g}}^2$. The cosmic rays fill a disk of the same radius, but with a different half-thickness h_{h} corresponding to the galactic halo. The total cosmic-ray energy in the Galaxy is $W_{\mathrm{cr}} = 2\, h_{\mathrm{h}}\, R_{\mathrm{g}}^2 w_{\mathrm{cr}}$. and the source power, $U_{\mathrm{cr}} = W_{\mathrm{cr}}/T_{\mathrm{cr}}$ where T_{cr} is the cosmic-ray lifetime in the system. In the diffusion model, one can put $T_{\mathrm{cr}} = h_{\mathrm{h}}^2/D$ where D is an effective diffusion coefficient (since we do not calculate D here, factors of the order of unity in the expression for T_{cr} are ignored). Thus,

$$w_{\mathrm{cr}} = \frac{U_{\mathrm{cr}} h_{\mathrm{h}}}{2\pi R_{\mathrm{g}}^2 D}\,. \tag{10}$$

Disregarding the discrete sources, we express the γ-luminosity of the Galaxy as

$$L_{\gamma,\,\mathrm{dif}} = A M_{\mathrm{g}} w_{\mathrm{cr}} = A\,\frac{U_{\mathrm{cr}} h_{\mathrm{h}} h_{\mathrm{g}} \rho_{\mathrm{g}}}{D}\,, \tag{11}$$

where A is some coefficient and ρ_g is the gas density in the disk ($M_g = 2\pi h_g R^2 \rho_g$). Like Houston, Riley & Wolfendale, we assume the γ-luminosity from discrete sources to be

$$L_{\gamma,s} = BU_{cr} \tag{12}$$

where B is an unknown coefficient. Then the total galactic γ-luminosity would be

$$L_{\gamma,dif} + L_{\gamma,s} = U_{cr}\left[A\frac{h_h h_g \rho_g}{D} + B\right]. \tag{13}$$

Expressing the ratio $L_\gamma / L_{\gamma\ dif}$ for our Galaxy in terms of $\alpha_G = B_G/[A_G(h_h h_g p_g/D)_G]$, we have for the ratio of the luminosities of LMC and the Galaxy

$$\lambda_\gamma \equiv \frac{L_{\gamma,LMC}}{L_{\gamma,G}} = r\,\frac{(h_h h_g \rho_g/D)_{LMC}(h_h h_g \rho_g/D)_G^{-1} + \alpha_G}{1+\alpha_G}, \tag{14}$$

where the coefficients A and B for LMC and for the Galaxy are assumed to be equal, and $r = U_{cr,\,LMC}/U_{cr,\,G}$.

The expression (14) differs from the one presented by Houston, Riley & Wolfendale only in the replacement of h_g by $h_h h_g$. This is no surprise as they assume the cosmic-ray outflow from the Galaxy to be determined by a gas disk of thickness h_g, and not h_h. In Equation (14) r, λ_γ and $\alpha \equiv \alpha_G$ are the same symbols as used by Houston, Riley & Wolfendale. Assuming as in their paper that

$$\begin{aligned} r &= 0.2, & D_{LMC} &= D_G, \\ \rho_{g,LMC} &\simeq \rho_{g,G}, & h_{g,LMC} &\simeq h_{g,G}, \end{aligned} \tag{15}$$

and setting $h_h = h_g$, we recover their results, *i.e.,*

$$\lambda_\gamma \simeq r = 0.2, \quad L\text{gLMC} = 0.2L\text{g,G} \simeq 3 \times 10^{41}\text{photons s}^{-1} \tag{16}$$

so long as $L_{\gamma,\,G}$ (E= 100 MeV) = 1.45×10^{42} photons^{-1} as in their paper. Using Equation (16), and R_{LMC} = 52 kpc,

$$F_{\gamma,LMC}\ (E_\gamma > 100\text{MeV}) \simeq 7 \times 10^{-7}\text{photons cm}^{-2}\ \text{s}^{-1} \tag{17}$$

As has been mentioned already, $\alpha \simeq 0.2$ for the Galaxy. If the ratio α is smaller for the LMC, the contribution from discrete sources will not change these estimates by more than a factor of 2. Therefore assuming for simplicity $\alpha = 0$ (*i.e.,* $L = L_{\gamma,\,dif}$), from Equations (11) and (16) we find that

$$\frac{w_{cr,LMC}}{w_{cr,G}} = \frac{(L_\gamma/M_g)_{LMC}}{(L_\gamma/M_g)_G} \simeq 2 \tag{18}$$

since $M_{g,\,LMC}/M_{g\,G} \simeq 0.1$.

We consider it more correct not to identify h_h with h_g, but to suppose, for example, that $h_h \sim R_g$ corresponding to a quasi-spherical halo. If we consider that according to Houston *et al.*, $R^2_{g,\,LMC}/R^2_{g,\,G} = M_{g,\,LMC}/M_{g,\,G}$, then $h_{h,\,LMC}/h_{h,\,G} \sim 1/3$. Consequently, we now obtain from Equations (14) and (15), the right-hand sides of equations (16)–(18) scaled down by a factor of three, (a still taken as zero), *i.e.,*

$$L_{\gamma,LMC} \sim 10^{41}, \quad F_{\gamma,LMC}(E_\gamma > 100\ \text{MeV}) \sim 2\text{–}3 \times 10^{-7}, \quad \frac{w_{cr,LMC}}{w_{cr,G}} \sim 1. \tag{19}$$

Although such an estimate may appear more reliable than Equations (16)—(18), it is clear

that it is accurate only to an order of magnitude. Further, the $L_{\gamma,\ \mathrm{LMC}}$ as given by either Equation (17) or (19) is not in contradiction with its value in the metagalactic models given by Equation (8) (see remarks after this equation). This is simply because the equality of energy densities (7), which must hold in the metagalactic models, is almost valid in Equations (17) and (19) also.

Thus it is clear that a value for $F_{\gamma,\ \mathrm{LMC}}(E_\gamma > 100\ \mathrm{MeV}) \sim (2\text{–}10) \times 10^{-7}$ photons $\mathrm{cm}^{-2}\ \mathrm{s}^{-1}$ would be compatible with both galactic and metagalactic models. A larger value like 10^{-6}, and preferably greater than $(2\text{–}3) \times 10^{-6}$, would rule out metagalactic models. So also will a very low value of $F_{\gamma,\mathrm{LMC}} \leqslant 10^{-7}$ even more strongly. At the end of their report, Houston, Riley & Wolfendale (1983) claim that if the latter case were true, the galactic models will also encounter difficulties. We also consider such a low value improbable, but if observed, it can in principle still be accommodated in galactic models under the assumption of a faster cosmic-ray exit in the LMC because of dissimilarity with the Galaxy.

As already mentioned, the ratio (9) in metagalactic models is fairly reliable and expected to be rather accurate. Therefore, a disagreement of the measurements with the result (9) will clearly disfavour metagalactic models (if the electronic contribution can be excluded). Note that the metagalactic models predict w_{cr} = constant in the entire galaxy; while in the galactic models the γ-ray flux is determined by the integral $\int n(r) w_{\mathrm{cr}}(r)\ \mathrm{d}^3 r$ over the galactic volume. Thus the galactic models can lead to Equation (9) only if the LMC and SMC are completely similar, in particular if their densities w_{cr} are equal. However, such an equality can hardly be expected to hold since the LMC and SMC differ greatly from each other as we discuss below.

Although the masses of neutral hydrogen in the two Clouds are almost equal, their visible sizes differ by more than a factor of 2 (6° and 2.5°; see Bok 1966) and the total masses obviously by more than a factor of 3 (Bok 1966) or perhaps even 10 (Long, Helfand & Grabelsky 1981). According to the latter, the frequency of Supernovae in the LMC is one per 110–340 years, while in the SMC it is only one per 1000 years (Mills *et al.,* 1982). Even more accurate are the ratios of explosion-frequencies in the Clouds and the Galaxy—they are $r_{\mathrm{LMC}} = 0.2$ (Houston, Riley & Wolfendale 1983) and $r_{\mathrm{SMC}} = 0.05$ (Mills *etal.,* 1982). Thus, if we accept the criterion used by the former, the mean cosmic-ray generation power in LMC is four times that in SMC. The total energy output from Supernovae is approximately a factor of 6–8 smaller in the SMC than in the LMC (Tarrab 1983).

All these fragmentary pieces of information cannot, without a detailed analysis, characterize clearly the differences between the Clouds. In our opinion, they show however that there is no reason to expect the fulfilment to a high accuracy of the equality $w_{\mathrm{cr},\ \mathrm{LMC}} = w_{\mathrm{cr},\ \mathrm{SMC}}$ in the framework of galactic models. Most probably, in galactic models $w_{\mathrm{cr},\ \mathrm{SMC}}$ would be considerably less than $w_{\mathrm{cr},\ \mathrm{LMC}}$, if we disregard possibilities like a substantial cosmic-ray overflow from one Cloud to the other along some magnetic tubes of force, *etc.* A determination of the ratio $F_{\gamma,\ \mathrm{SMC}}(>E_\gamma)/F_{\gamma,\ \mathrm{LMC}}(>E_\gamma)$ is therefore an important task, the outcome of which could clearly decide between the two models of cosmic-ray origin.

3. Antiprotons in the galactic cosmic rays

An understanding of the large flux of antiprotons found in cosmic rays near the Earth is an intriguing problem. In this paper we would like to supplement the model of galactic

antiproton sources (Ginzburg & Ptuskin 1981) with a discussion of the γ-radiation which inevitably arises in this model. But it seems reasonable to dwell first on the antiproton data and the model of Ginzburg & Ptuskin (1981).

The successful detection of antiprotons in cosmic rays has led to unexpected results (Golden *et al.* 1979, 1983; Bogomolov *et al.* 1979; Buffington, Schindler & Pennypacker 1981). The observed integrated antiproton flux was found to exceed significantly the flux of secondary antiprotons calculated from the standard model of cosmic-ray propagation in the interstellar medium (see, *e.g.* Stephens 1981a, b). Thus, the integrated intensity of secondary antiprotons in the leaky-box model with a mean matter thickness $X_e = \bar{p}vT_{cr} = 5$ g cm^{-2} traversed by cosmic rays in the Galaxy at energies $E_k = 0.1$–10 GeV is about $I_{\bar{p}} = 0.1$ (m^{2}s sr)$^{-1}$, (Gaisser 1982), while the measured value is 0.65 ± 0.2 (m^{2}s sr)$^{-1}$.

The energy spectrum of antiprotons was also unexpectedly different from the spectrum of secondary antiprotons generated in p-p collisions in that it was richer in low-energy particles. At energies of 130–320 MeV, the measured intensity (Buffington, Schindler & Pennypacker 1981) turns out to be larger by more than two orders of magnitude (the influence of solar modulation being disregarded). On the whole, the observed spectrum of antiprotons is quite similar to the proton spectrum in cosmic rays,

$$\frac{I_{\bar{p}}}{I_p} \simeq 6 \times 10^{-4} \text{ at } 1.4 \text{ GeV} < E_k < 13.4 \text{ GeV (Golden } et\ al. \text{ 1983)},$$

and

$$\frac{I_{\bar{p}}}{I_p} \simeq 2.2 \times 10^{-4} \text{ at } 0.13 \text{ GeV} < E_k < 0.32 \text{ GeV}$$

(Buffington, Schindler & Pennypacker 1981)

and unlike the spectrum of antiproton secondaries which falls sharply at low energies ($E_k \lesssim 3$ GeV).

Detailed data on secondary antiprotons generated in cosmic rays can be found, for example, in Stephens (1981a, b) and in Tan & Ng (1983).

Thus it is clear that the data on antiprotons are not in accord with the standard scheme which has been able to explain the high relative content of rare nuclei Li, Be, B, d, ^{3}He *etc.,* and also of positrons in cosmic rays. As is well known, all these particles are considered to appear as secondaries in the interaction of the proton-nuclear component of cosmic rays with the interstellar gas (or with the material in the cosmic-ray sources). The matter thickness necessary in this case is 5–10 g cm^{-2} at $E_k \sim 1$ GeV.

If we exclude exotic explanations of the anomalously large flux of antiprotons—such as their production during quantum evaporation of mini-black-holes (Kiraly *et al.* 1981), n–$\bar{n}$ oscillations (Sivaram & Krishan 1982), or acceleration of antimatter in galaxies (Stecker, Protheroe & Kazanas 1981), all of which for different reasons do not seem sufficiently convincing—there remains, perhaps, only one explanation: namely the generation of anti-protons as secondaries in compact sources with a high concentration of relativistic protons (Ginzburg & Ptuskin 1981; Eichler 1982; Mauger & Stephens 1983).

Taking the measurements mentioned above at face value, we discuss below a concrete model (Ginzburg & Ptuskin 1981) in which antiprotons are generated in young supernova remnants, and then in the next section present some estimates concerning the γ-radiation from such objects.

We assume that inside a supernova shell there exists a pulsar which accelerates protons at a rate proportional to its luminosity in low-frequency (magnetodipole) radiation, and that the power of the cosmic-ray source inside the shell is as given by Berezinsky & Prilutsky (1978; see also Apparao & Rengarajan 1971 and Cavallo & Pacini 1980):

$$L_i(t) = \frac{\lambda L_0}{(1+t/\tau)^2}, \tag{20}$$

where L_0 is the initial pulsar luminosity (at $t = 0$), λ the efficiency of acceleration and $\tau \simeq 5 \times 10^7$ s, the *initial* characteristic time of pulsar luminosity decrease due to magnetodipole braking.

Suppose that the protons accelerated by a pulsar leave the supernova shell quickly, *i.e.*, in a time $\delta t \lesssim t$. In leaving the shell, the thickness of the matter traversed by the accelerated particles is equal to

$$X(t) = \frac{3MK}{4\pi u^2 t^2} = 10^2 \left(\frac{K}{30}\right)\left(\frac{1.7\times10^7}{t}\right)^2 \left(\frac{M}{4\times10^{33}}\right)\left(\frac{10^9}{u}\right)^2 \text{g cm}^{-2}. \tag{21}$$

Here, M is the shell mass in g and u its velocity in cm s^{-1} and t is in s. The coefficient $K = \delta t\,(v/R_s)$ characterizes the entanglement of the particle trajectory inside the shell and is equal to the ratio of the time taken by a relativistic charged particle to pass through the shell to the mean free time; $R_s = ut$ is the shell radius at time t, and $v \simeq c$ is the particle velocity. In what follows we assume $M = 2M_\odot$, $u = 10^9$ cm s^{-1} and $K = 30$ (these values are somewhat different from those used by Ginzburg & Ptuskin 1981).

When the matter column density drops to $X_p \simeq 10^2$ g cm^{-2} (where X_P is the total attenuation path for relativistic photons), a considerable fraction of accelerated protons will have escaped. The age of the shell t_p at which it becomes transparent to relativistic protons, *i.e.*, $X(t_p) = X_p$, is

$$t_p = \left(\frac{3MK}{4\pi u^2 X_p}\right)^{1/2} \simeq 1.7\times10^7 \text{ s}. \tag{22}$$

The relative flux of antiprotons produced thus in a shell turns out to be essentially higher than the value $(I_{\bar{p}}/I_p)_l$ given by the leaky-box model. For $E_k \gtrsim 3$ GeV, we can write the following as an estimate:

$$\frac{I_{\bar{p}}}{I_p} = \alpha_p \eta \frac{X_p}{X_l}\left(\frac{I_{\bar{p}}}{I_p}\right)_l \tag{23}$$

where at $t_p > t_a \gg \tau$, the quantity $\eta = (X_a/X_p)\,[1 - (1 + X_p/X_a)^{-1/2}]$ with $X_a = 60\,[E_k\,(\text{GeV})]^{1/2}$ g cm^{-2} being the annihilation mean free path of the antiproton with an energy E_k; the quantity t_a is related to X_a by a formula analogous to (22), *i.e.*, by $X(t_a) = X_a$. For $\tau \gg t_p > t_a$ the value $\eta = t_p/\tau$. In Equation (23) it is assumed that pulsars do provide the fraction α_p of the total proton component of cosmic rays.

Proton generation in young supernova remnants with a large matter thickness not only makes it possible to obtain the required integrated antiproton concentration, but also allows for the enhancement of their intensity in the low-energy region by stochastic particle diffusion in energy, as well as by adiabatic and ionozation losses in an expanding turbulent shell. Besides, for such a large matter thickness, the anti-proton energy redistribution to the low-energy region due to the non-annihilating $\bar{p}$-p inelastic

interactions is also very important. This effect alone could explain a significant part of the antiproton intensity observed at low energies (Tan & Ng 1983).

Note that the adiabatic losses suffered by cosmic rays leaving the shell are not very large in the proposed model as compared, for example, with Mauger & Stephens (1983) and Berezinsky & Prilutsky (1978), who suggest a long confinement time for relativistic particles in shells. As a particle leaves the shell, the decrease in its momentum due to adiabatic expansion is described by the law $p(t) \propto 1/t$. Let a particle having a momentum $p(t_0)$ be emitted by the pulsar at a time t_0. In leaving the shell, the fractional decrease in $p(t_0)$ is equal to

$$\frac{|\delta p|}{p} = \frac{|p(t) - p(t_0)|}{p} = \frac{t - t_0}{t_0} = \frac{\delta t}{t_0} \simeq \frac{uK}{v}. \tag{24}$$

Hence

$$\frac{\delta p}{p} \sim 1 \left[i.e., \; p \sim p(t_0)/2 \right] \quad \text{when} \quad K = \frac{v}{u} = 30 \left[\frac{10^9}{u(\text{cm s}^{-1})} \right].$$

Now the quantity K is determined by physical conditions in the supernova remnant, or more precisely, by the structure of the magnetic field and turbulent motions in it. In Ginzburg & Ptuskin (1981) the value $K \simeq 10$ was used. A more probable value is perhaps the one that corresponds to $\delta t \sim R_s/u = t$, the exit time relevant for a shell with well-developed turbulence. The principal scale of turbulence in such a shell must be of the order of the system dimensions, and the velocity of random motions of the order of the shell expansion velocity. The relativistic particles are frozen in the strongly magnetized surrounding plasma and are carried out just for the time $\delta t \sim R_s/u$. It should be noted that intensive turbulence in a shell is already developed in the first 10^6 s after the explosion (Bodenheimer & Ostriker 1974). Thus the cosmic-ray leakage from a shell with a value of $K \simeq v/u$ will lead to a sufficiently enhanced low-energy antiproton spectrum, but, as is clear from Equation (24), not to strong adiabatic losses, which in turn would demand an increase in the source power.

4. Gamma-radiation from supernova-remnant sources of antiprotons

The data obtained from γ-ray astronomical studies also indicate that for supernova remnants to be antiproton sources, the cosmic-ray exit time δt must be relatively small *i.e.,* $\delta t/t \sim 1$ (remember that t is the age of the supernova remnant and $\delta t/t \sim 1$ when $K \sim v/u$). Otherwise, the γ-radiation would be excessive as discussed later.

The remnant pulsars, which accelerate protons according to the law (20) and provide the entire Galaxy with a cosmic-ray energy density $\alpha_p \omega_G$ must have an initial power

$$L_i(0) = \lambda L_0 = \frac{\alpha_p U_{cr}}{\nu_{SN} \tau} S \tag{25}$$

where $S \sim 1$ for $t_p \ll \tau$ and $S \sim t_p/\tau$ for $t_p \gg \tau$. U_{cr} 3×10^{40} erg s^{-1} is the total power in the Galaxy due to cosmic-ray sources; ν_{SN} is the galactic frequency of Supernovae leading to formation of pulsars surrounded by shells.

The total energy of cosmic rays contained in one shell and their energy density are equal to

$$W_{cr}(t) = L_i(t) t \frac{uK}{v} G; \quad w_{cr} = \frac{3W_{cr}}{4\pi (ut)^3} \tag{26}$$

where $G \simeq 1$ for $t \gg t_p$ and $G \quad t^2/t^2 = X_p/X\ (t)$ for $t \ll t_p$. Due to the decay of π^0 -mesons produced in the shell, the γ-ray luminosity of a young supernova remnant will be*

$$L_\gamma(t) = 4\pi\bar{q}_{\gamma,0} N \frac{w_{cr}}{w_G} = \frac{3\alpha_p K N \bar{q}_{\gamma,0} S G U_{cr}}{w_G u^2 v v_{SN} \tau t^2 (1 + t/\tau)^2} \tag{27}$$

where $N = M/(2 \times 10^{-24}$ g) is the number of gas particles (nuclei) in the shell (see Equation 2). Here and below, we assume the shell to be transparent to γ-radiation. This condition is fulfilled at times $t > 4 \times 10^6\ (M/2M_\odot)^{1/2}$ s. It is convenient to express the quantity L_γ in terms of the galactic luminosity in the diffuse γ-radiation $L_{\gamma,dif}$ which arises from the interaction of cosmic rays with the interstellar gas:

$$\frac{L_\gamma(t)}{L_{\gamma,dif}} = \frac{N w_{cr}}{N_G w_G} = \frac{\alpha_p X_p S}{v_{SN} \tau X_1} \frac{(t_p/t)^2 G}{(1 + t/\tau)^2}. \tag{28}$$

In order to derive the last equation, we have used Equations (21), (22), (25)–(27) and $X_1 = \bar{\rho} v T_{cr}$.

Thus, at $t \ll t_p$, $L_\gamma(t) \propto (1+t/\tau)^{-2}$. The shape of the light curve $L_\gamma\ (t)$ at this stage, when the shell is becoming transparent to relativistic protons, may be different since it depends on the details of the cosmic-ray propagation and the matter distribution in the shell. Relation (28) holds for turbulent cosmic-ray diffusion in a uniform shell.

For $t \gg t_p$, $L_\gamma\ (t) \propto (t_p/t)^2(1+t/\tau)^{-2}$; *i.e.*, at sufficiently large age ($t \gg t_p$, τ), $L_\gamma\ (t) \propto t^{-4}$.

For typical parameter values $\alpha_p = 1$, $\tau = 5 \times 10^7$ s, $t_p = 1.7 \times 10^7$ s, $v_{SN} = 1/30$ yr^{-1}, $X_1 = 5$ g cm^{-2}, we obtain the maximum flux value to be

$$L_{\gamma,max}(E_\gamma > 100\,\text{MeV}) \simeq 400\, L_{\gamma,dif} \simeq 6 \times 10^{44}\ \text{photons s}^{-1} \tag{29}$$

[the value of $L_{\gamma,\,dif}$ is 1.45×10^{42} photon s^{-1} (Houston, Riley & Wolfendale 1983).]

At a distance $R \simeq 3$ kpc we obtain from Equation (29)

$$F_{\gamma,max}(E_\gamma > 100\,\text{MeV}) = \frac{L_{\gamma,max}}{4\pi R^2} \simeq 0.6\ \text{photons cm}^{-2}\ \text{s}^{-1}. \tag{30}$$

The flux falls to the value 5×10^{-7} photons cm^{-2} s^{-1} for a supernova remnant age $t \simeq 30$ yr. Since there has been no supernova explosion in our Galaxy in the last 20–30 years (see, *e.g.,* the discussion by Palumbo 1983), there is no contradiction between a large γ-ray luminosity of antiproton sources in their bright phase and the available γ-ray data. A relatively fast cosmic-ray leakage from supernova remnants and a comparatively low value of $v_{SN}\tau \sim 5 \times 10^{-2}$ are the key factors which lead to a short duration for the bright phase of the γ-ray sources.

If the above picture of relativistic charged-particle outflow from a turbulent remnant is valid, and $K \sim v/u$, then one can obtain a sufficient antiproton flux with a low γ-ray luminosity. Indeed, as has been already mentioned, in order to explain the observed antiproton flux, it is necessary that the ratio t_p/τ is not small and may have to exceed 1/3 (see Equation 23). Only when this is so, do most of the particles accelerated inside the shell pass through a sufficiently thick layer of matter ($X \gtrsim X_n/3$). But then, it follows

* The contribution of bremsstrahlung to γ-radiation by the electron-positron component of cosmic rays inside the shell is negligible because large synchrotron and inverse Compton losses lead to very low concentrations of relativistic electrons and positrons in the source. At $t \sim 1$ yr the temperature is 10^4 K, for a magnetic field of ~ 0.17 Oe (Palumbo 1983).

from Equation (28) that the maximum luminosity of a supernova remnant also becomes large as it increases with t_p (for $t_p/\tau \gtrsim 1$).

Thus, the best choice is $t_p/\tau \sim 0.3$, which, with the chosen supernova remnant parameters, is realised just when $K \sim 30 \sim v/u$ (see Equations 21 and 22). It should be noted that such a choice of K also avoids catastrophic adiabatic cosmic-ray losses in the expanding remnant.

It is clear that the prediction of a large gray flux from a very young supernova remnant is an important feature of the above mechanism for antiproton production in compact sources and provides a test of the model. At least, some of the unidentified γ–ray sources observed by COS B may be just such sources as considered above (though not at a nearly stage of their evolution).We intend to return to this issue elsewhere. As the expected flux at even a metagalactic distance of say 1 Mpc, is $F_{\gamma,\,\max}(E_\gamma > 100\ \text{MeV}) \simeq 2 \times 10^{-6}$ photon $\text{cm}^{-2}\ \text{s}^{-1}$, the need to search for γ-ray flares associated with Supernovae is quite obvious (see also Berezinsky, Ginzburg & Prilutsky 1983).

In conclusion, our model for the production of a large number of antiprotons by compact sources such as supernova remnants inevitably requires an associated bright γ-ray phase. If such bright γ-ray sources are not discovered, another explanation for the large number of antiprotons observed in cosmic rays near the Earth is called for.

Acknowledgements

The authors would like to thank warmly V. Radhakrishnan and C. S. Shukre for helpful remarks on the manuscript.

References

Apparao, K. M. V., Rengarajan, T. N. 1971, *Proc. Indian Acad. Sci.,* **73,** 257.
Berezinsky, V. S., Ginzburg, V. L. 1981, *Mon. Not. R. astr. Soc.,* **194,** 3.
Berezinsky, V. S., Ginzburg, V. L., Prilutsky, O. F.1983, 18*th Int. Cosmic Ray Conf.,* Bangalore, **2,** 310.
Berezinsky, V. S., Prilutsky, O. F. 1978, *Astr. Astrophys.,* **66,** 325.
Bloemen, Y.B. *et al.* 1983, 18*th Int. Cosmic Ray Conf.,* Bangalore.
Bodenheimer, M., Ostriker, J. P. 1974, *Astrophys. J.,* **191,** 465.
Bogomolov, E. A., Lubyanaya, N. D., Romanov, V. A., Stepanov, S. V., Shulakova, M. S. 1979, 16*th Int. Cosmic Ray Conf.,* Kyoto, **1,** 330.
Bok, B. J. 1966, *A. Rev. Astr. Astrophys.,* **4,** 95.
Buffington, A., Schindler, S. M., Pennypacker, C. R. 1981, *Astrophys. J.,* **248,** 1179.
Cavallo, G., Paccini, F. 1980, *Astr, Astrophys.,* **88,** 367.
Dodds, D., Strong, A. W., Wolfendale, A. W. 1975, *Mon. Not. R. astr. Soc,* **171,** 569.
Eichler, D. 1982, *Nature,* **295,** 391.
Gaisser, T. K. 1982, *Proc. 2nd Moriond Astrophys. Meeting,* Les Arcs, France, BA-82-22.
Ginzburg, V. L., 1972, *Nature, Phys. Sci.*, **239,** 8.
Ginzburg, V. L. 1975, *Phil. Trans. R. Soc, London,* **A 277,** 463.
Ginzburg, V. L., Ptuskin, V. S. 1976, *Rev. Mod. Phys.,* **48,** 161; 675.
Ginzburg, V. L, Ptuskin, V. S. 1981, *Sov. Astr. Lett.,* **7,** 325.
Ginzburg, V. L., Syrovatskii, S. I. 1964, *The Origin of Cosmic Rays,* Pergamon Press.
Golden, R. L, Hozan, S., Mauger, B. G., Badwar, G. D., Lacy, J. L., Stephens,S. A., Daniel, R. R., Zipse, J. E. 1979, *Phys. Rev. Lett.,* **43,** 1196.
Golden, R. L., Nunu, S., Hozan, S. 1983, 17*th Int. Cosmic Ray Conf,* Paris, **13,** 89.
Hindman, J. V. 1967, Aust. J. Phys., **20,** 147.

Houston, B. R, Riley, P. A.,Wolfendale, A. W. 1983, 18*th Int. Cosmic Ray Conf.,* Bangalore, 1, 89.
Kiraly, P., Szabelski, J., Wdowczyk, J., Wolfendale, A. W. 1981, *Nature,* **293,** 120.
Long, K. S., Helfand, D. J., Grabelsky, D. A. 1981, *Astrophys. J.,* **248,** 925.
Mauger, B. G., Stephens, S. A. 1983, 18*th Int. Cosmic Ray Conf.,* Bangalore, **2,** 95.
Mills, B. Y., Little, A. G., Durdin, J. M., Kesteven, M. J. 1982, *Mon. Not. R. astr. Soc,* **200,** 1007.
Palumbo, G. G. C. 1983, *Space Sci. Rev.,* **36,** 293.
Sivaram, C, Krishan, V. 1982, Nature, **299,** 427.
Stecker, F. W., Protheroe, R. J., Kazanas, D. 1981, 17*th Int. Cosmic Ray Conf.,* Paris, **9,** 211.
Stephens, S. A. 1981a, *Astrophys. Space Sci.,* **76,** 87.
Stephens, S. A. 1981b, 17*th Int. Cosmic Ray Conf.,* Paris, **13,** 89.
Tan, L. C, Ng, L. K. 1983, 18*th Int. Cosmic Ray Conf,* Bangalore, **2,** 90.
Wolfendale, A. W.1981, in *IAU Symp.* 94: *Origin of Cosmic Rays,* Eds G. Setti, G. Spada & A. W. Wolfendale, D. Reidel, Dordrecht, p. 309.

J. Astrophys. Astr. (1984) **5,** 111–130

Spatial Interferometry in the Mid-Infrared Region

C. H. Townes *Department of Physics, University of California, Berkeley, CA 94720, U.S.A.*

(*Invited article*)

Abstract. The potential of high-resolution spatial interferometry for detailed mapping and precision astrometry in the mid-infrared region, somewhat analogous to interferometry now done in the microwave region, is discussed from an instrumental point of view. Some results from a prototype system and from tests of atmospheric properties are given. The design of a more advanced two-telescope system now under construction is outlined. This involves movable telescopes of 1.65 m aperture and of high precision, using heterodyne detection of infrared in the 10 μm atmospheric window.

Key words: interferometry—astronomical imaging—astrometry—infrared astronomy

1. Introduction

Some years of preparation and experimentation towards a large two-telescope infrared interferometer will come to fruition in 1984–1985. Preliminary work has included construction of a prototype system with a 5.5 m baseline, tests and observations of stars with surrounding dust shells, and astrometric tests. Technical developments include heterodyne detection techniques for the 10 μm atmospheric window, and engineering of two high-precision, 1.65 m telescopes specifically designed for infrared interferometry. These telescopes will be part of an observational system for both high angular resolution and precision astrometry. They will be mounted on trailers so that the inter-telescope spacing can be varied or the instrument moved from one observing site to another as appropriate. The two telescopes have Pfund-type optics in order to provide high rigidity and stability; the effective stability will also be much enhanced by careful monitoring of all critical telescope parameters with laser interferometers, which will allow correction for any changes. The purpose of the discussion below is to give the background, rationale, pertinent astronomical observations, expected performance, and general design of this interferometer system.

2. Characteristics of the atmosphere in the mid-infrared region

'Seeing' is a term to describe atmospheric effects on the quality of a stellar image in a telescope. As generally used, it is a somewhat subjective term, involving judgement of the average size of a stellar image and its motions while observations are being made. This blurring of a stellar image is primarily due to fluctuations in the index of refraction

of the atmosphere, which in turn are due to density variations associated with local temperature fluctuations, although they can be affected to a lesser extent by variations in the partial pressure of water vapour. For some time, there have been detailed theoretical discussions of the effects of local temperature fluctuations available on the basis of a randomly turbulent atmospheric model. However, only recently have good quantitative experiments been carried out to examine this type of theory in any detail, for example to check the wavelength dependence of seeing. At optical wavelengths, the seeing disk of a star in a large telescope may be anywhere from about $\frac{1}{4}$ arcsec to 10 or 20 arcsec in size. A disk no larger than 1 to 2 arcsec is generally considered to represent good seeing. This amount of angular blurring is approximately equal to the diffraction width due to the telescope aperture, $1.22\ \lambda/D$, if the aperture has a diameter of 10 cm. Here λ is the wavelength of visible light and D is the aperture diameter. The image size in a telescope of much larger aperture is clearly limited by seeing rather than by the diffraction beamwidth unless seeing is extraordinarily good. Based on random turbulence, for fixed turbulent conditions of the atmosphere, the diameter of a telescope which is diffraction limited increases as $\lambda^{6/5}$. Thus, at a 10 μm wavelength rather than the 0.5 μm of optical light, the diameter D of a diffraction-limited telescope should increase by a factor of approximately 38, to 3.8 m. In this case, the size of the seeing disk, which is proportional to λ/D, decreases as $\lambda^{-1/6}$. Hence, images produced at 10 μm should be expected to be approximately 1.8 times smaller in angular size than those in the visible region under similar atmospheric conditions. The latter conclusion was first clearly tested by Boyd (1978), who found that, at least under seeing conditions which gave a rather large stellar image, the ratio of the image size at 10 μm wavelength to that at 0.5 μm agreed very closely with the factor 1.8 predicted theoretically.

There are now a number of observational results using large telescopes near 10 μm wavelengths which make it clear that quite large apertures are in fact diffraction limited at this wavelength. Consider, for example, observations using heterodyne detection of 10 μm radiation. In this case a CO_2 laser is typically used as a local oscillator, which mixes with energy received by the telescope in a single diffraction mode, as in the case of normal heterodyne detection of microwaves in radio astronomy. Intensity of the received radiation has been examined carefully and found to correspond to essentially the full power striking the aperture. This shows that all of this power occurs in a single mode for telescopes as large as 3-m in diameter, the largest so far used with this mode of Operation (Betz 1981). A still more direct demonstration that telescopes of this size are diffraction limited at 9 μm has recently been provided by Bloemhof, Townes & Vanderwyck (1984). In this case, a linear array of very small infrared detectors was swept over the image of a star in the 3 m IRTF telescope on Mauna Kea. The resulting intensity distribution as a function of angle is shown in Fig. 1 for a star with no surrounding dust, α Boötis, and in Fig. 2 for two stars which are surrounded by dust shells. The width and shape of the central peak in all cases agree well with the diffraction pattern expected at 9 μm. The small bumps at the side of the image of α Boötis correspond to the side lobes of the diffraction pattern; those in the images of α Orionis and α Scorpii are primarily due to surrounding dust shells, and represent some of the fine-scale infrared intensity distribution which needs to be studied with interferometry.

Atmospheric effects in the infrared differ from those in the visible both because the wavelength is longer and because the index of refraction is different. Table 1 gives the index of refraction for dry air as a function of wavelength, and of water vapour of the same density. It can be seen that the index for refraction of air has a very much larger

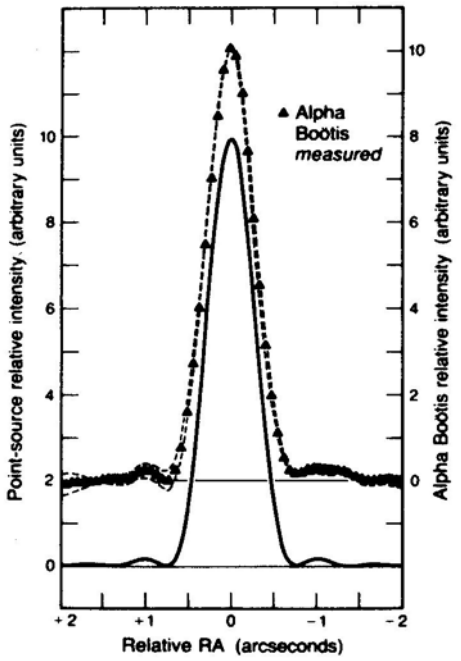


Figure 1. Profiles of telescope point-spread function as seen by a detector passing directly through the position of a point source. The solid line is the theoretical curve (Airy pattern) for a monochromatic 9 μm point source. Data points (offset for clarity) are the measured east-west profile through *a* Bootis (average of three scans). Dashed curves are 1-sigma error limits, *a* Bootis is believed to have negligible dust emission, and hence to be effectively a point source. The main peaks of the two profiles are virtually identical (Bloemhof, Townes & Vanderwyck 1984).

dispersion in the visible region than in the infrared. In addition, it is striking that in the centre of the 10 μm infrared region, the index of refraction for water vapour agrees precisely with that of air. Because of that, inhomogeneities in the humidity content of air will produce no disturbance in the propagation of 10 μm radiation if the air is at the same pressure and density. Of course, because of the difference in molecular weight, moist air of the same pressure and density as dry air must be at a slightly different temperature and cannot represent complete equilibrium. However, approach to an equilibrium from small temperature differences would be gradual and probably not troublesome so far as seeing is concerned.

The dispersion in the index of refraction in the visible region can be useful for possible correction to atmospheric seeing and refraction. For example, interferometry has been done simultaneously at two wavelengths in the visible region in order to correct some of the apparent variation in stellar position at a single wavelength (Shao, Colavita & Staelin 1984). It would be more difficult to provide a correction by this mechanism in the infrared region. On the other hand, the variation in water vapour content not only gives fluctuations in the index of refraction of air in the visible region, it also makes additional complications in the use of dispersion for the above type of corrections.

There is still another marked difference in the characteristics of seeing at 10 μm and those in the visible region—the timescales of fluctuations. Fluctuations involve motions

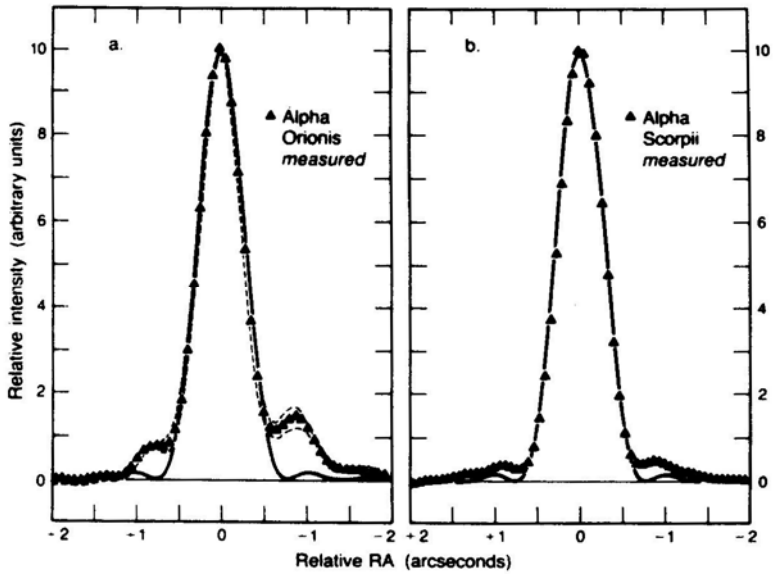


Figure 2. Profiles of supergiant stars with circumstellar dust shells (data points), superimposed on the theoretical point-spread function of Fig. 1 (solid line). (a) Measured east-west profile through α Orionis (average of three scans). Dashed curves are 1-sigma error limits. (b) Measured east-west profile through α Scorpii (single best-focus scan) (Bloemhof, Townes & Vanderwyck 1984).

Table 1. Index of refraction of dry air at NTP and of H_2O of the same density.

Wavelength μm	$(n-1)\times 10^4$ Air	H_2O
0.4	2.972	
0.5	2.934	4.02
0.6	2.915	
1.0	2.889	3.90
10	2.88	2.99
10.3	2.88	2.88
11	2.88	2.66
20	2.88	−3.00
10^4(1 cm)	2.88	89.6

of parts of the atmosphere within the path through which observations are being made. Their time variation is due generally to wind, or possibly convection. For given atmospheric conditions the displacements which affect infrared seeing must be larger than those which affect the visible region. The timescale for these displacements must be proportional to $\lambda^{6/5}$, or the distance over which the phase can change appreciably. The ratio $(\lambda_{IR}/\lambda_{vis})^{6/5}$ is approximately 38. Hence, instead of fluctuations which in the visible region are known to be characterized by frequencies of the order 10 Hz, at 10 μm wavelengths the frequencies should be in the range of $\frac{1}{4}$ Hz. This makes infrared seeing

fluctuations generally easier to compensate, assuming that the time for detection of fluctuations is adequately fast.

There are two other aspects of the atmosphere which are important to interferometry which are not quantitatively very well known. The first is the existence of atmospheric wedges which may persist over a long period of time. Such wedges give unequal paths through the atmosphere to the two receiving telescopes in the interferometer and thus change the apparent position of a star. Experience at the Naval Observatory in Washington has been that the RMS variation between apparent positions of stars from night to night is approximately 0.1 arcsec, substantially more than expected from random errors in individual measurements. These variations have in the past been attributed to atmospheric wedges. However, there are no other measurements on the atmosphere nor theoretical explanations which can account for wedges of this magnitude lasting over such a long period of time. Hence, it is now suspected that much of this variation is associated with temperature or other effects in the particular apparatus used. Pressure gradients in the atmosphere parallel to the earth's surface do exist, and are known from various meteorological measurements. In addition, gradients of humidity can provide some modest wedges in the optical path length through the atmosphere. Still other types of wedges can be produced by the heat generated in a large city (Hughes 1979). However, numerical estimates of the sizes of these effects indicate that they should be substantially smaller than 0.1 arcsec—probably not larger than 0.01 arcsec under most conditions. For the purposes of high precision astrometry, wedges which produce deflections of 0.01 arcsec are still significant and objectionable. Hence, methods of measuring and compensating for them are important.

For interferometers in the optical region, it is expected that compensation for atmospheric wedges can be obtained by measuring interferences at two different wavelengths, spaced widely enough so that the dispersion in the atmospheric index of refraction can be used to measure and correct for wedges. As noted above, the dispersion cannot be very easily used this way in the infrared region. Hence, the best method of measurement and compensation for atmospheric wedges when interferometers in the 10 μm region are used is probably to measure, by interferometry, the position of a star as the earth turns over a wide range of angles, from the zenith down to one or two hours above the horizon. If there are any atmospheric wedges which last over a long period of time, the increased effect they have on the apparent position (or a differential time delay to the two receiving telescopes as a star approaches the horizon) can be detected and the effect of a wedge thus measured. If wedges exist which last over a much shorter period of time, of course the atmospheric variations may be averaged in order to minimize any resulting error in position. It is the longer-term wedges, lasting over an entire series of measurements, which are somewhat more difficult to deal with, and to which the above discussion primarily applies. They will last long enough that their measurement and correction should be practical.

The second aspect of the atmosphere which is not known quantitatively but which is important to the success of interferometry with long baselines is the outer scale of turbulence. A randomly turbulent atmosphere has variations in density on all scales, including very long distances. However, the atmosphere is finite in size and hence there must be some maximum scale on which turbulence can occur. This limit is probably not set by the scale of the atmosphere itself, but rather by rates of equilibration of density variations, the size of obstructions on the earth's surface which would cause or damp out variations, or the distance above the earth at which measurements are made. In the

higher atmosphere, the outer scale of turbulence is generally believed to be set by the difference in buoyance of blocks of air of different temperature, and the consequent rapid rise or fall of air which has any large-scale variation of density. If this mechanism provides a limit to the scale of turbulence, one can estimate the outer scale of turbulence to be in the range 10 to 100 m.

For two telescopes separated by a distance ρ and observing the same stellar object, theory for a randomly turbulent atmosphere predicts that the difference in path lengths through the atmosphere along the two lines of sight should increase as $\rho^{5/6}$. Such a law should apply for distances smaller than the outer scale of turbulence, after which the increase of path length difference with ρ should be slower, and eventually reach a limiting value for very large ρ. In an interferometer operating at 10 μm wavelength, distances between telescopes can appropriately be as large as hundreds of metres. A baseline of 1 km would provide an angular resolution of about 10^{-8} radians, which is clearly within the range of resolution desired, and involves distances far greater than the outer scale of turbulence. Even 1/10 or 1/100 of this distance may still be considerably beyond the outer scale. Hence, knowledge of this outer scale of turbulence under conditions during which astronomical observations would normally be made is quite important to the future of infrared interferometry, and perhaps also for interferometry in the visible region. A recent experiment which will be described below indicates that this outer scale of turbulence is no more than about 2 m within a distance of a few metres from the earth's surface. Very likely the outer scale is larger in the higher atmosphere, but it could turn out to be no larger than 10 m, or possibly less. In that case, phase fluctuations between two telescopes in an interferometer would not increase greatly as the distance is increased beyond the 5.5-m separation for which we already have experience from a prototype 10- μm interferometer.

3. Sensitivity and the number of stars available for study

To obtain high resolution in an astronomical field of view, clearly an adequate signal-to-noise ratio must be obtained for signals from a very small solid angle. Furthermore, the accuracy of measurement of the position of a star depends both on the accuracy with which the parameters of baseline of an interferometer are known and on the signal-to-noise ratio obtainable in measurement of the interference fringe pattern. The latter determines the precision with which the phase of a fringe may be measured. In most infrared work, noise comes primarily from heat radiation impinging on the detector from sources associated with the telescope, earth, and atmosphere. In the case of heterodyne detection, there is a substantially larger noise due to quantization of the radiation field and the uncertainty principle. This sets a limit $\Delta n \Delta\phi \geqslant \frac{1}{2}$ to the uncertainty of numbers of quanta n and of phase ϕ. For linear detection, which allows determination of the phase of a wave within some limits ($\Delta\phi < \infty$) there must therefore be an uncertainty in the measurement of intensity. In most simple linear systems, and for normal heterodyne detection in particular, uncertainties $\Delta\phi$ are of a magnitude that makes this noise correspond to one random photon per mode of the radiation field per cycle per second. For a given bandwidth $\Delta\nu$, the noise power in a diffraction-limited telescope is hence $h\nu\sqrt{\Delta\nu}$. Thus, for a heterodyne detector using photodiodes, after a 1-s measurement the signal-to-noise ratio for the detection of a single star in a single

telescope is given by

$$\left(\frac{S}{N}\right)_{t=1} = \frac{\eta P(\nu)\sqrt{\Delta\nu}}{h\nu} \tag{1}$$

where η is the quantum efficiency of the detector and $P(\nu)$ is the power per unit bandwidth received by the telescope. $\Delta\nu$ is the total bandwidth of the system, or twice the bandwidth of the receiver since both heterodyne sidebands would normally be used. When the heterodyne signals from two separate telescopes are combined to provide an interference, the signal-to-noise ratio after one second for the combined signal is the square of the above, or

$$r = \left(\frac{S}{N}\right)^2_{t=1}.$$

For a single telescope, the signal-to-noise ratio increases as the square root of t, the time of observation; for an interference fringe, it increases as t. Hence, under ideal conditions where the phase of the fringe is subjected to no arbitrary fluctuations due to the atmosphere, the signal-to-noise ratio for an interferometer is

$$r = \frac{\eta^2 P^2(\nu)\Delta\nu t}{(h\nu)^2}. \tag{2}$$

Under less ideal conditions, when an imperfect atmosphere causes fluctuations in the phase of the fringes, the signal-to-noise ratio is necessarily poorer than that given by the above expression. A simple model of atmospheric fluctuations could assume a constant phase for a period of time τ_c and then a sudden arbitrary change of phase, or almost equivalently, an exponential distribution of the time during which the phase is coherent and is then followed by a sudden arbitrary change, *i.e.* the coherence time distribution given by a probability exp $(-t/\tau_c)$. These cases are equivalent to a number t/τ_c of independent measurements each of length τ_c, so that for such an atmosphere the signal-to-noise ratio is

$$r = \frac{\eta^2 P^2(\nu)\Delta\nu}{(h\nu)^2}(t\tau_c)^{1/2}. \tag{3}$$

It is of course assumed that $t > \tau_c$. For small baselines, τ_c may be very long and expression (2) applies; for long baselines it is more likely that expression (3) applies. Probably for many actual cases, the appropriate expression will be somewhere between (2) and (3).

In the ideal case where the quantum efficiency η is unity and the atmosphere produces no important disturbances of the phase, an acceptable signal-to-noise ratio of 5 after an averaging time of one second can be shown from expression (2) to be produced by a star of flux of $71/A$ Jy into the telescope. A is the area of the telescope in square metres. This assumes a double sideband width of 4000 MHz, which is practical in fast heterodyne detectors. After measurement time t of one hour (t = 3600), the flux required for an adequate fringe detection reduces to $1.2/A$ Jy. For the telescopes presently under construction, having an aperture of 1.65 m diameter, this is 0.6 Jy as a limiting flux after observations of one hour under ideal conditions. However, deviations from the ideal are important. The quantum efficiency for actual systems, including losses in the optics and imperfect detection, has normally been as high as η = 0.25 at a bandwidth of

3,000 MHz. A value $\eta = 0.25$ may hence be used for real estimates of sensitivity, although perhaps as much as a factor of two further improvement can be obtained in the future. The losses in sensitivity due to imperfect seeing are of course highly dependent on the particular seeing conditions, and are more difficult to specify quantitatively. For moderately good conditions, past experience indicates that τ_c is probably a few seconds for fairly long baselines. Very much longer values of the coherence time, τ_c, have been obtained with baselines of 5.5 m and, as was discussed above, the maximum frequency of fluctuation of phase is probably not greater than $\frac{1}{4}$ Hz under normal conditions. Hence, the value of τ_c of three seconds for baselines as long as many tens of metres is probably a conservative estimate for good conditions, and one can hope for much longer times under very favourable conditions. For observation times as long as one hour, this decreases the sensitivity of an interferometer by a factor of 5.9. Thus, a reasonably conservative estimate of the stellar flux for which a good ($r = 5$) signal-to-noise can be obtained after one hour's observation is 24 times greater than the ideal value given above, or 14 Jy.

The brightest stars in the mid-infrared provide a flux of ~ 3000 Jy at the earth in the 10 μm range. For example, the total infrared flux from α Orionis in the 10 μm region is 4800 Jy. However, only approximately 0.6 of this comes from the stellar disk itself, the remainder coming from the surrounding dust which for stars as close as α Orionis would normally be resolved out by an interferometer of moderate baseline. If 3000 Jy is the flux of the brightest stars, an interferometer of 1.65 m aperture operating under ideal conditions could detect stars 5000 times weaker, while the rather conservative estimate allowing for imperfect detection and atmospheric fluctuations would indicate that stars ~ 200 times weaker than the brightest ones could be detected after one hour's observation. The disk of α Orionis corresponds to a 10 μm magnitude of – 5, so that a star 200 times or 5.8 mag weaker would be + 0.8 mag at 10 μm. A hot star of this magnitude would have a diameter of about 0.002 arcsec, which is just the limit of resolution of the interferometer with a baseline of 1 km. This indicates that longer baselines would be useful only if more sensitivity is obtained.

It is easy to obtain a rough estimate of the total number of stars which can successfully be measured with an interferometer of the type discussed here. In the visible wavelength range, stars of magnitudes ~ 0 to about+ 6, or a range of 6 mag, can be seen by the naked eye. This is approximately the same range of intensities which would be well detected by the interferometer under construction, so that the number of objects for which it can be reasonably used must be comparable with the total number of visible stars, or a few thousands. For observations carried out over a period much longer than one hour, or for conditions closer to the ideal, substantially weaker objects in larger numbers should be observable. For astrometry, of particular interest is the number of stars easily measurable in the infrared region which are identical with those in the FK4 list of stars measured in the visible region. Also of interest is the number which overlap with the list of stars reasonably detectable by a radio interferometer. Examination of a short list of the brightest infrared objects indicates that about 30 per cent of these correspond to the FK4 list, and there would hence be a substantial overlap between any list of a few thousand stars detectable in the infrared and the FK4 list, or other stars typically used for astrometry in the visible region. So far, only rather few stars have been detected in the radio region, but probably all of those within one magnitude of α Orionis, one of the brightest red giants, can be detected as a result of thermal radiation alone. Almost all of those for which SiO masers have been detected

are detectable with an infrared interferometer; these are of some interest because the SiO masers are believed to be rather close to the stellar disk and hence to the central stellar position.

Heterodyne detection has two primary advantages for interferometry: (1) it is the most sensitive detection available for a very narrow bandwidth, and (2) after detection the resulting signals are in the radio range and can be amplified and used flexibly in a variety of ways without degradation of the signal-to-noise ratio. The possibility of amplification and multiple use of the detected signal can be particularly important if an interferometer involving an array of telescopes and use of multiple baselines is wanted. A very narrow band is useful in allowing a relatively easy variable relative delay between signals from the two telescopes in order to obtain a white-light fringe. Conversion to radio frequencies provides a further easing of the delay-time precision required in order to determine the phase accurately, as will be discussed below. If an interference is desired over a range of frequencies $\Delta\nu$, the total path length from a stellar source through the two individual telescopes to the point of interference must be equal to an accuracy substantially better than $c/\Delta\nu$. For a heterodyne bandwidth of 4 GHz, $c/\Delta\nu$ is 7.5 cm, so that a precision in the path length better than 1 cm is usually adequate. This is fairly easy to provide by delaying the signal after heterodyne detection in a variable-length RF cable. A broader bandwidth would be more demanding of such a delay line. For example, a 10 per cent bandwidth at 10 μm would require delay lines with a precision of about one wavelength, or 10^{-3} cm. This can be achieved, but must be done in the infrared region with the primary signal because the bandwidth is so large. For baselines as long as 100 m, a variable optical path to this precision represents a substantial undertaking.

For precise measurement of phase, which is important for astrometry, or for mapping of regions which are not symmetric with respect to inversion, the two path lengths involved in interferometry must also be maintained to a precision substantially better than a wavelength of the signal carrier, regardless of how narrow the bandwidth may be. Thus, paths of the infrared signal from star to heterodyne detection must be maintained to a precision $(\Delta\phi/2\pi)\lambda$, where $\Delta\phi$ is the limit in phase uncertainty desired and λ the wavelength. After heterodyne detection, the wavelength λ which applies is that of the radiofrequency, or about 7.5 cm, and the path length precision can be relaxed. If a variable delay line is constructed so that an interference may be obtained directly between the two infrared signals, then for accurate phase determination the delay-line must be controlled to a small fraction of a wavelength regardless of how narrow the bandwidth may be.

In spite of the convenience of heterodyne detection and the use of narrow bandwidths, very likely larger bandwidths and direct interference of infrared signals and their detection in photodetectors or photodiodes will eventually be wanted in order to obtain an increase in sensitivity. After some work and observations with heterodyne systems, this type of system will probably be very seriously considered.

For direct detection, *e.g.,* by a photodiode at the point where two infrared beams from two receiving telescopes interfere, expressions broadly similar to those above for the signal-to-noise ratio apply. However, expression (1) is then replaced by

$$\left(\frac{S}{N}\right)_{t=1} = \eta \frac{P(\nu)\Delta\nu}{h\nu \exp(h\nu/2kT)}. \tag{4}$$

Here the temperature T is the effective temperature of radiation leaking into the

telescope, which presumably would never be more than 300 K, the room temperature. For$T$= 300 K and frequencies in the 10 μm region, exp ($hv/2kT$) is approximately 10. In addition to this gain of 10 in sensitivity, the bandwidth may be much wider than in the case of heterodyne detection, perhaps as wide as 100 wave numbers, or 3×10^3 GHz. In principle, this allows an additional gain in sensitivity of 27. Thus, ideally, direct detection and use of a wide bandwidth can improve sensitivity by a factor of about 300 over that given above for heterodyne detection. This corresponds to a detection of stars 6 mag still weaker, or ones of magnitude about + 7 in the 10 μm region. Thus, it seems reasonable to expect in the future that interferometry can be done on stars at least 10^4 times weaker than the brightest, which would of course open up the possibility of study of a very large number of additional objects, including essentially all the stars normally used for astrometry, a large fraction of the infrared sources within dark molecular clouds, and the centres of external galaxies. It may also justify baselines larger than 1 km.

It is important to note that with narrow bandwidths, direct detection cannot—in the foreseeable future—come anywhere near achieving its theoretical sensitivity, which from expressions (1) and (4) can be seen in principle to be better than that for heterodyne detection even when narrow bandwidths are used. The presence of residual detector and amplifier noise will always make heterodyne detection more sensitive than direct detection at very narrow bandwidths. However, direct detection is attractive for the broader bandwidths, and while all of its potential will be difficult to achieve in a practical system, this potential is great enough to make a system with direct detection a very attractive future goal.

In the above discussion of signal-to-noise relations, we have considered primarily the problem of detection of interference fringes, not measurement of the fringe phase. Simple detection or measurement of fringe visibility is adequate to map any object with inversion symmetry. However, either for more complex mapping or for astrometry, fringe phase must be determined. For this purpose, the phase fluctuations due to the atmosphere must not be too large. (We assume here that no additional uncontrolled phase variations are produced by the optics employed.) Errors in phase measurements due to atmospheric fluctuations involve a number of different conditions, and it is useful to discuss the more important types of effects separately.

If phase fluctuations due to the atmosphere are negligible, uncertainty in phase measurement would be due primarily to detector or amplifier noise. When the observations have provided a reasonably good signal-to-noise ratio (*e.g.* $S/N \geqslant 5$), the uncertainty in phase is approximately $\Delta\phi = N/S$. For a projected baseline of length D, the uncertainty in angular position of the star is then

$$\varepsilon = \frac{\Delta\phi}{2\pi}\frac{\lambda}{D} = \frac{\lambda}{2\pi D}\left(\frac{S}{N}\right)^{-1} \tag{5}$$

In this case, it is clearly advantageous to make D as large as is practical. The size of D would be limited primarily by the uncertainty in the star's position before the measurement in question, which should be appreciably smaller than λ/D. For an initial uncertainty in stellar position of 0.1 arcsec, and for λ = 10 μm, D should hence be less than 20m.

A very different, and more common, case is where substantial phase fluctuations are produced by the atmosphere. If the phase fluctuations are sufficiently slow and the signal strong enough, the phase changes may be systematically followed and averaged.

In that case, the astrometric accuracy would depend only on how well the atmospheric variations average out, which is not presently known very well. However, if the signal is not strong enough to follow all phase changes, the phase error has a different character. While the distribution of atmospheric fluctuations is not well known, for present purposes a Gaussian distribution should be an adequate representation, with the probability of a phase error ϕ_{At} proportional to $\exp[-(\phi_{At}/\phi_0)^2]$ where ϕ_0 is a constant. If the signal-to-noise ratio is not low ($\gtrsim 5$), the phase error in this case is

$$\Delta\phi' = \frac{\exp(\phi_0^2/4)}{(S/N)},$$

where *S/N* would be the signal-to-noise ratio in the absence of atmospheric fluctuations. The angular error is then

$$\varepsilon' = \frac{\Delta\phi'}{2\pi}\frac{\lambda}{D} = \frac{\lambda \exp(\phi_0^2/4)}{2\pi D(S/N)}. \tag{6}$$

The error ε' in expression (6) has a minimum as a function of *D* because ϕ_0 increases with *D* If ϕ_0 has the form $\phi_0 = aD''$, then the minimum occurs at $\phi_0 = \sqrt{2/n}$. For a random atmosphere, *n* would be 5/6; beyond the outer scale of turbulence, *n* would be smaller. If $n = 5/6$, $\phi_0 = \sqrt{12}/5$, or 88.8°. Hence, for minimum error in position measurement, *D* should be increased until the atmospheric phase fluctuations are about 90°, in which case the phase fluctuations have degraded the precision by only a factor of about 1.8. Of course, there may be an error in the average phase of the atmosphere as well as variations which may be represented by the symmetric probability distribution $\exp[-(\phi_{At}/\phi_0)^2]$. This average error can be minimized only by longer averaging or, in case of a very long-standing atmospheric wedge, by careful observation of the variation in phase as a function of angle from the zenith.

The determination of phase as well as amplitude of interference fringes should provide, as in radio astronomy, complete mapping of arbitrary distributions of infrared intensity. The resolution of this mapping would be between ~ 0.2–0.002 arcsec, with the latter resolution adequate to map convection cells on the surfaces of the larger stars.

4. Experimental results with a small prototype system

As indicated above, a small prototype interferometer system was set up at the Kitt Peak National Observatory using the two auxiliary McMath solar telescopes (Johnson 1975; Storey 1979; Sutton 1979; Townes & Sutton 1981). The aperture of each was defined by a 75-cm mirror which was a coelostat, so that normally the projection effects made the aperture somewhat less than this mirror size. In addition, the telescopes were not built for high stability, since they were designed to be used primarily for solar spectroscopy. Nevertheless, they were convenient for a first test of heterodyne interferometry in the 10 μm region and provided extremely useful results.

The interferometer built around the auxiliary McMath telescopes was primarily used for the resolution of dust shells around stars. However, it also provided an opportunity to test the use of such systems for astrometry. Fig. 3 shows the measured phase of fringes from the star *o* Ceti compared with the theoretical position of fringes expected from an assumed baseline chosen to approximately fit the observed fringe rate. No

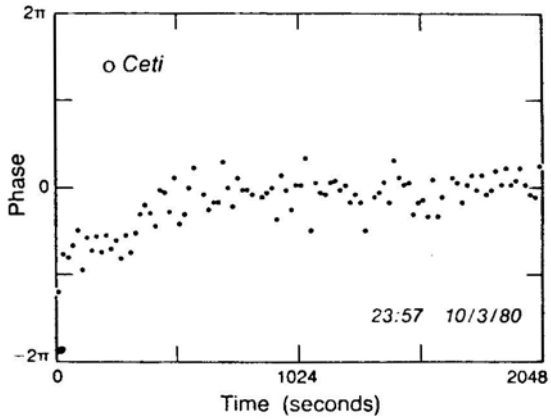


Figure 3. Observations of the phase of the interference fringe for *o* Ceti. Each data point represents 20 s of integration (Sutton, Subramanian & Townes 1982).

single fixed baseline would fit the fringes precisely over the time of observation, since either the atmosphere or the telescope positions and dimensions were changing slightly with time. Nevertheless, it can be seen that the fringes could be measured and tracked, and that the phase fluctuated randomly about a reasonably smooth progression of values. The signal-to-noise ratio obtained was in reasonable agreement with expectations from the theoretical expressions given above. In this particular figure, it is estimated that the fluctuations due to noise in the detection circuits are approximately one-half the amplitude of those observed, the remainder being presumably due to the atmosphere. A suitable average of these fluctuations should allow determination of the fringe phase to a precision of about 1/100 cycle on a statistical basis.

A series of measurements were made over a two-week period to examine the night-to-night consistency with which the separation between three stars could be measured (Sutton, Subramanian & Townes 1982). For this consistency check, the single fixed east-west baseline was adequate. The stars used were α Orionis, *o* Ceti, and R Leonis. These were at rather similar declinations and separated from each other in right ascension by approximately three hours. Each star was tracked for one to three hours before a switch was made to the successive star. Measurement of the phase of fringes for each of the three stars then allowed a series of measurements of separation of the stars in angle with an uncertainty of some integral number of fringes. Since the fringe spacing was approximately 0.38 arcsec and the precision of measurement turned out to be one quarter of this, the integral number was assumed to be the same each night. If a real measurement of stellar positions rather than a consistency check had been the goal, then a knowledge of the stars' position from other measurements to a precision appreciably better than 0.38 arcsec would have allowed the observations to provide a refinement of the relative stellar positions. However, we were primarily interested in the consistency of measurements during a night and from night to night. Fig. 4 illustrates the phase measurements obtained for the three stars as a function of time. These phases are all compared with a nominal fixed baseline which represented an approximate fit. However, it is clear from Fig. 4 that the baseline could not have been constant. There is some evidence of cycling of baseline orientation on a timescale approximately equal to the rotation of bearings, and in addition, of a general change of the baseline with time—

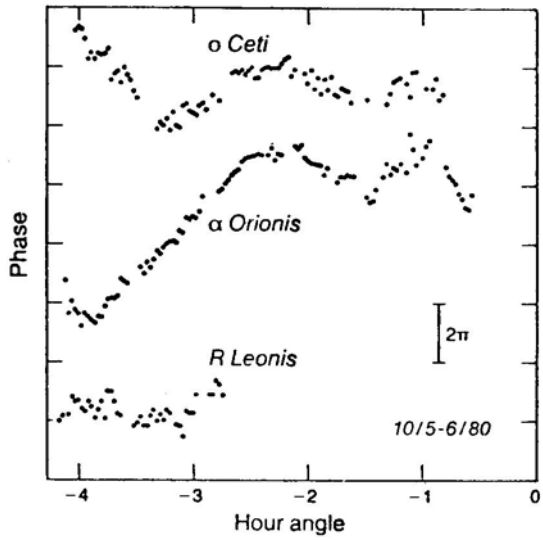


Figure 4. Phase measurements on *o* Ceti, α Orionis, and R Leonis for a single night (Sutton, Subramanian &Townes 1982). Each point represents 100 s of integration. The deviations at large hour angles are probably due to systematic mechanical distortions.

probably due to temperature changes, differences in rotation axes of the telescopes, or differences in the elastic distortion of the telescopes as a function of position. Even though the effective baseline changed as a function of the direction of pointing, it was hoped that the distortion of telescopes would be similar for each star if phase comparisons were made on the same night at identical hour angles, since the stars were not very different in declination and could be measured at the same hour angles. Fig. 4 indicates that the patterns of fringes did in fact reproduce to some extent with hour angle, particularly near the zenith. The differences in fringe phase were measured at all points of similar hour angles at which fringe phases could be obtained. Tables 2 and 3

Table 2. Differences in right ascension between *o* Ceti and α Orionis measured with prototype interferometer (Sutton, Subramanian & Townes 1982).

Date	Average positional difference (arbitrary zero point) arcsec	Standard deviation of samples arcsec	Number of samples
Sept. 22–23	0.03	0.11 (in 512 s)	10 (512 s each)
Sept. 24–25	0.25	0.05	4
Sept. 26–27	0.13	0.18	12
Sept. 27–28	0.18	0.10	9
Sept. 28–29	0.06	0.10	12
Oct. 1–2	0.06	0.05	8
Oct. 2–3	0.05	0.10	6
Oct. 3–4	0.24 (0.04)	0.08 (0.16)	8
Oct. 4–5	0.07	0.04	4
Oct. 5–6	0.16	0.04	8
Total	0.12	0.08 (night-to-night)	10 nights

Table 3. Differences in right ascension between α Orionis and R Leonis measured with prototype interferometer (Sutton, Subramanian & Townes 1982).

Date	Average positional difference (arbitrary zero point) arcsec	Standard deviation of samples arcsec	Number of samples
Oct. 1–2	0.11	0.04 (in 512 s)	7 (512 s each)
Oct. 2–3	−0.07	0.06	4
Oct. 3–4	0.10	0.09	6
Oct. 4–5	−0.02	0.04	4
Oct. 5–6	−0.06	0.06	7
Total	0.01	0.09 (night-to-night)	5 nights

give results of such measurements. It is seen from Table 2 that on a given night the RMS variation between α Orionis and *o* Ceti corresponded to approximately 0.08 arcsec, and that the variation from night to night during the ten nights of observation was essentially the same. Only five nights of observation were obtained for R Leo. However, Table 3 shows that also in this case the fluctuations in measured separation between α Orionis and R Leonis were almost identical in magnitude to those of the separation between α Orionis and *o* Ceti. Since normally the variations from night to night are about 0.1 arcsec with the present best available zenith tubes, these results are quite encouraging. Nevertheless, they are primarily limited by the structure of the telescopes used, which made the effective baseline very sensitive to temperature, to pointing position, and perhaps to the positions of individual balls in the bearing races. Hence, these variations should be regarded as upper limits, and it is believed they can be very much improved by telescopes especially designed for an astrometric interferometer. It is reasonable to expect a precision of 0.01 arcsec.

5. Telescope design and construction

A general arrangement of the telescopes which are presently under construction is indicated in Fig. 5. Each involves a 2-m flat mirror rotating about two axes and a 1.65-m paraboloid, with the optic axis horizontal in order to make the telescope sturdy and compact. Such an arrangement has been called a Pfund-type telescope. These telescopes are compact enough that the entire system can be mounted on a trailer and moved in order to vary baselines. This mobility allows them to travel over highways and hence to move from one observatory to another as need arises. The two telescopes would be moved into position by a tractor, and then each mirror set on kinematic mounts on a concrete pad on the ground, so that during observation they are not supported by the trailer. Probably, observations will be made without changing telescope positions over a period of a week or so before any telescope is moved to a new position in order to provide a new range of baselines.

As with almost all large telescopes, some part of the sky is obscured by vignetting. The flat mirror is moved in azimuth around a vertical axis ± 55° and in altitude around the horizontal axis—which approximately intersects the vertical axis—from 0° to 55°. There is, furthermore, a limit of 65° on the total combined angle of motion from the optic axis. This allows a full view of the horizon to either side of the telescope, and a view

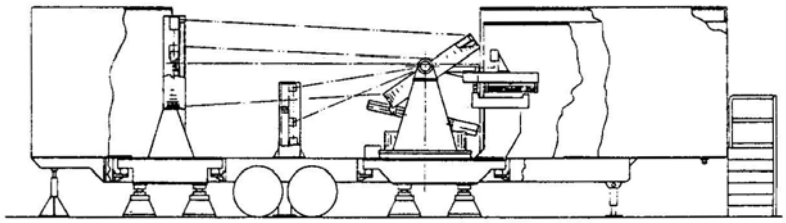


Figure 5. Schematic diagram of the Pfund-type telescope mounted on a trailer. Lines between mirrors represent beams of laser interferometers used for steering and for path-length monitoring. The post between the mirrors is of invar and set in bedrock to provide a reference position, and lines between it and the movable mirror are laser interferometer beams which monitor the telescope position with respect to bedrock.

overhead to 20° beyond the zenith. Separation between the mirrors is 4.8 m, so that the paraboloid subtends an angle of ± 10° from the axis at the surface of the flat and obscures only a small part of the sky. Any direction in the sky can be observed with almost full aperture by appropriate orientation of the axis of each telescope. However, it is likely that no more than two orientations would normally be required or used, the two being with the optic axes pointed in one of two opposite directions, probably north and south.

Optical paths through each telescope and the baseline are continuously monitored with helium-neon laser interferometers to a precision better than 0.5 μm. This monitoring will be described more fully below.

The local oscillators on each telescope are CO_2 lasers, which are locked in phase in order to maintain a fixed relative phase of detection for the two telescopes. This is done by sending a beam from the laser on one telescope to the other telescope where it is compared with the radiation from the second laser and the phase of the latter is controlled with a fast, feedback circuit. A similar arrangement was used in the prototype system built at the Kitt Peak National Observatory. However, since baselines will be long enough for appreciable changes to occur in the optical path between the two telescopes (*e.g* . due to atmospheric changes), part of the CO_2 laser beam will be returned along the original path and its relative phase on return automatically adjusted by a variable element in the path. This will ensure constant phase difference between the two telescopes.

At present, the four large mirrors of the telescopes are in the process of being ground and polished, and the mirror mounts and rotating axes are under construction. It is hoped the two telescopes can be finished by the fall of 1984 and initial tests be made shortly thereafter.

6. Control of the interferometer baseline and of telescope tracking

Positions of the telescopes are monitored with respect to bedrock and the optical path-lengths within the telescopes are monitored by the use of helium-neon laser interferometers. In addition, the large flats are pointed by use of interferometric measurements between the parabola and the flat in order to determine angular positions. Pointing of the telescope depends only on the orientation of the optic axis of

the fixed parabola and the relative angle between parabola and flat, which is measured by four interferometer beams on the left, right, upper, and lower edges of the mirrors.

We have examined the precision with which distances of a few metres can be measured in the open atmosphere at locations and under conditions during which astrometry might reasonably be carried out. The results showed that fluctuations in atmospheric density will allow measurements of the optical path-lengths in the telescope to adequate precision and provide excellent guiding of the flat (Gibson *et al*. 1984). In order to make measurements of fluctuations in optical path-lengths over the dimensions of the telescope, two Michelson-type interferometers were set up in the open atmosphere using a helium-neon laser as a source. One arm of each interferometer was quite short while the other was of 1-m path-length horizontal to the ground and at various distances above the surface. Measurements were made at the Lick Observatory at various times and places during three different nights, and with varying seeing conditions in the nearby 3-m Shane telescope. Fig. 6 illustrates the nature of the observed path-length fluctuations. It can be seen that there was a substantial correlation in the fluctuations between two paths separated by not more than one or two metres. However, the amount of correlation and the frequency of these variations indicated clearly that the outer scale of turbulence at distances only a few metres from the ground was no greater than about 2 m. For distances much larger than this, most of the fluctuations would hence be uncorrelated. Tables 4 and 5 show that the RMS fluctuations in our 1-m path-lengths were typically about one quarter wavelength, or 0.15 μm. The most remarkable results of this series of measurements was the relative independence of the size of the RMS fluctuations on varying conditions—the direction and amount of wind, location of the measurements on the mountain, and seeing observed in the nearby telescope. Since the amount of correlation between two paths was not high for distances greater than about two metres, a good approximation for distances greater than one metre is to assume independence of the various fluctuations, so that the RMS fluctuation over a longer path is approximately $1.5 \times 10^{-6} \sqrt{L}$ cm, where the path length L is measured in centimetres. Differences in two nearby pathlengths would be no more than $\sqrt{2}$ times this value, and somewhat less if there is correlation between some of the fluctuations in the two paths.

To monitor the interferometer baseline, an invar post will be set in bedrock at a position between the two mirrors. From a triangle of three positions attached to this post, three interferometer beams, each separated by an angle of about 30°, will converge on a cat's-eye retroreflector attached to the flat approximately at the centre of rotation. This establishes a convenient position within the telescope for which small motions due

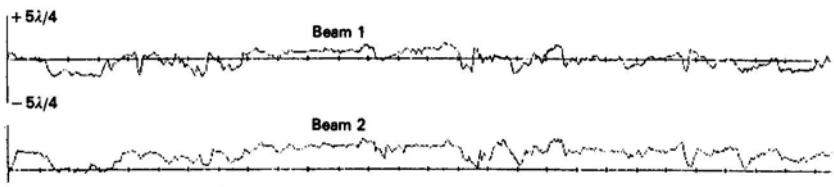


Figure 6. Variation in optical path-length of two 1-m horizontal distances approximately one metre and two metres above ground surface. Ordinates represent path length variations, with scale maxima of ± 5/4 λ, where λ is 632.8 nm. Abscissae represent time, with maximum time of 100 s (Gibson *et al*. 1984).

Table 4. Fluctuations in optical path-lengths, due to atmospheric density variations of two horizontal, fixed distances one metre in length. RMS fluctuations are given in fractions of a wavelength of the interferometer light used (632.8 mm). Note that in many cases the fluctuations in the difference between light paths are less than those of a single path, showing correlations in atmospheric density variations (Gibson *et al.* 1984).

Height of light paths in metres		Light path orientation with respect to wind					
		Perpendicular to wind velocity			Parallel to wind velocity		
Lower path	Upper path	Number of runs	Average fluctuations in individual paths	Fluctuations in the difference between two separate paths	Number of runs	Average fluctuations in individual paths	Fluctuations in the difference between two separate paths
1.09	1.75	7	0.17 ± 0.02	0.12 ± 0.01	10	0.16 ± 0.01	0.14 ± 0.02
1.09	2.77	9	0.18 ± 0.02	0.20 ± 0.02	4	0.22 ± 0.03	0.26 ± 0.04
2.16	2.77	3	0.26 ± 0.04	0.15 ± 0.01	0		

Table 5. Grand averages of fluctuations in horizontal light paths under various conditions. Note similarity of results under all conditions (Gibson *et al.* 1984)

	Number of runs	RMS Fluctuations in units of $\lambda = 0.6328\,\mu$m
All runs perpendicular to wind velocity	49	0.20
All runs parallel to wind velocity	29	0.18
All runs 1.09 m above ground level	39	0.18
All runs 1.75 m above ground level	17	0.20
All runs 2.77 m above ground level	19	0.23
All runs with seeing of 1–2 arcsec	29	0.20
All runs with seeing of 3–4 arcsec	44	0.18
All runs with seeing of 5–8 arcsec	2	0.25

to such errors as bearing run-out or temperature drifts can be measured with respect to bedrock. Tides in bedrock typically produce a relative motion between two points about 1 μm per 100 m distance. Other, long-term, variations in good quality bedrock are less than about 1 μm per 100 m over a few days. Hence, a 100-m baseline in bedrock should be fixed in direction on earth during any moderate period of observation within a precision of 1 μm relative motion of the two ends, or an angle of 2×10^{-3} arcsec. Interferometers attached to the invar post should monitor the relative position of the telescopes with respect to bedrock to a precision of about 0.5 μm. While there may be changes in position of the fiduciary cat's-eye on the flat due to bearing errors, temperature changes, or mechanical strains, these are monitored and allowed for by continuous correction of the baseline. While no large changes of this type are expected, any change up to one or two millimetres would be easily corrected for by simultaneously recording data and corrections.

The optical path-length between parabola and flat is continuously monitored by a helium-neon interferometer beam approximately parallel to the optic axis, between the parabola and the above cat's-eye near the centre of rotation of the flat. The optical path to the focus behind the flat is similarly monitored by another interferometer beam from the parabola to the focal table. Since optical and infrared path-lengths both depend on atmospheric density, and in a slightly different way, barometric pressure, temperature, and humidity also need to be monitored, though not to a high precision.

For astrometry, the most critical parameters—in addition to a good atmosphere—are the optical path-lengths within the telescope, and changes in baseline length or orientation during observations. The former should be monitored and any changes corrected by the above method to a precision better than 0.5 μm, and the latter monitored and corrected for to a precision better than 1 μm.

Good pointing accuracy of the telescopes is not critical for astrometric precision. On a visible star, pointing accuracy can be monitored and corrected by an observer and for a non-visible star an infrared tracker may be used. Nevertheless, accurate pointing is a considerable convenience, particularly for measuring objects inside a dust cloud. Hence, servos controlled by the interferometric measurements of relative changes in distance between the two mirrors will be used to control the pointing to a precision of 0.1 arcsec. We hope to have tracking to this precision over periods at least as long as 15 min in order to minimize guiding problems for the operator.

7. Summary and discussion

It seems clear that interferometers can be constructed in the infrared region with baselines up to hundreds of metres which will allow mapping and astrometric measurements parallel to those which are now undertaken with interferometers in the microwave region. Two mobile telescopes of 1.65 m aperture are now under construction for this purpose. Telescope pointing will be controlled and all the critical dimensions monitored by He-Ne distance interferometers. Initially, detection will be done by heterodyne techniques, using CO_2 lasers as local oscillators and very fast HgCdTe photodiode detectors.

The principal uncertainties in performance of such an interferometer come from uncertainties in our knowledge of atmospheric path-length fluctuations. However, tests with a prototype interferometer system and other studies of the atmosphere indicate that a good performance can be obtained in the 10 μm atmospheric window. Use of the instrument when built will provide information on atmospheric fluctuations to a precision and over path-length separations which have not previously been testable, and hence a good evaluation of future potentialities for this type of astronomy. A few thousand infrared objects should be available for study with the instrument planned. It is expected to provide resolution in some cases of angles as small as about one milliarcsec and an astrometric precision of about ten milliarcsec. Since the instrument will measure fringe phase as well as intensity, maps of complex infrared fields may be obtained with a very high angular resolution.

It is hoped that experience with the two telescopes under construction will justify further developments, which are likely to proceed in two directions. Firstly, an optical delay line may be constructed and a direct detection technique may be used, which, while more awkward than heterodyne detection, will allow a considerable increase in sensitivity and hence the examination of additional astronomical objects, including a number of extragalactic ones. Secondly, the heterodyne detection will allow easy use of multiple element interferometers and of phase closure, and perhaps experience will justify enlargement of the interferometer to a number of telescopes. This would have the same advantage as a multiple element radio interferometer, for which the number of baselines and hence the speed increases as $n(n+1)/2$, where n is the number of telescopes.

Acknowledgements

Research and development reported here has been assisted by the U.S. Office of Naval Research under Contract N00014-82-C-0700 and by the California Space Group under grant CS 39–79.

References

Betz, A. L. 1981, *Laser Spectroscopy V,* Eds A. R. W. Mckellar, T. Oka, B. P. Stoicheff, Springer-Verlag, Berlin, p. 81.

Bloemhof, E. E., Townes, C. H., Vanderwyck, A. H. B. 1984, *Astrophys. J.,* **276,** L 21.

Boyd, R. W. 1978, *J. Opt. Soc. Am.,* **68,** 877.

Gibson, G., Heyman, J., Lugten, J., Fitelson, W., Townes, C. 1984, Preprint.

Hughes, J. A. 1979, in *IAU Symp. 89: Refractional Influences in Astrometry and Geodesy,* Eds E. Tengström and G. Teliki, D. Reidel, Dordrecht, p. 13.
Johnson, M. A. 1975, *Ph.D. Dissertation,* University of California.
Shao, M., Colavita, M., Staelin, D. 1984, Preprint.
Storey, J. W. V. 1979, in *Proc. IAU Coll 50: High Angular Resolution Stellar Interferometry,* Eds J. Davis and W. J. Tango, University of Sydney, p. 20.
Sutton, E. C. 1979, *Ph.D. Dissertation,* University of California.
Sutton, E. C., Subramanian, S., Townes, C. H. 1982, *Astr. Astrophys.,* **110,** 324.
Townes, C. H., Sutton, E. C. 1981, *Scientific Importance of High Angular Resolution at Infrared* and Optical Wavelengths, Eds M. H. Ulrich & K. Kjär, ESO, p. 199.

J. Astrophys. Astr. (1984) **5,** 131–138

Spectra of the Ammonium Radical: The Schüler Bands

G. Herzberg *Herzberg Institute of Astrophysics, National Research Council of Canada, Ottawa, Ontario, Canada K1A 0R6*

(*Invited article*)

Abstract. The main bands of the Schüler system of ND_4 and NH_4 have been observed at high resolution. On the basis of these spectra, Watson, in a separate paper, has analysed the ND_4 main band showing that it represents a $^2F_2 \rightarrow {}^2A_1$ transition of a tetrahedral molecule. The observed wavenumber data for both ND_4 and NH_4 are presented; the latter have not yet been analysed. Isotopic bands for $^{15}ND_4$,$^{14}ND_3H$, $^{14}ND_2H_2$,$^{14}NDH_3$ have also been obtained and as previously pointed out confirm the assumed carrier of the spectrum. The much weaker bands accompanying the main Schüler band on the short and long wavelength sides are photographed at medium resolution. The interpretation of these bands in terms of the vibrational levels of upper and lower states is briefly discussed.

Key words: molecular spectra—ammonium radical

1. Introduction

Twenty-one years ago we carried out some experiments at the National Research Council of Canada to find an absorption spectrum of NH_4, the ammonium radical. These experiments were motivated by the thought that NH_4, if it exists as a free radical, must have absorption bands in the yellow or green region analogous to the D lines of Na with which NH_4 is iso-electronic and that these absorption bands might account for the diffuse interstellar lines. In these old experiments, diffuse absorption features were looked for in the flash photolysis of NH_3 mixed with H_2 in the hope that NH_4 could form from the photodissociation products of NH_3, that is, either from NH_3 + H or from NH_2 + H_2. Not surprisingly (in hindsight) these experiments were unsuccessful. It was only 20 years later that emission spectra of NH_4 and its isotopes were identified (Herzberg 1981, referred to here as Paper 1; Herzberg & Hougen 1983, referred to here as Paper 2): they are the Schuster and Schüler bands of ammonia that had been known for 109 and 26 years respectively, but not recognized as spectra of the ammonium (NH_4) radical until our work. The NH_4 radical, like H_3, may be called a Rydberg radical since only the Rydberg states are stable; the ground state is unstable.

None of the emission bands of NH_4 coincides with any of the diffuse interstellar lines. Nevertheless the NH_4 spectrum may well have an astronomical importance since NH_4^+ is almost certainly present in the interstellar medium and the recombination e + NH_4^+ will at least temporarily lead to Rydberg states of NH_4 which might be observed in emission. In addition, NH_4 may play a role in cometary spectra: an

unidentified feature (Spinrad 1980, personal communication) in the spectrum of Comet Bradfield (1979b) does coincide with the head of the Schüler band of NH_4. Needless to say that the spectra of ammonium have considerable interest from the point of view of studies of molecular structure.

Very recently Watson (1984) has succeeded in analyzing the main Schüler band of ND_4 on the basis of advance information on some of the high resolution spectra obtained in the present work. Here we shall describe these and other more recent spectra and discuss their interpretation in a preliminary way. A more detailed interpretation of the weaker bands will have to await the production of spectra of higher resolution.

2. Experimental methods

The methods suitable for producing the spectra of Rydberg radicals vary greatly. While the only way in which the H_3 radical has been produced is by a Cu hollow cathode discharge, such a source does not produce the spectrum of NH_4. Rather, a discharge at high pressure (see below) is needed. On the other hand the spectrum of He_2, the oldest known Rydberg radical, can be produced by either of these methods.

The following methods have proved useful for the production of the spectra of the ammonium radical.

(1) The simplest method is an ordinary Tesla discharge (see McKeever *et al.* 1979) as produced for example by a leak tester or preferably by the sturdier model BD20 (Electro-Technic Products Co., Chicago). The intensity of such a discharge is, of course, rather weak and fairly long exposure times are required. Such a Tesla discharge can be used with pressures up to 500 torr; several of the spectra reproduced below were obtained in this way. This method is obviously not suitable for photo-electric recording of the spectrum or for studies with an infrared spectrometer.

(2) A second method of exciting these spectra is by means of an ozonizer discharge as described by Wulf & Melvin (1939) and D'Silva, Rice & Fassel (1980). In its simplest form the space between two concentric glass tubes is filled with NH_3 or ND_3; two pieces of Al foil, one on the outside of the outer tube and one on the inside of the inner tube, serve as electrodes. It can be operated by ordinary a.c. from a transformer or more effectively by a radio-frequency resonance circuit (4 MHz, 100 watt). In later experiments a quartz cell with a rectangular cross section was used (2 mm × 12 mm) to the sides of which the electrodes were fastened. These ozonizer discharges can be operated at pressures up to atmospheric or even above.

(3) A third method is the use of ordinary a.c, d.c. or microwave discharges in tubes of Standard design but these cannot be operated at pressures much above 40 torr and are therefore not very effective in exciting the Schüler bands. Also, while the first two methods show an enormous intensity difference in favour of ND_4 compared to NH_4 (up to a factor of 20) such a difference does not arise with the d.c, a.c. or microwave excitation. However, this method is superior and was used for studies in the infrared since unlike the Tesla discharge it does not upset the electronics of the spectrometer.

(4) A fourth method favoured by Schüler, Michel & Grün (1955), particularly for what we now call the Schüler bands, is excitation by an electron beam. We have set up this method using a commercial 'electron gun' (Debe Associates, Burwell, U.K.) and confirmed its effectiveness. With ND_3 at 500 torr the electron beam shows a deep red

colour, corresponding to the 6749 Å main band which dominates the spectrum. Here also the factor 20 arises in the intensity ratio of ND_4 relative to NH_4.

(5) A fifth method used an apparatus developed by Huber & Sears (1984) in this laboratory, following Droege & Engelking (1983), to study emission spectra of free radicals in a supersonic jet in order to reduce the rotational temperature and thus simplify the spectrum. A corona-like d.c. discharge was produced between an anode (at about 800 V), positioned just behind the nozzle, and the wall of the expansion chamber acting as the cathode. A low-temperature spectrum of ND_4 was readily produced.

(6) Very recently, Hunziker and his associates at IBM Research Laboratory, San José, have succeeded in obtaining the main Schüler band in absorption by a laser frequency modulation technique using photochemical production of ND_4 from ND_3 by photosensitization with Hg (Whittaker *et al.* 1984).

(7) Still more recently, Huber & Alberti (1984) in our laboratory have obtained the main Schüler band of ND_4 in absorption by the flash discharge technique using a mixture of ND_3 and Ar. The trick to accomplish this result was to time the source flash (Lyman tube) before the final electrical breakdown in the absorption tube.

3. The Schüler band under high resolution

Unlike the Schuster bands, the Schüler bands are strongest at the highest pressure at which spectra can be taken (of the order of 1 atm). Under these conditions the principal Schüler band (6636Å for NH_4, 6748 Å for ND_4) is almost a factor 100 stronger than the principal Schuster band. With the additional factor 20 in favour of ND_4 it was relatively easy to obtain spectra of the Schüler band of ND_4 with our 10-m grating-spectrograph. One of the spectra is reproduced in Fig. 1 in two sections. It can be seen that many lines even at this dispersion appear quite sharp; broad lines are in all probability groups of unresolved lines. For NH_4, because of the lower intensity, the highest resolution which we could obtain was that of a 3-m grating-spectrograph. Such a spectrum is shown in Fig. 2. It can be seen that here the lines appear broader than those of ND_4 in Fig. 1 in spite of the lower dispersion. The similarity in the structure of the bands of the two isotopes is very obvious. In particular, the spacing of the two band heads at the longward end is very nearly the same. It seems very likely that the doubling of the head corresponds to spin-splitting similar to that in the Na D lines.

The measured wavenumbers of the lines of the Schüler bands of ND_4 and NH_4 are listed in Tables 1 and 2 respectively. Herzberg & Hougen (1983) attempted an analysis of the ND_4 Schüler band on the basis of the assumption made in the earlier papers that the $3p\ ^2F_2$ state is the *lower* state of the transition, since at the time it was assumed that the ground state of the molecule, $3s\ ^2A_1$, is the lower state of the Schuster band. It was Watson who first attempted an analysis on the basis of the reverse assumption, namely that $3p\ ^2F_2$ is the upper and $3s\ ^2A_1$ is the lower state, an assumption that was strongly suggested by the neutralized ion beam experiments of Porter and his associates (Gellene, Cleary & Porter 1982) and confirmed by the absorption experiments of Whittaker *et al.* (1984) and Huber & Alberti (1984). On the basis of this assumption a fine fit of a simulated band with the observed spectrum was obtained yielding provisional constants for upper and lower state as described in more detail in Watson's paper. A few of Watson's assignments are included in Table 1 for the purpose of orientation. Attempts to do the corresponding analysis for the NH_4 Schüler band have

so far failed. There is a strange gap between the branches forming the two heads and the rest of the band suggesting that in NH_4 there is a predissociation in the upper state for low N values.

In Fig. 3 a comparison is shown between the Schüler bands of $^{14}ND_4$ and $^{15}ND_4$ obtained with the 3-m grating-spectrograph. In these spectra the band heads are strongly over-exposed so that the tails of the bands are more developed. For the main part of the band there is a small but constant shift of 1.5 cm^{-1}. Except for this shift the

Table 1. The wavenumbers of lines in the main Schüler band of ND_4.

ν_{vac}	I^a	Assignments[b]	ν_{vac}	I^a	Assignments[b]
14812.039	7 h	P_1^c	14829.291	2	$R_2(3)F_2$
812.256	3		829.966	8	
812.625	10 h	P_1	830.502	3 sh	
813.270	12 b		830.779	5 as	
813.750	8		831.545	3.5 s	$R_1(3)F_2$
814.138	7	$P_1(6)F_2^1$	832.257	10 as	
814.314	6	$P_1(6)F_1$	833.027	5 s	$Q_2(7)F_2^1$, $R_1(4)F_1$
814.681	6	$P_1(5)F_2$	834.218	5 b	$R_2(6)F_2^2$
815.219	6	$P_1(4)F_1$	834.654	9 d	
815.790	3	$P_1(3)F_2$	835.026	3 sh	
816.322	1	$P_1(2)F_2$	835.543	4 s	$Q_2(8)F_1^1$
816.722	0 as		836.311	8 b	
816.962	0 as	$P_1(1)F_1$	837.218	7 s	$Q_1(11)F_2^2$
817.579	0.5 b	$P_2(14)F_2^1$	838.021	3 sh	
818.293	6 h	P_2^d	838.330	7 b	
818.428	6		838.624	3 sh	
818.736	4	$P_2(10)F_2^1$	839.572	3	
819.031	10 h	P_2^e	839.891	8 b	
819.438	11 ov		840.160	5 b	
819.767	12 ov		840.819	6 b	
820.046	11		841.232	2	
820.383	9	$P_2(7)F_2^1$	841.683	5 s	
820.718	3		842.091	6	
821.049	7	$P_2(6)F_2$	842.277	6	
821.232	7		842.648	5 b	
821.568	3		843.471	5	
821.700	3	$P_2(5)F_2$	843.668	7	
821.913	3		844.159	6	
822.538	6 s	$Q_1(4)F_1$, $P_2(4)F_1$	844.341	6	
823.264	1.5		844.578	3 sh	
823.358	1·5		845.078	2	
823.818	0.5 ov		845.288	3	
824.103	5 s	$Q_1(5)F_2$, $P_2(2)F_2$	845.682	4	
824.726	0.5 b		846.199	4	
825.853	6 s		846.344	4	
826.621	1 s	$Q_2(3)F_2$	846.611	4	
827.017	0.5 as	$R_2(1)F_1$	846.805	4	
827.769	7 s		846.972	3	
828.028	1 sh		847.195	3	
828.810	2		847.354	3	
829.077	2	$Q_2(5)F_2$	847.838	5 s	

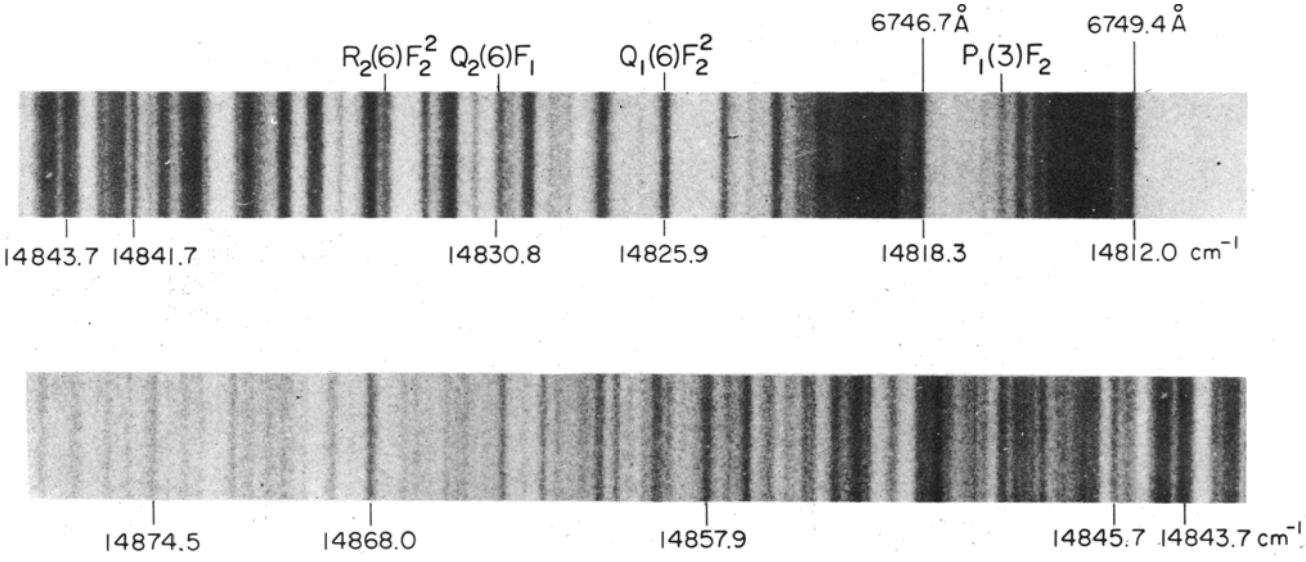


Figure 1. Fine structure of the main Schüler band of ND_4 obtained with a 10-m grating-spectrograph using a Tesla discharge in ND_3 at 100 torr pressure. Watson's assignments for a few single lines are given.

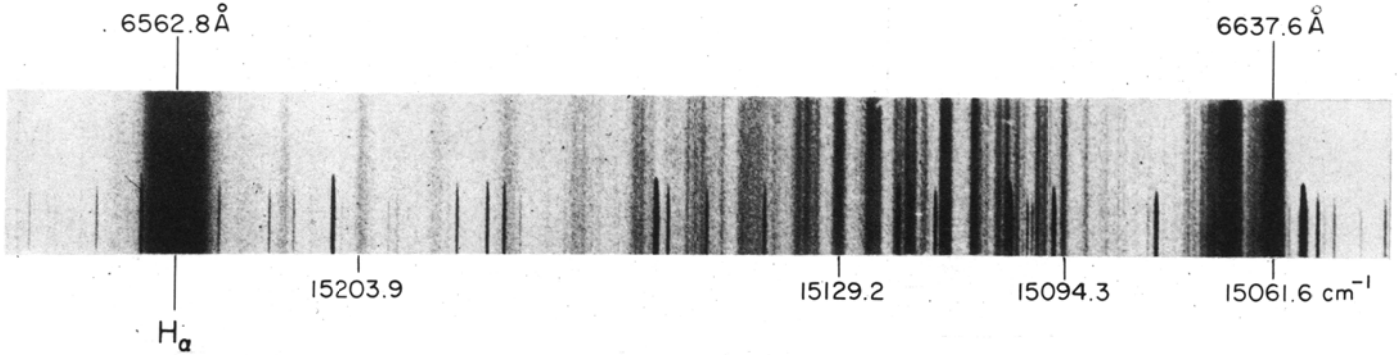


Figure 2. Fine structure of the main Schüler band of NH_4, obtained with a 3-m grating-spectrograph using an ozonizer discharge in NH_3 at 100 torr pressure.

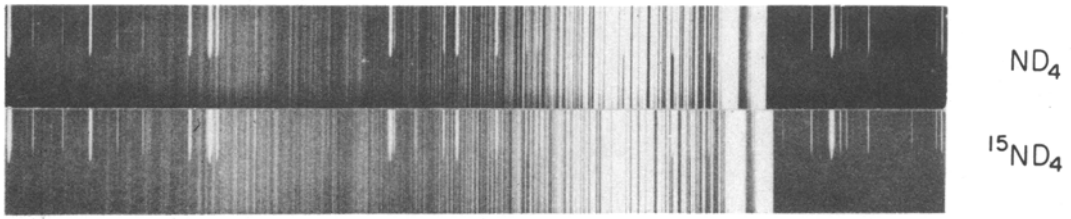


Figure 3. Comparison of the main Schüler bands of $^{14}ND_4$ and $^{15}ND_4$ photographed with a 3-m grating-spectrograph using an ozonizer discharge in $^{14}ND_3$ and $^{15}ND_3$ at 100 torr pressure.

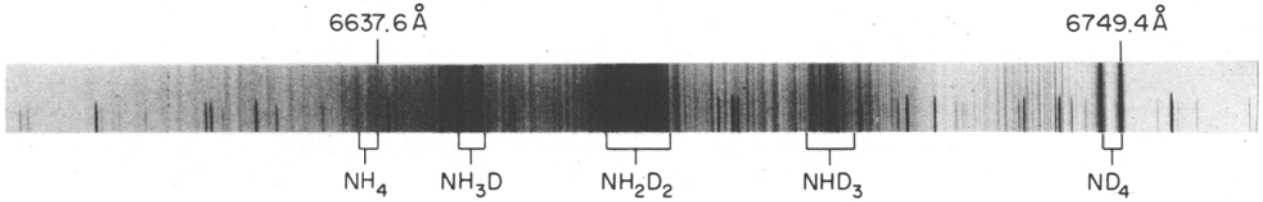


Figure 4. Spectrum of a 60 : 40 mixture of NH_3 and ND_3 in an ozonizer discharge at 100 torr showing the main Schüler bands of the intermediate isotopes NH_3D, NH_2D_2, NHD_3 in addition to NH_4 and ND_4 (3-m grating-spectrograph).

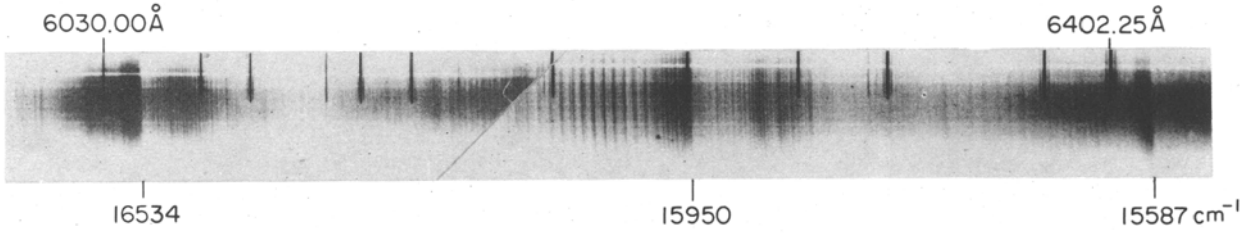


Figure 5. The weaker bands of the Schüler system of ND_4 shortward of the main band obtained with a Zeiss grating-spectrograph using electron beam excitation at 400 torr pressure. Reciprocal dispersion : 1.82 Å mm^{-1}.

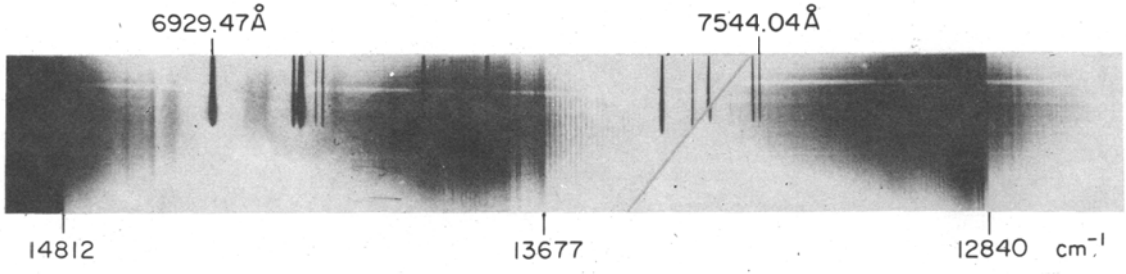


Figure 6. The weaker bands of the Schüler system of ND_4 longward of the main band obtained with a Zeiss grating-spectrograph using electron-beam excitation at 400 torr pressure. Reciprocal dispersion : 5.39 Å mm^{-1}.

ν_{vac}	I^a	ν_{vac}	I^a	ν_{vac}	I^a
14848.174	3	14859.445	5 as	14872.197	3 as
848.391	3.5	859.803	2 s	873.162	1 as
848.608	4	860.249	2	873.628	1 s
849.028	5	860.694	4 s	874.115	0.5 as
849.593	3	861.207	5 s	874.538	3 as
849.892	4 s	861.617	2.5	874.662	1.5 sh
850.102	2	861.752	2.5	875.294	2 s
850.281	2	861.886	1 sh	875.918	2 as
850.559	1.5 sh	862.118	1 b	876.223	1 s
850.720	2.5 b	862.432	2 as	876.906	2 b
851.017	8 s	862.883	4 s	877.884	2 b
851.390	4	863.419	1.5	878.822	1.5
851.559	3 sh	864.087	4 s	879.477	2 s
852.386	3 as	864.567	1 b	880.353	4 s
853.089	3	864.888	0.5 s	881.442	2 b
853.405	5 as	865.226	2 s	882.312	1 b
853.634	3 sh	865.447	1.5 s	882.927	1 b
854.070	4 as	865.761	2 s	883.703	1 b
854.790	3	866.287	1.5 s	884.054	1 as
854.971	3	866.670	2 b	884.845	1 s
855.342	1.5 sh	867.024	2 s	885.589	0.5
855.518	2 as	867.847	0.5 sh	886.342	3 s
855.619	1 sh	867.945	1 sh	887.309	2 s
856.077	2.5 b	868.051	5 as	888.405	1.5
856.317	1	868.592	0	889.176	2 s
856.755	6 as	869.169	3 as	889.998	1.5 b
856.928	2 sh	869.629	0	890.465	0.5 b
857.291	2.5	870.017	1 b	891.212	0 b
857.729	1 sh	870.430	1 sh	892.077	0
857.937	5 as	870.645	2 as	892.738	1 as
858.277	1.5	871.089	2 as	895.873	1 as
858.641	1 s	871.634	1	896.577	0 b
859.096	3 s	871.812	1	897.188	0.5 b

[a] Abbreviations—h: band head; s: sharp; as: asymmetric; b: broad; u: unresolved; ov: overlapping lines; sh: shoulder; d: diffuse.

[b] Only a few assignments are given; for the detailed assignments see Watson (1984). Lines shortward of 14839 cm^{-1} have not yet been assigned.

[c] Band head formed by longward tetrahedral components of P_1 lines.

[d] Band head formed by longward components of P_2 lines.

[e] Band head formed by central components of tetrahydral clusters.

lines in the two bands are virtually identical. However, above 14890 cm^{-1} differences appear. It seems that here a hot band of the type (1–1) is overlapping the tail of the main band and that because of the difference of the α's in upper and lower state larger and more variable isotope shifts arise than in the (0–0) band.

In Fig. 4 the Schüler spectrum of a 60:40 mixture of NH_3 and ND_3 is shown which confirms the statement in Paper 1 that there are three intermediate isotopes as required if the molecule has four equivalent H atoms. A corresponding spectrum for the Schuster bands was shown in Paper 2. While it seems certain that the Schuster system also belongs to NH_4 its interpretation in terms of the electronic states of NH_4 is as yet mysterious. Raynor & Herschbach (1982), Gellene, Cleary & Porter (1982) and Watson

Table 2. Wavenumbers of lines in the main Schüler band of NH_4.

ν_{vac}	I^a	ν_{vac}	I^a	ν_{vac}	I^a
15061.61	10 h	15100.68	1.5	15141.88	2
62.63	10	101.58	4	143.40	2.5
64.76	3	103.20	8	144.24	3
65.62	2	104.44	6 as	147.33	4 s
68.45	10 b, h	106.37	2	150.45	3 u
69.54	8	107.37	2	151.57	3 b
71.74	5	108.23	10	152.87	4 s
72.74	3	112.21	10	154.93	0.5
74.10	1	112.86	10	156.27	0.5
74.98	2 s[b]	115.76	2	158.50	2
75.50	1.5	116.28	2	160.06	5
77.08	0	117.54	8	160.91	5 u
78.80	0	118.56	8	162.31	0.5
80.74	0.5	119.71	2.5	164.08	1
81.14	0.5	120.39	3	166.57	1
82.83	0	123.11	8	167.76	2
83.34	0	124.33	9	169.01	2 u
84.94	1	125.68	1.5	170.52	1
85.78	2 s	129.17	10	171.75	0
87.64	2	131.03	0	174.85	0.5 b
88.39	0	132.68	3	180.71	3 u
90.76	4	134.20	5	190.32	1.5
94.26	8	135.47	4	191.91	2 b
96.42	1.5	136.95	0	202.14	2 b
97.35	4 as	138.42	1.5	203.86	3 b
98.41	5	139.24	2	215.83	2
				216.48	2

[a] See Table 1 for abbreviations.
[b] Possibly an NH_2 line.

(1984) have suggested that the Schuster band belongs to the Schüler system. While it is certainly possible that the lower state is a vibrationally excited ground state level such that it is much closer to the top of the potential barrier leading to strong predissociation, I find it difficult to assume that the upper state belongs to $3p\ ^2F_2$ because of its anomalous ζ value (see Paper 2) and the difference in pressure dependence. It seems more natural to assume that the upper state is a $3d$ state as originally suggested by Watson and confirmed in Paper 2. Even with this assumption it is hard to understand that there is no sign of the transition to the vibrationless ground state.

4. The weaker Schüler bands at medium resolution

Schüler, Michel & Grün (1955) have recorded under very low resolution additional bands besides the principal bands in both the Schüler and Schuster systems. From their stick diagrams one might be led to believe that these additional bands are comparable in intensity to the principal bands. On our plates these additional bands in the Schüler system have intensities less than 1/50 of the principal bands. Figs 5 and 6 show

Table 3. Wavenumbers of the band heads shortward and longward of the main band of ND_4.

ν_{vac}	I	Description
18462	2	central peak of a group of irregular diffuse lines (may belong to Schuster system)
17228.7	20	Q head of main Schuster band—see Paper 2 for details
16534	10	Q head shaded to violet; partially resolved structure on both red and violet sides
15950	20	Q head shaded to violet; partially resolved structure on both red and violet sides
15587	15	Q head shaded to violet; poorly resolved P and R branches
14930	20	head shaded to violet; (1–1) band overlapping tail of main band
14818	200	P heads; main Schüler band, see Section 3.
14812		
14797	10	head shaded to violet; 'hot' band
14758	5	head shaded to violet; 'hot' band
14731	2	head shaded to violet; 'hot' band
14646	1	line-like head, shaded to violet; 'hot' band
14582	1	line-like head, shaded to violet; 'hot' band
14522	1	broad head shaded to violet; 'hot' band
13749	10	head shaded to violet
13736	4	double head, line-like
13729	4	
13686	5	double head, accompanied by simple branch at longer λ
13677	5	
12855	20	double head, line-like, accompanied by partially resolved structure at shorter λ
12848	20	
12840	10	head shaded to violet
12814	2	head shaded to violet; 'hot' band
12798	1	line-like head; 'hot' band
12785	1	broad line-like head; 'hot' band

spectrograms of the ND_4 bands shortward and longward of the main band respectively as obtained with a small Zeiss grating-spectrograph. In Table 3 the observed band heads are listed.

The bands at the shortward side of the main band which were not observed by Schüler *et al.* must obviously correspond to transitions from vibrationally excited levels of the upper state. Except for the band heads very close to the main band the lower state is presumably the same as for the main band *i.e.* 0, 0, 0, 0. Three levels at 775, 1138 and 1722 cm^{-1} above T_0 would thus seem to be established. It is not trivial to establish which vibrations these numbers correspond to. Since certainly the upper $3p\ ^2F_2$ state is a Rydberg state, the vibrational frequencies would be similar to those of ND_4^+ except for the changes caused by vibronic interactions which might indeed be large. Only the vibrational frequencies of NH_4^+ have so far been established in the liquid (see Herzberg 1945). From these one would predict for ND_4^+, $\nu_1 = 2146$, $\nu_2 = 1192$, $\nu_3 = 2330$, $\nu_4 = 1039$ cm^{-1}. None of these frequencies gives good agreement with the intervals observed for the Schüler system. One must, however, consider that, since the upper electronic state is $^2F_2$, for all but ν_1 strong vibronic interactions might lead to large splittings. If the difference 1722 were interpreted as $2\nu_4$ there would be seven resulting

vibronic states of which three are of species F_2, that is, can combine with the ground state.

On the other hand it must be remembered that in an allowed electronic transition only totally symmetric vibrations (here v_1) are Franck-Condon allowed while non-totally-symmetric vibrations occur only with $\Delta v_a = 0, \pm 2, \pm 4$ of which only the first one ($\Delta v_a = 0$) is strong. Only through vibronic interactions, and if the vibronic symmetry is F_2, can $\Delta v_a = \pm 1$ transitions occur. All this is in accordance with the low intensity of the observed bands compared to the (0–0) band.

Vibronic interactions play a much smaller role for the electronic ground state $3s\ ^2A_1$. Therefore the weak bands on the longward side of the main band are easier to understand than those on the shortward side: the band at 12850 probably corresponds to v_1'' and gives 1960 cm^{-1} for this vibration while the band at 13680 corresponds to v_2'' (= 1134 cm^{-1}). The corresponding bands of NH_4 are at 12515 and 13486 cm^{-1} (from Schüler, Michel & Grün 1955) yielding v_1''(H) = 2552 and v_2'' (H) = 1581 cm^{-1}. For both ND_4 and NH_4 the breathing vibration (v_1) comes out substantially smaller than for the ion in agreement with expectation for a fairly weakly bound ground state.

In addition to the (1–1) band of ND_4 mentioned earlier there area number of band-heads longward of the main band (see Table 3).They are observable only when the main band is very strongly overexposed. These band heads clearly belong to sequence bands. Their assignment is impossible at present as long as the fundamentals in upper and lower state are not definitively identified.

All the ND_4 bands discussed here have interesting fine structures which however cannot be analyzed with the present resolution. It is interesting to note that there are fairly extensive branches of nearly equidistant lines in the bands at 16534, 15950 and 13750 with spacings of 3.8, 9.9 and 10.6 cm^{-1} In view of the uncertainty of the values of the electronic and vibrational ζ's it is not possible at the present stage to draw conclusions about the B values but the order of magnitude is correct. Further analysis must await the availability of spectra of substantially higher resolution.

References

Alberti, F., Huber, K. P., Watson, J. K. G. 1984, *J. molec. Spectrosc.,* (submitted).

Droege, A. T., Engelking, P. C. 1983, *Chem. Phys. Lett.,* **96,** 316.

D'Silva, A. P., Rice, G. W., Fassel, G. W. 1980, *Appl. Spectrosc.,* **34,** 578.

Gellene, G. I., Cleary, D. A., Porter, R. F. 1982, *J. chem. Phys.,* **77,** 3471.

Herzberg, G. 1945, *Molecular Spectra and Molecular Structure II. Infrared and Raman Spectra of Polyatomic Molecules,* D. Van Nostrand, New York.

Herzberg, G. 1981, *Faraday Disc. chem. Soc,* No. 71, 165 (Paper 1).

Herzberg, G., Hougen, J. T. 1983, *J. molec. Spectrosc.,* **97,** 430 (Paper 2).

Huber, K. P., Sears, T. 1984, in preparation.

McKeever, M. R., Sur, A., Hui, A. K., Tellinghuisen, J. 1979, *Rev. scient. Instrum.,* **50,** 1136.

Raynor, S. Herschbach, D. R. 1982, *J. phys. Chem.,* **86,** 3592.

Schüler, H., Michel, A., Grün, A. E. 1955, *Z. Naturforsch.* **10a,** 1.

Watson, J. K. G. 1984, *J. molec Spectrosc,* (submitted).

Whittaker, E. D., Sullivan, B. J., Björklund, G. C., Wendt, H. R., Hunziker, H. E. 1984, *J. chem. Phys.,* **80,** 961.

Wulf, O. R., Melvin, E. H. 1939, *Phys. Rev.,* **55,** 687.

J. Astrophys. Astr. (1984) **5,** 139–148

The Ooty Synthesis Radio Telescope: First Results

G. Swarup *Radio Astronomy Centre, Tata Institute of Fundamental Research, P.O. Box 8, Ootacamund 643001*

(*Invited article*)

Abstract. A 4-km synthesis radio telescope has recently been commissioned at Ootacamund, India for operation at 327 MHz. It consists of the Ooty Radio Telescope (530 m × 30 m) and 7 small antennas which are distributed over an area of about 4 km × 2 km. It has a coverage of about ± 40° in declination δ. The beam-width is about 40 arcsec × 90 arcsec at $\delta = 0°$ and about 40 arcsec × 50 arcsec at $\delta = 40°$. The sensitivity attained for a 5:1 signal-to-noise ratio is about 15 m Jy after a 10-hour integration.

The observational programmes undertaken and some of the results obtained recently are summarized. The radio halo around the edge-on spiral NGC 4631 is found to have a larger scale-height at 327 MHz than is known at higher frequencies. Mapping of interesting radio galaxies at 327 MHz is being carried out; preliminary results for 0511–305 (~2 Mpc) and 1333–337 (~ 750 kpc) are summarized. The very-steep-spectrum radio source in the Abell cluster A85 is found to be resolved; since it has no obvious optical counterpart, it is conceivable that it is a remnant of past activity of a galaxy that has drifted away in about 10^9 years.

Key words: radio telescopes—galaxies, radio—galaxies, spiral

1. Introduction

Many galactic and extragalactic radio sources have extended features of low surface brightness which arise from the oldest relativistic electrons in the universe. Studies of such diffuse features offer a powerful means of understanding the evolution of radio sources, particularly concerning the sites of particle acceleration and energy losses. As these components generally have steep spectra, they become increasingly more prominent at longer wavelengths. Hence, their detailed study requires a radio telescope operating at wavelengths of a metre or more and having sufficiently high resolving power, sensitivity and dynamic range. Such a telescope is also useful for many other types of studies, such as, undertaking radio source surveys, monitoring metre-wavelength variable sources, studying flare stars, observing radio recombination lines *etc.*

In this paper, we briefly describe a 4-km aperture synthesis radio telescope recently set up at Ootacamund, India, for operation at 326.5 MHz. Several observational programmes that have been initiated and some of the results obtained are also described.

2. The Ooty Synthesis Radio Telescope

As shown in Fig. 1, the Ooty Synthesis Radio Telescope (OSRT) consists of 8 antennas located in an area of about 4 km × 2 km (Bagri *et al.* 1984). This configuration was chosen to take advantage of the large collecting area and the long north-south aperture of the Ooty Radio Telescope (ORT) which is described below. As the ORT is 530 m long, it was necessary to install only a small number of remote antennas to achieve a good coverage in the spatial-frequency (*u, v*) plane of the synthesis telescope. For a 10-hour observing run with the 4-km OSRT, the maximum level of positive or negative sidelobes of the synthesized beam is below 15 per cent (depending upon the declination), thus permitting a reasonable CLEANing of the dirty maps (Högbom 1974). Though the remote antennas are of much smaller size than the ORT, the sensitivity attained with the OSRT in 10 hours is ~ 15 mJy (for a 5:1 signal-to-noise ratio). The beamwidth of the OSRT is about 40 arcsec × 90 arcsec at declination $\delta = 0°$ and about 40 arcsec × 50 arcsecat $\delta = 40°$. The system parameters are summarized in Table 1.

The ORT consists of a 530 m long and 30 m wide steerable parabolic cylindrical antenna mounted equatorially (Swarup *et al.* 1971; Sarma *et al.* 1975). Its effective collecting area is about 8000 m^2. The antenna can be steered mechanically in hour angle (HA) from $-\ 04^h07^m$ to $+\ 05^h26^m$. An array of 1056 co-linear dipoles is placed along the focal line of the cylindrical antenna, and is grouped into 22 modules containing 48 dipoles each. The antenna beam is steered in declination by appropriately phasing the dipole array. A 4-bit diode-controlled phase-shifter with a loss of about 0.7 dB is placed after each dipole. A branched transmission line system with a loss of 0.3 dB connects the 48 dipoles. An RF amplifier of noise temperature around 120 K follows each of the modules of 48 dipoles. The system temperature of the ORT is about 250 K. The outputs of the 22 modules of the ORT are brought to a central receiver room at an IF of 30 MHz and combined there.

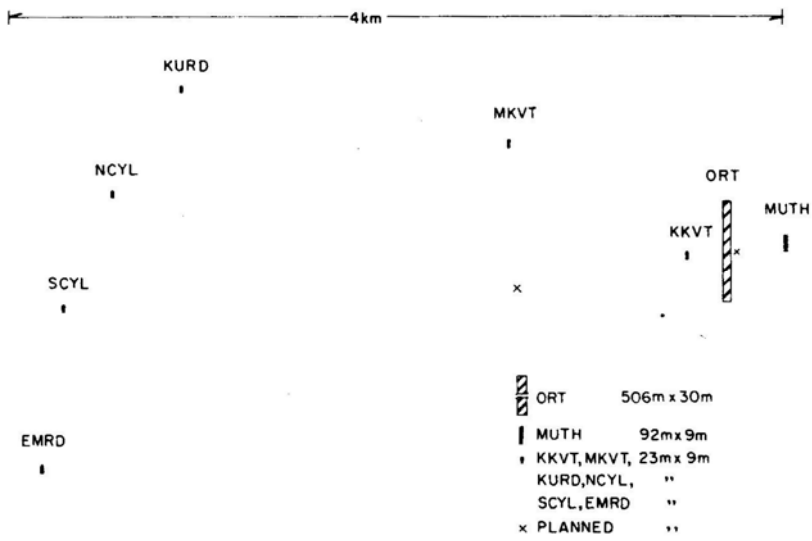


Figure 1. Configuration of the 4-km Ooty Synthesis Radio Telescope (OSRT).

Table 1. Parameters of the Ooty Synthesis Radio Telescope.

Antennas (Illuminated aperture)	ORT (506 m × 30 m) One parabolic cylinder (92 m × 9 m) Six parabolic cylinders (23 m × 9 m)
Baseline	4 km (EW), 2 km (NS)
Frequency	326.5 MHz
Bandwidth	3.5 MHz
Sky coverage HA	-04^h07^m to 05^h26^m (ORT) -05^h30^m to 05^h30^m (other antennas)
DEC	$\pm 40°$
Field of view (FWHP)	3° × 0°.6 (EW × NS); 3° × 3° with declination scanning
Synthesized beamwidth ($\beta_\alpha, \beta_\delta$)	40 arcsec × 90 arcsec at $\delta = 0°$ 40 arcsec × 50 arcsec at $\delta = 40°$
Sensitivity	15 mJy (S/N ratio = 5:1) for 10-hour integration

The antenna located about 360 m to the east of ORT (Fig. 1) is a 92 m long × 9 m wide equatorially-mounted parabolic cylinder, containing 4 modules of 48 dipoles each. All the other 6 antennas, which are located at distances ranging from 200 m to about 4 km west, are 23 m × 9 m parabolic cylinders. These are steered mechanically from about -5^h30^m to $+5^h30^m$ (some of them from -6^h to $+6^h$). The dipole array and RF electronics for these antennas are similar to those of the ORT. The pointing of the OSRT antennas in hour angle and declination is done using a telemetry system (Sankararaman, Subramanian & Balasubramaniam 1982).

Local oscillator signals are transmitted to antennas located up to about 300 m distance by cables buried underground. For more distant antennas, the local oscillators are phase-locked to a central oscillator and any path variations are measured using a round-trip phase measurement scheme (Bagri, Narayana & Venkatasubramani 1984). The 30-MHz IF signals from the remote antennas are brought to the central receiver room, converted to video signals, and correlated with the output of the ORT. The observations described in this paper have been made over the last few months using only a 5-baseline digital correlator system, which correlates the output of the entire ORT with those of 5 other antennas at a time.

The field of view of the OSRT is determined by the voltage radiation pattern of the entire ORT, which is 170 arcmin (EW) × 8 arcmin (NS). However, for mapping a large field of view (up to 170 arcmin × 170 arcmin) with the existing 5-baseline correlator system, a declination scanning technique has been developed by Pramesh Rao & Velusamy (1983). Using diode phase shifters, the declination of the ORT is changed in steps of about 3 arcmin every 200 ms. For a 170 arcmin field of view, the declination scanning is completed in 12 s, during which period there is negligible change of the u and v Fourier components of the source brightness distribution being observed. The Fourier transform of the visibilities measured during every scan gives the true visibility function of the source at Δv intervals, equal to the width of the ORT divided by the number of declination settings. This scheme also allows calibration of the OSRT using any relatively compact source located in the field for determining instrumental and ionospheric phase variations.

For the purpose of measuring low spatial frequency components, it is planned to install a 28 m × 9 m cylinder about 50 m to the east of ORT. Another antenna is to be installed a kilometre away for improving the (u, v) coverage. A 256-channel correlator

system is presently being installed for intercorrelating 14 antenna-outputs, *e.g.* 5 outputs of the ORT (each from 4 of the 22 modules) and 9 of the small antennas. The division of the ORT would widen the NS field of view of the OSRT to about 0°.6. These additions will improve the capability of OSRT considerably.

3. Scientific programmes and results

A summary of the observational programmes being undertaken currently with OSRT and some of the results obtained so far are given below.

3.1 *Galactic Radio Sources*

The wide-field mapping technique described above has been used for mapping many galactic sources, including the galactic centre. Velusamy & Pramesh Rao (1982) have mapped W 50 at 327 MHz and have examined the radio emission from the distant lobe-like features separated by ~ 2° along the jet axis of SS 433. Map of IC 443 at 327 MHz was presented by Pramesh Rao & Velusamy (1983). Recently, P. D. Jackson & T. Velusamy have observed a 6° × 6° region to detect any continuum emission at 327 MHz associated with an expanding HI shell in Puppis (Stacy & Jackson 1982). However, the OSRT maps of several of these galactic sources suffer from the missing short spacings ($< 80\ \lambda$) and can be completed only after the nearby antenna (at 50 m from the ORT) becomes available.

3.2 *Nearby Spiral Galaxies*

S. Sukumar & T. Velusamy are observing systematically several edge-on and face-on nearby spiral galaxies, particularly at low or southern declinations. The aim of this programme is to (i) separate the thermal and nonthermal emission in the disk unambiguously, (ii) study the synchrotron emission as a function of distance (z) from the plane of the galaxy, and (iii) derive detailed models for the origin and propagation of relativistic electrons in the disk and halo of nearby spiral galaxies. Fig. 2 shows a 327-MHz map of NGC 4631, a well-known edge-on galaxy. The radio map has been restored with a beam of 60 arcsec × 60 arcsec. The optical image of NGC 4631 is marked by a broken line and is shaded by dots. Sukumar & Velusamy (1984) have made a comparison of the 327-MHz map with the Westerbork maps at 610 MHz and 1412 MHz (Ekers & Sancisi 1977), and with the Effelsberg map at 10.7 GHz (Klein & Wielebinski, personal communication), smoothing all the four maps to similar resolutions. Their results show a halo of larger scale height at 327 MHz than seen at higher frequencies. The spectral index α between 327 MHz and 10.7 GHz steepens from about –0.6 ($S \propto \nu^{\alpha}$) over the galactic disk to – 1.5 at about 6 kpc above the disk.

3.3 *Extended Radio Galaxies*

Several strong and nearby radio galaxies such as Fornax A are being mapped for studying the interaction of their outer components with the circumgalactic material. Gopal-Krishna, S. Lakshmi, A. K. Singal & S. Krishnamohan are making a search for giant radio galaxies and are also mapping radio galaxies with size > 500 kpc.

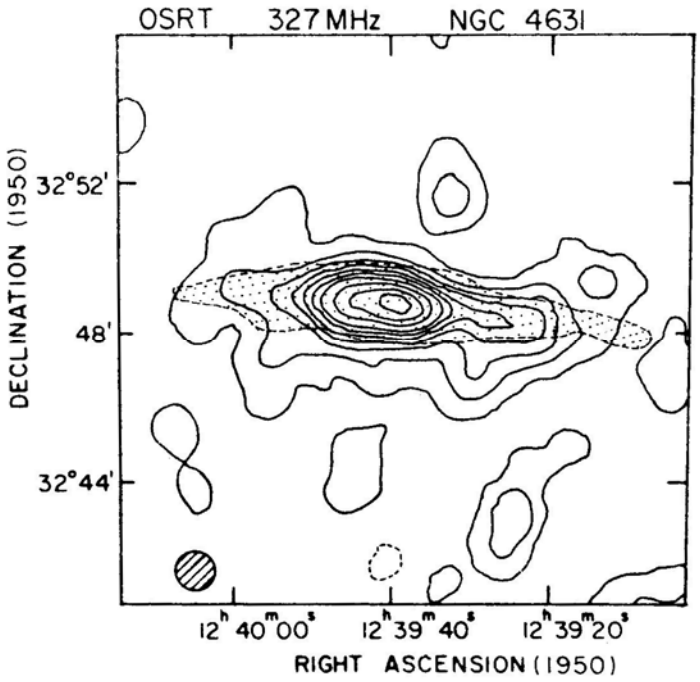


Figure 2. A327-MHz map of NGC 4631 made with a 60-arcsec Gaussian beam. The contour levels are – 15,15,45,75,105,135,165,195,245,295 and 345 mJy/beam. The optical image of the galaxy is marked by a broken line and is shaded by dots.

In Fig. 3 is shown a 327-MHz OSRT map of the radio source 0511 – 305 made with a 60-arcsec beam, revealing a complex structure. The source was mapped earlier at 408 MHz with a 2.8-arcmin beam (Schlizzi & McAdam 1975). The present observations resolve each lobe into a few peaks lined up along a curved ridge. These ridges show an inversion symmetry about the 18.5-mag galaxy 'c' which, therefore, might be the correct optical identification. Also, this galaxy lies almost exactly midway between the outermost radio peaks. For an estimated redshift of 0.25 based on the apparent magnitude, the overall size of the source would be more than 2 Mpc.

An OSRT map of another large radio galaxy 1333–337 is shown in Fig. 4. The parent galaxy IC 4296 is elliptical with $m_r \sim 11$ and a redshift $z = 0.0129$ (Evans 1963). Earlier, this source has been mapped at 1415 MHz with 50-arcsec resolution (Goss *et al.* 1977) and at 408 MHz with a 2.8-arcmin beam (Schlizzi & McAdam 1975). The outermost peaks are seen to be separated by 33 arcmin which corresponds to ~ 750 kpc. The south-eastern lobe has a complex structure. The inner radio double is embedded within the main body of the galaxy, and has a notably sharp boundary to the north-eastern side as seen in the 327 MHz map. This is consistent with the motion of the galaxy against the intergalactic medium, as proposed by Goss *et al.* (1977) in order to account for the bend in the large-scale structure.

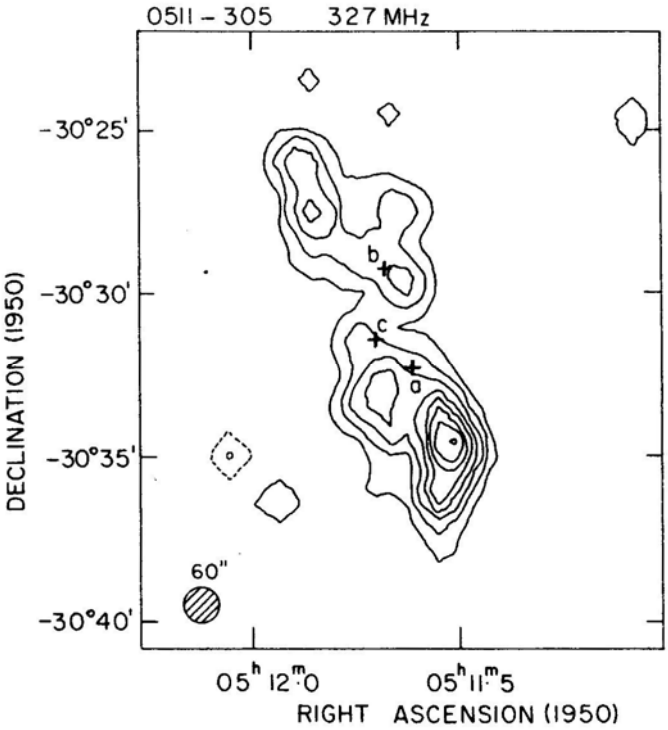


Figure 3. The OSRT map of the source 0511–305, restored with a 60-arcsec beam. The contour levels are – 60, –30, 30, 100, 200, 350, 500, 700, 900 and 1200 mJy/beam. The three plus marks denote the positions of optical objects (a) 17-mag N galaxy; (b) 17.5-mag galaxy and (c) 18.5-mag galaxy (Schlizzi & McAdam 1975).

3.4 *A Very-Steep-Spectrum Source in Abell 85*

It is known that radio sources with a very steep spectral index ($\alpha \lesssim -1.2$) tend to occur in rich clusters of galaxies (Baldwin & Scott 1973), predominantly of Bautz-Morgan (BM) class I which are dominated by a giant cD galaxy (Slee, Wilson & Siegman 1982). Such clusters are also found to be X-ray emitters, implying the presence of hot thermal gas which can confine relativistic electrons for a long period of time and thereby allow synchrotron losses to steepen the radio spectrum. There are only a limited number of known, very-steep-spectrum sources, which have been found either from the association of 4C sources with rich clusters (*e.g.* McHardy 1979), or from Culgoora observations at 80 and 160 MHz of sources known to be associated with clusters from work at higher frequencies (*e.g.* Slee & Siegman 1983). With the high sensitivity of the OSRT, it is proposed to search for steep spectrum sources by surveying a large number of clusters of BM-I type.

In Fig. 5(a) is shown a radio map of the A bell cluster A 85 made by V. K. Kapahi & Ravi Subrahmanyan at 327 MHz with a 3-arcmin gaussian beam. For comparison, a 2695 MHz Effelsberg map made with a 4.4 arcmin beam (Waldthausen *et al.* 1979) is shown in Fig. 5(b). The source 1 is coincident with the brightest elliptical cD galaxy of 12.8 mag. Bright elliptical galaxies are also seen near radio sources 2 and 5. The source 4 is not seen at 327 MHz, and may be a background fiat spectrum source.

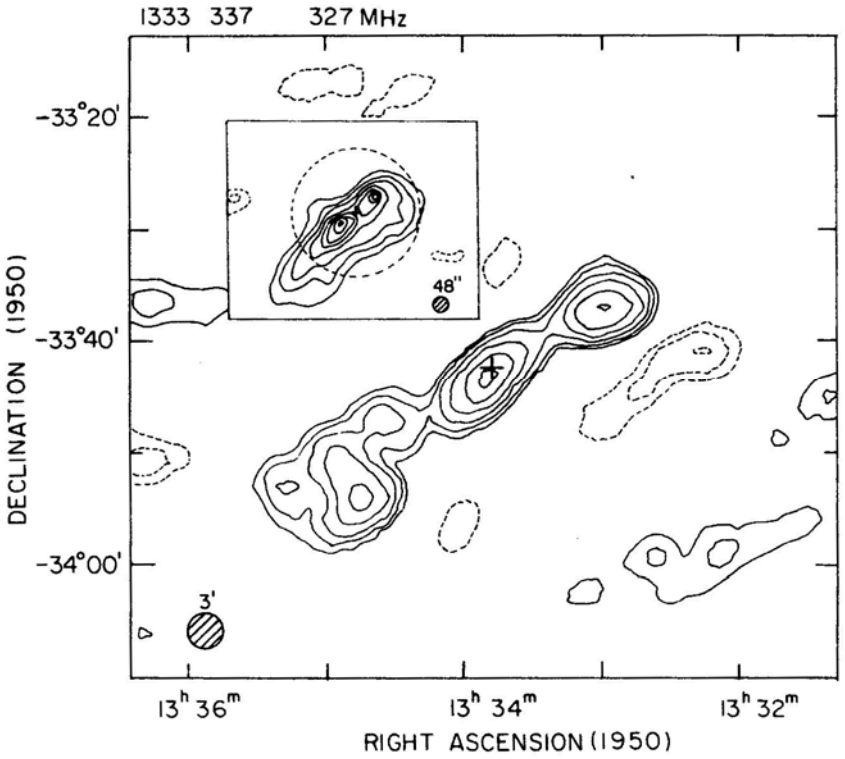


Figure 4. The OSRT map of the source 1333–337 restored with a 3-arcmin beam. The contour levels are –500, –250, –120, 120, 250, 500, 1000, 1800 and 3000 mJy/beam. The plus sign marks the position of the centre of the optical galaxy IC 4296 (Goss *et al.* 1977).The inset is a 327-MHz map of the central component, made with a resolution of 50 arcsec, using mainly the larger antenna spacings of the OSRT. The contours are drawn at –75, –40, 40, 100, 200, 350, 500, 650, 800 and 900 mJy/beam. The outline of the optical galaxy with a diameter 160 arcsec (= 60 kpc) is indicated with a circle encompassing the inner double source.

The 327-MHz map shows a very-steep-spectrum source (VSSS) coincident with the Molonglo position of radio source 0038–096 (Mills & Hoskins 1977). Its spectral index $\alpha = -1.94$ between 80 and 408 MHz, which is consistent with its being absent (<20 mJy) in the Effelsberg map at 2695 MHz. A preliminary 327-MHz map with a 40-arcsec resolution shows that the source is considerably resolved.

Abell 85 is a distance class 4, Richness class 1, BM class I cluster with a redshift of 0.0518 (Sarazin, Rood & Struble 1982). The X-ray emission is smooth and centrally peaked around the galaxy with an X-ray core radius of 3.25 arcmin (Jones *et al.* 1979). Assuming $\alpha = -2$ between 10 MHz and 10 GHz for the VSSS, the total radio luminosity $L = 6 \times 10^{42}$ erg s^{-1} (for $H_0 = 50$ km s^{-1} Mpc^{-1}). This gives an estimated equipartition field $B \sim 3.5 \times 10^{-6}$ gauss and source age $> 8.4 \times 10^{8}$ yr, assuming that the frequency of break in the spectrum $\nu_c < 30$ MHz and the ratio of the energy of protons to that of electrons is unity.

The tenth brightest galaxy in the cluster is of 15.7 mag. Since strong radio sources are generally associated with the brightest galaxies in a cluster, it is curious that there is no obvious optical identification for this very-steep-spectrum source. The source is

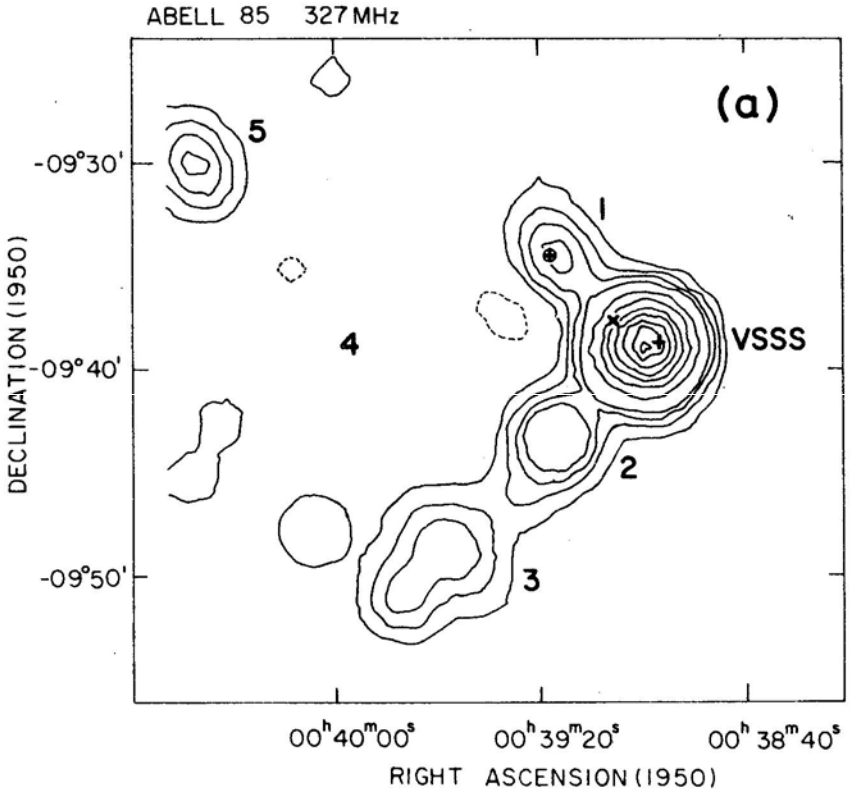


Figure 5. (a) A 327-MHz map of the cluster A bell 85 made with a 3-arcmin beam. The contour levels are: –50, 50, 100, 150, 200, 400, 600, 800, 1000, 1200, 1400 and 1600 mJy/beam. The symbol × shows the estimated centre of the cluster (Waldthaussen *et al.* 1979), ⊕ the position of the cD galaxy and + the Molonglo position of 0038–096. (b) A 2695-MHz map of Abell 85 with a 4.4-arcmin beam made by Waldthaussen *et al.* (1979). Symbols are same as in Fig. 5(a). The contour levels are: 0, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90 mJy/beam.

separated from the cD galaxy by 6 arcmin corresponding to ~ 550 kpc. If the radio source was created by the cD galaxy, the speed of the cD galaxy would have to be about 550 km s^{-1} to travel a distance of 550 kpc in 10^9 yr, the estimated age of the source. This speed is considerably higher than generally estimated for cD galaxies. Kapahi & Subrahmanyan have considered therefore the possibility that the radio source was caused by another bright elliptical galaxy which may have since moved away, or even been cannibalized by the dominant cD galaxy in the potential well of the cluster. Higher resolution observations of this and other similar clusters could throw more light on this question.

4. Conclusions

Observations have been summarized of some of the radio sources mapped with a 4-km synthesis radio telescope which has recently been completed for operation at 327 MHz. The dynamic range of the maps is usually limited to about 20 to 30 due to phase

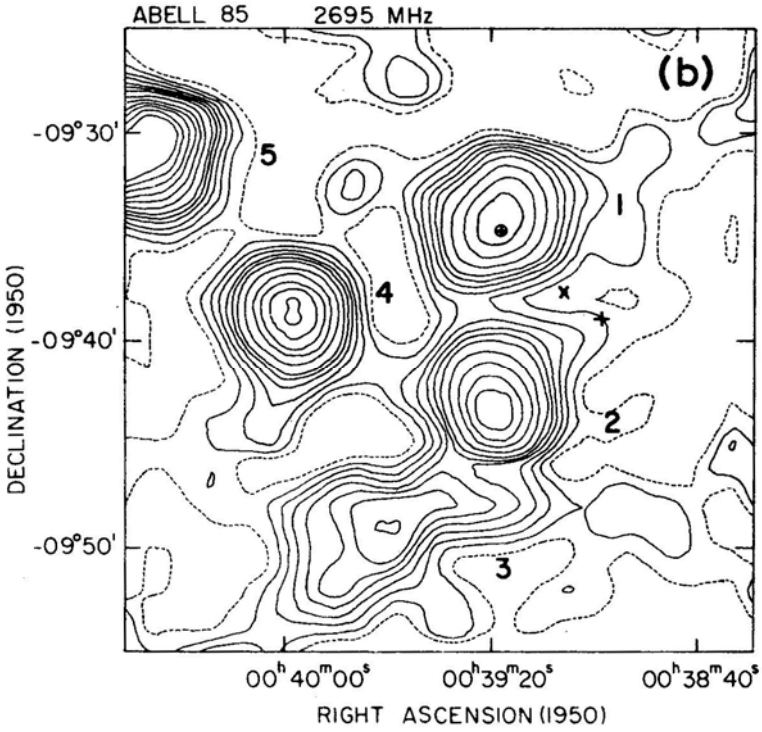


Figure 5. Continued.

variations caused by ionospheric irregularities. Because of the relatively small collecting area of most of the antennas, 'self-calibration' schemes cannot in general be applied except for a few strong sources. To improve the dynamic range, suitable calibration schemes are therefore being evolved utilizing the rapid steering capability of the antennas in declination. With an improved dynamic range, the 40-arcsec beam would be valuable for studying diffuse features in a variety of galactic and extragalactic radio sources.

Acknowledgements

The OSRT is the result of the coordinated and persistent efforts of the entire staff of the TIFR Radio Astronomy Centre. I thank all my colleagues who have kindly provided their results in advance of publication.

References

Bagri, D. S. *et al.* 1984, in preparation.
Bagri, D. S., Narayana, D. L., Venkatasubramani, T. L. 1984, in preparation.
Baldwin, J. E., Scott, P. F. 1973, *Mon. Not. R. astr. Soc.,* **165,** 259.
Ekers, R. D., Sancisi, R. 1977, *Astr. Astrophys.,* **54,** 973.
Evans, D. S. 1963, *Mon. Not. astr. Soc. Sth Africa,* **22,** 140.

Goss, W. M., Wellington, K. J., Christiansen, W. N., Lockhart, I. A., Watkinson, A., Frater, R. H., Little, A. G. 1977, *Mon. Not. R. astr. Soc.,* **178,** 525.
Högbom, J. A. 1974, *Astr. Astrophys. Suppl.,* **15**, 417.
Jones, C., Mandel, E., Schwarz, J., Forman, W., Murray, S. S., Harnden, Jr., F. R. 1979, *Astrophys. J.,* **234,** L21.
McHardy, I. M. 1979, *Mon. Not. R. astr. Soc.,* **188,** 495.
Mills, B. Y., Hoskins, D. G. 1977, *Austr. J. Phys.,* **30,** 509.
Pramesh Rao, A., Velusamy, T. 1983, in *Measurement and Processing for Indirect Imaging,* Ed. J. A. Roberts, Cambridge University Press, p. 000.
Sankararaman, M. R., Subramanian, N., Balasubramaniam, R. 1982, *J. Instn Electronics Telecommun. Engr.,***28,** 216.
Sarazin, C. L., Rood, H. J., Struble, M. F. 1982, *Astr. Astrophys.,* **108,** L7.
Sarma, N. V. G., Joshi, M. N., Bagri, D. S., Ananthakrishnan, S. 1975, *J. Instn Electronics Telecommun. Engr.,***21,** 110.
Schilizzi, R. T., McAdam, W. B. 1975, *Mem. R. astr. Soc.,* **79,** 1.
Slee, O. B., Siegman, B. C. 1983, *Proc. astr. Soc. Austr.,* **5,** 114.
Slee, O. B., Wilson, I. R. G., Siegman, B. C. 1982, *Proc. astr. Soc Austr.* **4,** 431.
Stacy, J. G., Jackson, P. D. 1982, *Nature,* **296,** 42.
Sukumar, S., Velusamy, T. 1984, in preparation.
Swarup, G., Sarma, N. V. G., Joshi, M. N., Kapahi, V. K., Bagri, D. S., Damle, S. H., Ananthakrishnan, S., Balasubramanian, V., Bhave, S. S., Sinha, R. P. 1971, *Nature, Phys. Sci.,* **230,** 185.
Velusamy, T., Rao, A. P. 1982, in *IAU Symp.* 101: *Supernova Remnants and their X-ray Emission,* Eds P. Gorenstein & J. Danziger, D. Reidel, Dordrecht, p. 465.
Waldthausen, H., Haslam, C. G. T., Wielebinski, R., Kronberg, P. P. 1979, *Astr. Astrophys. Suppl.* **36,** 237.

J. Astrophys. Astr. (1984) **5,** 149–158

Brightness, Polarization and Electron Density of the Solar Corona of 1980 February 16

K. R. Sivaraman, M. Jayachandran, K. K. Scaria, G. S. D. Babu,
S. P. Bagare & A. P. Jayarajan *Indian Institute of Astrophysics, Bangalore 560034*

Received 1983 September 1; accepted 1984 January 30

Abstract. During the eclipse of 1980 February 16 we photographed the solar corona at an effective wavelength of 6300 Å. Using a quadruple camera we also obtained the coronal pictures in polarized light for four Polaroid orientations. We have used these observations to derive the coronal brightness and polarization and from these the electron densities in the corona out to a distance of about 2.5 $R_\odot$ from the centre of the disc. The coronal brightness matches well with that of the corona of 1958 October 12.

Key words: solar eclipse—solar corona: brightness, polarization, electron density

1. Introduction

The solar eclipse of 1980 February 16, with its belt of totality within 800 km of Bangalore, made it possible to conduct several experiments for the study of the solar chromosphere and corona. Two camps were set up, one at Hosur about 40 km south of Hubli and a second one at the state central farm at Jawalgere about 50 km east of Raichur.

The long-focus camera for the broad-band photography of the corona, the polarigraph and a coronal spectrograph were located at the camp at Hosur (latitude 15° 00′ 12″ N; longitude $5^h\ 00^m\ 36.3^s$ E) about 4 km south of the central line of totality. The duration of totality at the site was 164 s. The second contact was predicted at $10^h\ 13^m\ 9.0^s$ U.T. The skies were clear many days before and on the eclipse day and the winds were exceptionally low. In this paper we present the results obtained from the polarigraph and the coronal photographs with the long-focus camera.

2. The observations

2.1 *Broad-band Photometry*

We photographed the corona through a broad-band filter with peak transmission around 6300 Å (Wratten 25) using the same arrangement as during the eclipse of 1970 March 7 at Mexico (Bappu, Bhattacharyya & Sivaraman 1973). This consisted essentially of an *f* /48, 6.0-m focal length camera fed by a siderostat of 30 cm aperture

Table 1. Details of the exposures of the broad-band coronal photographs.

Plate No.	Duration of exposure s	Developer
1	Instantaneous	Promicrol
2	2	Promicrol
3	10	Promicrol
4	30	Promicrol
5	50	Promicrol diluted

and a rotating plate holder that accommodated six 8 × 10 inch plates of Kodak IIIa-F emulsion. The shutter of the camera located in front of the objective was operated from the focal plane end by the observer who changed the plate after each exposure in a pre-programmed sequence. The scheme of these exposures are set out in Table 1. Photometric standards were impressed on plates belonging to the same box of IIIa-F emulsion in the laboratory with a Kodak step-wedge with eleven steps, their densities ranging from 0.86 to 3.08.The exposures with the standard wedge of the same duration as for the eclipsed Sun established a relative photometric scale. This in turn was related to the mean brightness of the solar disc through exposures of the disc made on the day after the eclipse around the same time, through a Kodak neutral-density filter of density 4.7 and a 12.7 mm diameter diaphragm over the objective. Each photograph was processed along with its step-wedge calibration and the absolute calibration plates, details of which are also presented in Table 1.

2.2 *The Polarigraph*

The experiment designed to record the polarization of the corona from the limb to about $3R_{\odot}$ was carried out with a quadruple lens camera with identical Zeiss lenses of

Table 2. Details of the exposures with the quadruple lens polarigraph.

Exposure No.	Duration of exposure s	Frame number in each exposure for the four polaroid positions			
		A ↔	*B* ⤡	*C* ↕	*D* ⤢
1	Instantaneous	1	2	3	4
2	1	5	6	7	8
3	2	9	10	11	12
4	4	13	14	15	16
5	6	17	18	19	20
6	12	21	22	23	24
7	20	25	26	27	28
8	Instantaneous	29	30	31	32
9	1	33	34	35	36
10	3	37	38	39	40
11	5	41	42	43	44
12*	7	45	46	47	48

*Diamond Ring appeared.

1 m focal length working at *f*/10 and supported on an equatorial mount. Four polarizers cut from the same sheet of HN 32 polaroid together with four Wratten 25 Kodak gelatine filters were mounted near the focal plane such that the axis of each of the polaroids was inclined at 45° with reference to its neighbour. The four polarized images of the corona were photographed in one shot of exposure on 70-mm Kodak 2485 film. This film in 150-foot roll was mounted in the film transport of the camera so that the entire sequence of exposures planned during the progress of the eclipse could be photographed without any interruption. One member of the team operated the shutter and another the film transport.

In Table 2 we present the details of the exposures made after the announcement of the second contact in the camp. The impression of the standard wedge was established on that part of the film in the same roll which remained unused after the exposures during the eclipse. The entire film roll was developed in a developing tank, a few days after the eclipse.

3. Reductions and results

3.1 *Absolute Surface Brightness*

Among the five broad-band photographs obtained within the high-resolution image, the 10-s and 30-s exposures were chosen for evaluating the surface brightness distribution in the corona. We derived the isophotes by the equidensitometry technique

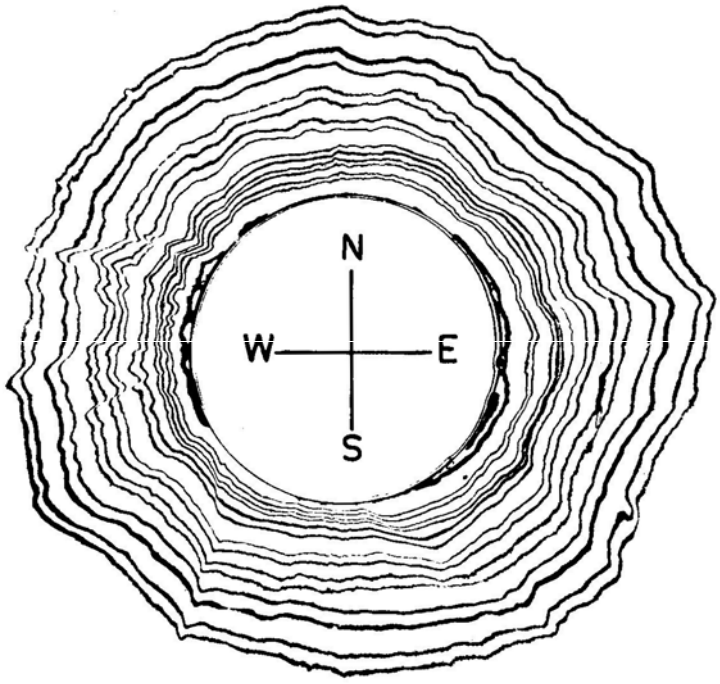


Figure 1. Isophotes of the solar corona. The outermost isophoteis 1 and the numbers increase towards the solar limb (see Table 3).

based on the Sabattier effect, which has been employed with considerable success earlier (Bappu, Bhattacharyya, & Sivaraman 1973), using the high contrast Indu Graphic Arts film. We made the final composite by combining together the family of contours of each plate. The two plates had two of their density contours common between them. By perfectly overlapping these common density contours we ensured the matching of the rest of the isophotes of one plate with those of the other. The composite thus derived and shown in Fig. 1 has isophotes running from 1.01 to 2.53 $R_\odot$ from the centre of the Sun.

To obtain the intensity gradients along the polar and equatorial directions, we made microdensitometer scans along the two diameters approximately at position angles 0°, 90°, 180° and 270° and derived the intensity values on a relative scale using the step-wedge calibration curve. The mean of the four values represented the intensity level of each of the isophotes on an arbitrary photometric scale. We then transformed these intensities from the relative scale to values expressed in terms of the mean brightness of the solar disc $\bar{B}_\odot$ using the absolute calibration, adopting the value $\bar{B}_\odot = 1.98 \times 10^5$ stilb outside the Earth's atmosphere (Allen 1973).

We have estimated the brightness level of the sky contribution from traces made diagonally from one end of the plate to the other across the moon. The sky brightness at about $5R_\odot$ has a value of 1.5×10^{-10} of average brightness of the solar disc. In Table 3 we give the intensities of the isophotes in units of 10^{-8} of the mean brightness of the solar disc after correction for the sky brightness. The brightness of the corona along the N, S, E, W are represented in Fig. 2 and those along the equatorial and polar diameters in Fig. 3.Waldmeier's values of the brightness of the corona of 1958 October 12 along the same directions, extracted from the compilations of Hata & Saito (1966) are also shown in Fig. 3 for comparison. It is interesting to note the close agreement between our present values and those of Waldmeier in view of the fact that both these eclipses

Table 3. Intensities in the equatorial direction (mean of east and west) of the corona of 1980 February 16.

Isophote No.	$r/R_\odot$*	$B/\bar{B}_\odot$†
1	2.244	6.59
2	2.148	7.45
3	1.973	9.86
4	1.882	11.84
5	1.747	16.57
6	1.690	19.99
7	1.591	30.26
8	1.536	41.32
9	1.433	79.83
10	1.380	110.49
11	1.332	151.55
12	1.294	184.02
13	1.223	264.57
14	1.188	329.41

* Equatorial radius in the units of solar radius.

† Intensities in the units of 10^{-8} times the mean brightness of the solar disc.

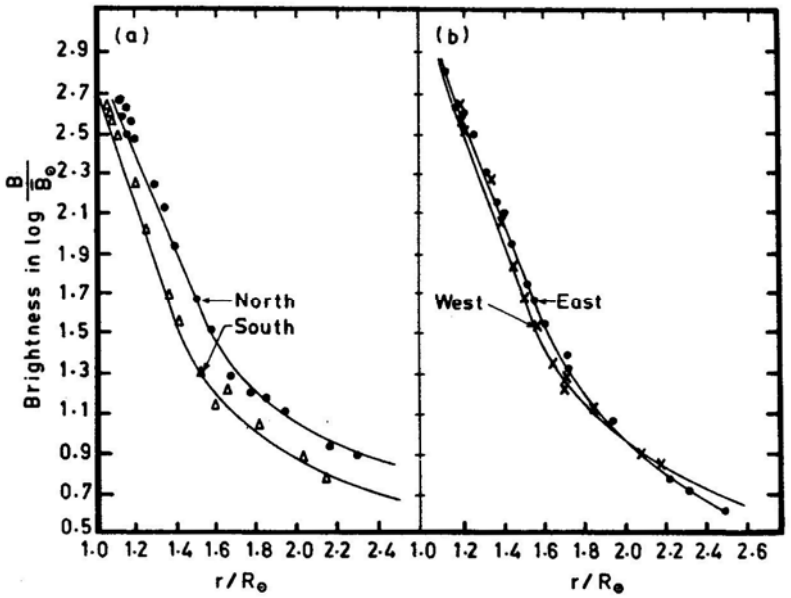


Figure 2. Brightness of the corona (K + F) along the north (filled circles), south (open triangles), east (filled circles) and west (crosses) directions, expressed in units of 10^{-8} of the mean brightness of the solar disc ($\bar{B}_{\odot}$), corrected for sky brightness. The curves represent the 19th-degree polynomial fit.

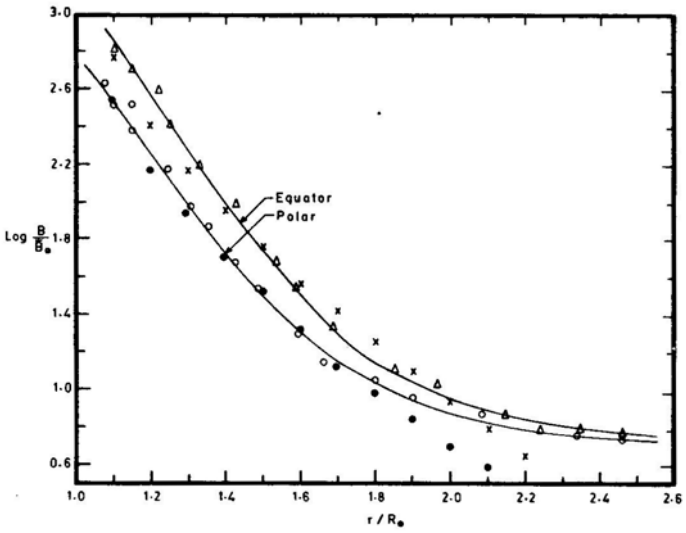


Figure 3. Brightness of the corona (K + F) along the equatorial and polar diameters. The curves represent the 19th-degree polynomial fit to the observed points (open triangles and open circles). Waldmeier's results (crosses and filled circles) for the corona of 1958 October 12 eclipse are plotted for comparison.

occurred at epochs of two of the most prominent maxima of solar activity. The radial brightness distribution along the equatorial and polar directions can be represented by a polynomial of the form

$$B = \sum C_n r^{-n}. \tag{1}$$

By numerous trials using various degrees of polynomials starting from the 30th degree and employing the method of least squares we determined the polynomials representing the best fit with our observations. We find that only the coefficients listed below are significant and by the addition of the intermediate terms the quality of the fit does not improve.

Equatorial:

$$\frac{B_{eq}}{\bar{B}_\odot} = -\frac{0.21\times10^7}{r^{19}} + \frac{0.83\times10^7}{r^{17}} - \frac{0.14\times10^8}{r^{15}} + \frac{0.13\times10^8}{r^{13}} - \frac{0.78\times10^7}{r^{11}} + \frac{0.28\times10^7}{r^9} - \frac{0.62\times10^6}{r^7} + \frac{0.75\times10^5}{r^5} - \frac{0.37\times10^4}{r^3} \tag{2}$$

Polar:

$$\frac{B_{polar}}{\bar{B}_\odot} = \frac{0.10\times10^7}{r^{19}} - \frac{0.45\times10^7}{r^{17}} + \frac{0.83\times10^7}{r^{15}} - \frac{0.84\times10^7}{r^{13}} + \frac{0.52\times10^7}{r^{11}} - \frac{0.19\times10^7}{r^9} + \frac{0.44\times10^6}{r^7} - \frac{0.54\times10^5}{r^5} + \frac{0.29\times10^4}{r^3} \tag{3}$$

These polynomial curves are also shown in Fig. 2. The above coefficients are optimum for the best fit possible. The mean departure we have tolerated between the observations and the polynomial fit is 10 per cent for $r/R_\odot \leqslant 1.5$ and 5 per cent for $r/R_\odot > 1.5$. J. Dürst (1981, personal communication) has confirmed that such a polynomial as above represents very well the brightness distribution in the corona.

3.2 *Polarization*

From the data secured, exposures 1, 2, 3 and 10 were considered most worthy of future analysis. The 12-s exposure shows the weak contribution from the sky background which increases rapidly in higher exposures. We designate the four images through the polaroids as *A, B, C* and *D,* the first one having its electric vector parallel to the polar axis of the Sun and the subsequent ones in steps of 45° with reference to *A*. The light from the corona is linearly polarized due to Thomson scattering and can be divided into two orthogonal components characterised by a maximum and a minimum. If this maximum intensity makes an angle θ with the direction of the electric vector of *A,* then for each point in the corona the polarization P and the position angle θ can be computed from the standard relations,

$$P = \left[\left(\frac{A-C}{A+C}\right)^2 + \left(\frac{B-D}{B+D}\right)^2\right]^{\frac{1}{2}}, \tag{4}$$

$$\theta = \tan^{-1}\left(\frac{B-D}{A-C}\right). \tag{5}$$

The films were scanned using the PDS microdensitometer system at the Kitt Peak National Observatory by one of us (K.R.S) with a 100-μm square aperture. For each polarized image, the digitised output of the microdensitometer consisted of a 320 × 320 term matrix. These were found to be highly resolved and hence five consecutive density values were block averaged and converted into an array of 64 × 64 points of relative intensity values. Such four arrays derived from the four images corresponding to the four orientations of the polaroids, belonging to one exposure were now combined according to Equation (4) and the polarization values in per cent were computed point by point. Even a small error in the assignment of position angles would introduce local differences between the four measures. We have taken extreme care in marking the N-S and E-W fiducial directions on each of the images using their enlarged transparencies and then transferring these onto the originals. While adopting this procedure we ensured that the misalignment errors in the orientations among the 4 images did not exceed 1°.

We repeated the above analysis for three more sets of exposures. The agreement between the four sets are so good (within 10 per cent) that we derived the mean values of

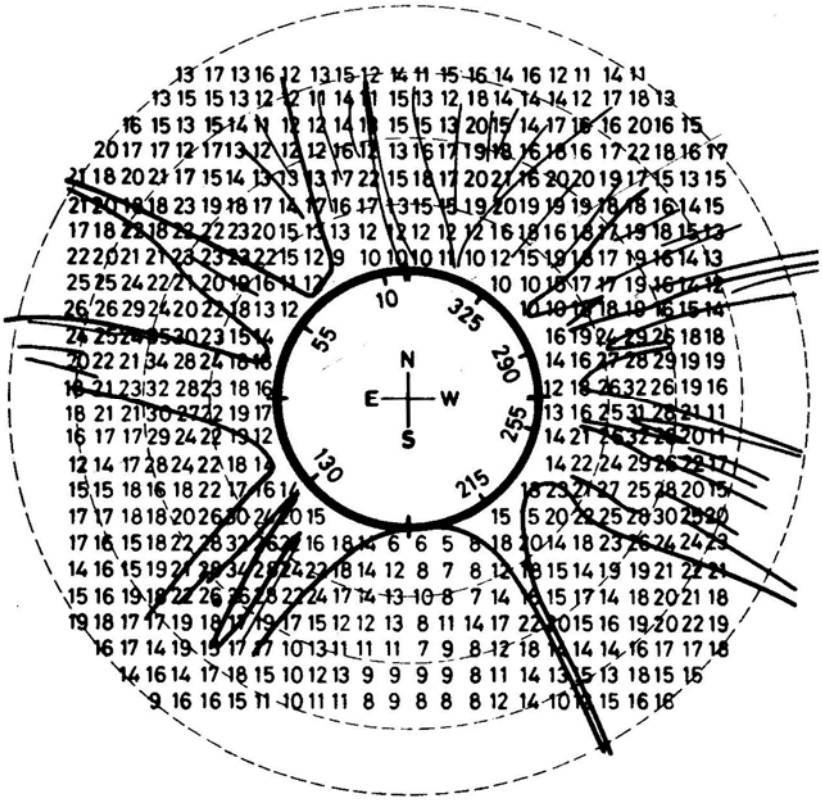


Figure 4. Polarization P $(K + F)$ of the corona. Each number represents the polarization in per cent over an area of 3.3 $(\text{arcmin})^2$. Superposed is a schematic representation to scale of the coronal streamers. Notice the very low polarization near the south. The four circles in dashed lines represent the radial distances in terms of the solar radius ($r/R_\odot$ = 1.5, 2.0, 2.5 and 3.0) from the Sun's centre.

polarization point by point from the four sets. To measure the sky light we made tracings on every film diagonally across and reaching upto 8 solar radii from the centre of the moon. The mean sky brightness from each set of exposures has values around $4.5 \times 10^{-10}\ \bar{B}_{\odot}$. The mean value of the sky polarization was found to be 0.055 and inclination 7° to the vertical. The polarization values derived for the corona are those after correction for the sky polarization for the respective sets of exposures. Since the neighbouring values of polarization in the 64 × 64 configuration were almost identical over most part of the corona, we once again averaged the polarization to obtain a 32 × 32 matrix formation. Thus each value of the polarization represents the mean over a fictitious aperture of area 1000 μm × 1000 μm corresponding to 200 square arcsec on the corona. These mean polarization values are presented in Fig. 4. On these are superposed a schematic scale representation of the coronal features to enable identification of the prominent coronal features and the polarization values associated with them. The angles of deviation of the magnetic vector from the radial direction were also computed and it was found that the maximum departure from the radial direction does not exceed ± 8°.

3.3 *Electron Density*

Our observations provide the intensity and polarization of the combined K and F corona represented by $(K + F)$ and P_{K+F} respectively at 32 × 32 points within the corona. We have used these to separate the K and F components in the customary way and have calculated the electron densities N_e (r) of the K corona. If we assume that the polarization of the F component is zero for the distances involved in our observations, then these quantities are related to the unknown quantity (K), the intensity of the K corona and the polarization P_K of the K corona alone by the relation

$$\frac{P_{K+F}}{P_K} = \frac{K}{K+F}. \tag{6}$$

Following the procedure adopted by von Klüber (1958) we computed

$$K_t - K_r = K P_K = (K+F) P_{K+F} \tag{7}$$

where the subscripts t and r denote the tangential and radial intensity components of K. $K_t - K_r$ can be represented by a power series of the form

$$K_t - K_r = \sum_s h_s r^{-s} \tag{8}$$

where r is the apparent distance of a given point from the centre of the Sun in the units of solar radius. The electron density can then be expressed by a power series of the form

$$N_e(r) = \sum_s \frac{h_s}{a_{s+1}} \cdot \frac{1}{r^{s+1}} \cdot \frac{1}{C} \cdot \frac{1}{A(r) - B(r)}. \tag{9}$$

Using the intensity values at every 10° interval in position angles derived from the isophotes, along with the polarization values at the corresponding points, we have computed the values of N_c (r) using the relation (9). We adopted the value $C = 3.44 \times 10^{-6}$ cm^3 and the values of a_s, $A(r)$ and $B(r)$ appropriate for the limb-darkening coefficient $q = 0.75$, from the compilations of van de Hulst (1950). In Fig. 5 we present the values of N_e (r) so derived over the entire corona at every 10° in position angle. The

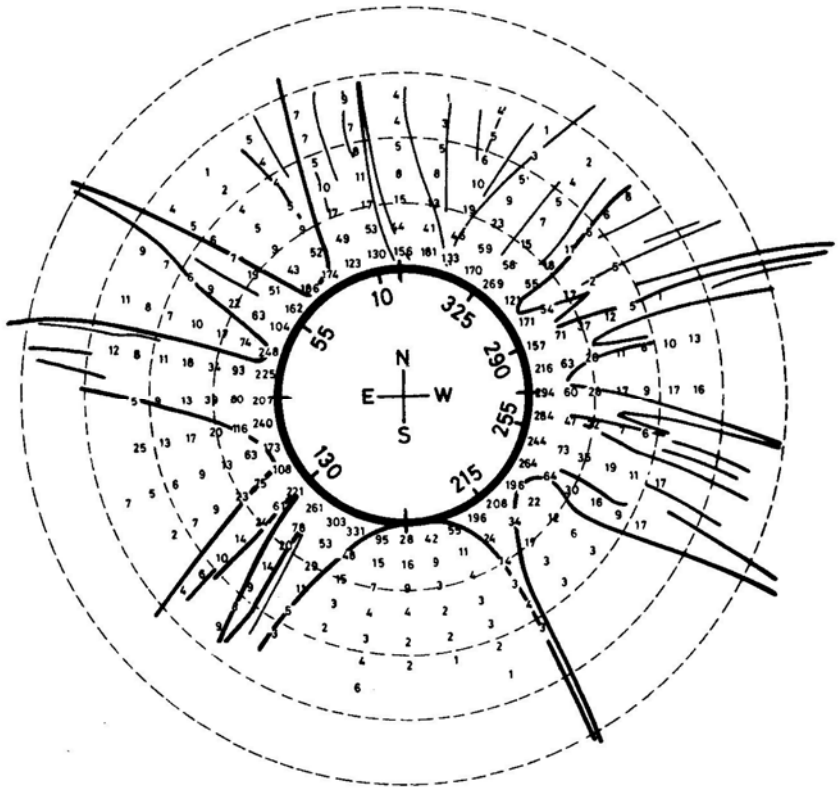


Figure 5. Distribution of the electron density N_e (cm^{-3}), in units of 10^{-7}, in the corona of 1980 February 16. The four dashed circles are at the radii $r/R_\odot$ = 1.5, 2.0, 2.5 and 3.0, respectively, from the centre of the Sun.

electron densities along N,S,E and W directions are plotted in Fig. 6.Dürst(1982) has derived the electron densities, representative of the mean background corona by averaging the values for 8 position angles free of coronal holes or streamers for the same eclipse. We find that a plot of these values included in Fig. 6, lies midway between our curves for the north and the equatorial directions. This brings out the agreement between the two independent observations of the same eclipse. The bright streamers have electron densities in the range 3.5–4.0 × 10^7 cm^{-3} at $1.5R_\odot$,while the region in the south between position angles 170° and 190° has extraordinarily low values in the range 3.5–6.0 × 10^6cm^{-3} similar to those in a coronal hole. In Fig. 6 the enhancement of N_e beyond 2.0 $R_\odot$ in the west, is caused by the streamer in this direction. The presence of such enhancements due to streamers have been illustrated well by Dollfus, Laffineur & Mouradian (1974) in their study of the electron density models of streamers. The brightness enhancement due to this streamer can also be identified in the brightness gradient curve in the west direction beyond 2.0 $R_\odot$ in Fig. 2.

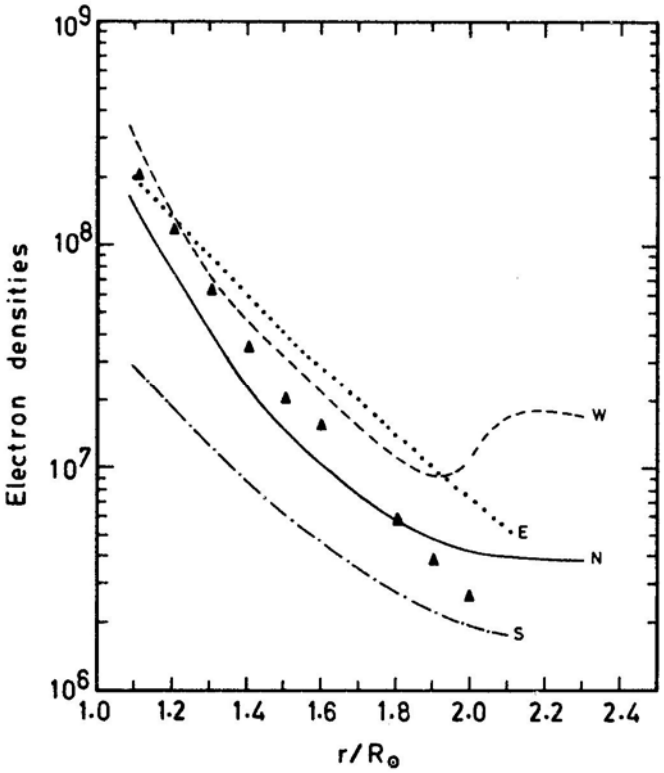


Figure 6. Electron density in the corona along the north, south, east and west directions. The electron density values of Dürst (filled triangles) for the same eclipse are also plotted for comparison. Notice the enhancement of N_e in the west at $r/R_\odot \sim 2$ due to a strong streamer.

Acknowledgements

The authors wish to thank the late Dr M. K. V. Bappu who was responsible for creating all the facilities for setting up these experiements. K.R.S. is thankful to Dr J. Dürst—formerly with the Institut für Astronomie, ETH Zentrum Zürich—for very valuable discussions during the latter's visit to India in January 1981. K.R.S. also wishes to express his gratitude to Mr Charles Mahaffey of the Kitt Peak National Observatory, Tucson, for his valuable help at the PDS microdensitometer system and in computations with the Cyber at KPNO. We are also thankful to Mr P. M. S. Namboodiri who helped us with the electron density computations.

References

Allen, C. W. 1973, *Astrophysical Quantities,* 3 edn, Athlone Press, London.
Bappu, M. K. V., Bhattacharyya, J. C., Sivaraman, K. R. 1973, *Pramana,* **1,** 117.
Dollfus, A., Laffineur, M., Mouradian, Z. 1974, *Solar Phys.,* **37,** 367.
Dürst, J. 1982, *Astr. Astrophys.,* **112,** 241.
Hata, S., Saito, K. 1966, *Ann. Tokyo astr. Obs.,* Ser. 2, **10**, 16.
van de Hulst, H. C. 1950, *Bull. astr. Inst. Netherl.,* **11,** 135.
von Klüber, H. 1958, *Mon. Not. R. astr. Soc.,* **118,** 201.
Waldmeier, M. 1959, Z. *Astrophys.,* **48,** 9.

J. Astrophys. Astr. (1984) **5,** 159–168

Spot Activity in the RS CVn Binary UX Arietis

M. B. K. Sarma & B. V. N. S. Prakasa Rao *Nizamiah Observatory, Centre of Advanced Study in Astronomy, Osmania University, Hyderabad 500007*

Received 1983 October 19; accepted 1984 January 25

Abstract. Photoelectric observations of the RS CVn type non-eclipsing binary UX Arietis obtained at Nizamiah Observatory during the observing seasons of 1975–76, 1981–82 and 1982–83 are presented. The light curve of UX Ari showed a distortion wave with an amplitude in V varying from 0.02 mag during 1975–76 to 0.15 mag during 1982–83. An analysis of the available data shows that the light maximum is almost constant. It is also evident that the light-curve minimum decreases as the wave amplitude increases. The constant light at maximum, $V = 6.51 \pm 0.03$ indicates the unspotted photospheric brightness. It is also suggested that the variation in mean V brightness is mainly due to spot activity and not due to intrinsic variation.

Key words: photoelectric photometry—RS CVn variables—stars, individual

1. Introduction

UX Arietis (HD 21242), an RS CVn type system, is a double-lined spectroscopic binary with an orbital period of 6.44 d. Carlos & Popper (1971) have classified the components as G5 V and K0 IV. This system displays quasi-sinusoidal light variations modulated with the orbital period. A series of light curves have been obtained at a number of observatories (Guinan *et al.* 1981 and the references therein). Guinan *et al.* (1981) have showed that the wave minimum had retrograde motion during 1972–77 while it appeared to have direct motion during early and late 1980. Zeilik *et al.* (1982) have reported that the minimum phase of the distortion wave has remained steady for a year (early to late 1981) and will perhaps have retrograde motion again.

The amplitude of the wave varied non-uniformly having the range 0.09 to 0.05 mag between 1974.9 to 1980.0 and increasing from 0.06 to 0.17 mag during late 1980 (Guinan *et al.* 1981). The non-uniform variation of the amplitude of the wave has been attributed to the variation in the total spot area or to the change in spot location or both. On the other hand, as we have already reported (Sarma, Prakasa Rao & Ausekar 1983, hereafter SPA) the behaviour of the phase of light minimum and amplitude of the distortion wave is systematic and not erratic as reported by Zeilik *et al.* (1982). It is evident from Fig. 2 of SPA that the motion of the phase of wave minimum is direct in the beginning of a cycle and becomes retrograde after the spots migrate below the co-rotating latitude (the latitude that is rotating in synchronism with the orbital period of the system). The amplitude of the wave is constant during one spot cycle. The period of this cycle is estimated to be ~ 5–6 yr.

In this communication we present photometry of UX Ari in *B* and *V* bands for three observing seasons and discuss the wave migration, the nature of spot activity and intrinsic variability of this system.

2. Observations

UX Ari was observed during 1975–76, 1981–82, and 1982–83 seasons using the 38-cm Grubb refractor of the Nizamiah Observatory with an unrefrigerated RCA 1P 21 & EMI 9502B photomultipliers. The photocurrent was amplified with a GR 1230A DC amplifier and recorded on a Honeywell-Brown strip-chart recorder. Standard Johnson *B, V* filters were used. 62 Ari and HR 999 were used as comparison and check stars respectively. The observations of the comparison star (62Ari) were used for determining the nightly extinction coefficients. The observations were made differentially with respect to the comparison star 62 Ari and were transformed to the standard Johnson system. The constancy of Δm (HR 999–62 Ari) values suggests that the comparison star remained constant during the period of our observations within ± 0.02 mag in *B* and *V* passbands. The transformation coefficients, the magnitude and colour of the comparison star for the three years of observations are given in Table 1.

The phases were calculated using the following ephemeris given by Landis *et al* (1978)

$$\mathrm{HJD} = 2440133.75 + 6^{\mathrm{d}}.43791\,E \tag{1}$$

where zero phase corresponds to conjunction with the cooler component in front and the period is the orbital period determined spectroscopically by Carlos & Popper (1971). All the observations *i.e.* Δm (variable – comparison) are grouped according to phases and normal points are formed for each passband. These are given in Tables 2 and 3 for yellow and blue passbands. The plot of ΔV versus phase is shown in Fig. 1. In order to determine the amplitude of the wave and the phase of its light minimum, individual observations are fitted to the truncated Fourier series (Sarma & Ausekar 1980),

$$l = A_0 + A_1 \cos\theta + A_2 \cos 2\theta + B_1 \sin\theta. \tag{2}$$

The values of Am adopted to normalize *l* to 1 for each light curve are given in Table 4. The coefficients A_0, A_1, A_2 and B_1 are determined for each year of observation using the least-squares method. These coefficients, along with the full amplitude

Table 1. Transformation coefficients.

Year	Photomulti-plier	ε	μ	V of C_1 (mag)	$(B-V)$ of C_1 (mag)
1975–76	RCA 1P21	−0.057 ±0.010	1.412 ±0.010	5.54	1.08
1981–82	EMI 9502B	−0.123 ±0.008	1.444 ±0.008	5.54	1.09
1982–83	EMI 9502B	−0.121 ±0.005	1.507 ±0.008	5.55	1.09

Table 2. UX Arietis: Normal points in *V* band.

HJD 2440000 +	Phase	ΔV	*N*
1975–76			
2799.1237	.0123	1.017	6
.1611	.0181	1.012	8
.1886	.0224	1.019	4
2787.2273	.1644	1.009	10
2755.1304	.1788	1.001	5
.1553	.1827	1.005	4
.1804	.1866	1.008	4
2801.1043	.3199	1.009	1
.1214	.3226	0.996	6
.1526	.3274	1.000	6
.1754	.3310	0.993	3
2757.1262	.4888	1.002	2
.1476	.4921	1.006	4
.1853	.4980	1.003	6
.2393	.5064	0.992	6
.2747	.5119	0.988	2
2784.1022	.6790	0.982	3
.1467	.6859	0.991	8
.1903	.6927	0.998	5
2759.1735	.8068	1.003	8
.2395	.8171	1.015	6
.2690	.8216	1.001	2
2772.1625	.8244	1.001	9
.2163	.8327	0.998	8
2785.1097	.8355	1.029	5
.1674	.8444	1.013	10
1981–82			
4930.1732	.0280	1.119	4
.1975	.0318	1.125	5
4931.1817	.1846	1.091	12
.2262	.1916	1.080	4
5002.1529	.2086	1.042	3
.1820	.2131	1.018	8
4996.2094	.2854	1.057	8
4951.1501	.2863	1.053	4
.1774	.2906	1.073	2
5009.1659	.2979	1.070	6
.1957	.3025	1.060	6
5003.1069	.3568	1.007	8
.1615	.3653	1.006	7
4997.1919	.4380	1.031	4
4952.1286	.4383	1.016	2
4997.2156	.4417	1.030	7
4952.1601	.4432	1.022	6
5010.1428	.4497	1.014	2
.1677	.4535	1.036	9
4959.2191	.5397	0.991	2
.2543	.5452	0.995	10
.2938	.5513	0.979	4
4972.2590	.5652	0.966	5
.2932	.5705	0.972	2
4998.2091	.5960	1.001	12
4928.1816	.7186	1.043	2
.2288	.7260	1.025	7
4986.1850	.7283	1.028	4
.2167	.7332	1.042	14
4935.1419	.7998	1.050	2
.1782	.8054	1.092	12
.2151	.8112	1.100	7
4993.2067	.8190	1.090	4
.2315	.8228	1.100	7
4929.3387	.8984	1.169	2
.3818	.9051	1.165	13
4936.1571	.9575	1.148	7
.1990	.9640	1.174	9
1982–83			
5355.2072	.0485	1.113	2
.2414	.0538	1.098	4
5323.2191	.0798	1.078	1
.2545	.0853	1.056	14
.2976	.0920	1.057	5
5375.1139	.1406	1.042	3
5356.1609	.1966	0.994	10
.1967	.2022	0.987	5
5324.2353	.2376	0.999	4
.2821	.2449	0.994	12
.3260	.2517	0.975	4
5318.3023	.3161	0.965	10
.3345	.3211	0.997	2
5325.2532	.3957	0.975	11
.3027	.4034	0.986	8
5377.1358	.4546	0.944	3
5319.2630	.4653	0.967	9
.3213	.4744	0.970	6
5326.2363	.5484	0.943	4
.2801	.5552	0.951	12
.3127	.5603	0.961	2
5339.2365	.5678	0.926	5
.2691	.5728	0.913	7
5314.3055	.6952	0.962	7
.3390	.7005	0.951	2
5353.2011	.7369	0.940	5
.2293	.7413	0.936	3
5354.1826	.8894	1.123	2
.2163	.8946	1.131	6
5322.2382	.9274	1.134	4
.2851	.9347	1.157	13
.3255	.9410	1.152	2
5374.1309	.9879	1.136	2
.1582	.9922	1.108	4

Table 3. UX Arietis: Normal points in *B* band.

HJD 2440000 +	Phase	ΔB	N
1975–76			
2799.1238	.0123	0.838	6
.1592	.0178	0.830	7
.1864	.0220	0.844	5
2787.2269	.1643	0.821	10
2755.1301	.1787	0.829	4
.1562	.1828	0.838	4
.1804	.1866	0.811	4
2801.1211	.3225	0.805	8
.1588	.3284	0.818	4
.1756	.3310	0.815	3
2757.1272	.4890	0.820	2
.1463	.4919	0.818	3
.1856	.4980	0.811	6
.2390	.5063	0.801	6
.2739	.5117	0.792	2
2784.1020	.6789	0.824	3
.1465	.6859	0.806	8
.1911	.6928	0.804	6
2759.1735	.8068	0.800	8
.2397	.8171	0.812	6
.2683	.8215	0.798	2
2772.1636	.8246	0.799	9
.2190	.8332	0.805	7
2785.1102	.8355	0.837	5
.1675	.8444	0.817	10
1981–82			
4930.1744	.0282	0.925	4
.1991	.0320	0.921	4
4931.1798	.1844	0.864	13
.2270	.1917	0.869	4
5002.1529	.2086	0.848	1
.1863	.2138	0.834	9
4996.2113	.2857	0.858	8
4951.1460	.2857	0.868	3
.1787	.2908	0.826	2
5009.1667	.2981	0.867	5
.1962	.3026	0.835	6
5003.1094	.3572	0.868	6
.1620	.3653	0.851	7
.2011	.3714	0.859	2
4997.1926	.4381	0.850	4
4952.1300	.4386	0.798	3
.1617	.4435	0.816	5
4997.2162	.4418	0.848	5
5010.1422	.4496	0.853	1
.1664	.4533	0.833	10
4959.2187	.5396	0.791	1
.2522	.5448	0.793	11
.2948	.5515	0.795	4
4972.2035	.5566	0.778	1
.2683	.5666	0.815	3
.2941	.5707	0.786	2
4998.2099	.5961	0.797	12
4928.1822	.7187	0.778	2
.2296	.7261	0.815	7
4986.1858	.7284	0.832	4
.2172	.7333	0.827	12
4935.1804	.8058	0.914	12
.2151	.8112	0.896	5
4993.2078	.8192	0.894	4
.2320	.8229	0.910	6
4929.3406	.8987	0.953	2
.3835	.9053	0.944	11
4936.1578	.9576	0.946	8
.1934	.9631	0.984	9
1982–83			
5355.2117	.0492	0.858	3
.2415	.0538	0.875	3
5323.2197	.0799	0.882	1
.2546	.0853	0.876	15
.2965	.0918	0.871	7
5375.1191	.1414	0.828	1
5356.1596	.1964	0.774	9
.1962	.2021	0.789	5
5324.2361	.2378	0.761	4
.2789	.2444	0.823	12
.3244	.2515	0.825	5
5318.3005	.3158	0.792	9
.3330	.3208	0.815	3
5325.2559	.3962	0.800	12
.3035	.4036	0.772	8
5377.1248	.4529	0.818	1
5319.2603	.4648	0.811	8
.3155	.4734	0.785	8
5326.2369	.5486	0.814	4
.2808	.5554	0.774	12
.3134	.5604	0.756	2
5339.2291	.5666	0.733	7
.2701	.5730	0.724	7
5314.3070	.6955	0.783	9
.3402	.7006	0.770	2
5353.1999	.7367	0.780	5
.2337	.7420	0.754	6
5354.1833	.8895	0.895	1
.2143	.8943	0.900	7
5322.2395	.9276	0.952	4
.2861	.9348	0.953	11
.3242	.9408	0.953	3
5374.1299	.9878	0.931	1
.1600	.9924	0.936	7

Table 4. Fourier coefficients.

Epoch (Year)	Filter	Δm at $l = 1$ (mag)	A_0	A_1	A_2	B_1	Amplitude (mag)	ϕ_{min}
1976.0	*V*	1.004	+1.0011	−0.0087	−0.0039	+0.0009	0.019	0.994
			±0.0009	±0.0014	±0.0012	±0.0011	±0.003	±0.006
	B	0.814	+0.9994	−0.0090	−0.0064	+0.0042	0.024	0.019
			±0.0009	±0.0014	±0.0013	±0.0012	±0.004	±0.006
1982.0	*V*	1.060	+0.9932	−0.0708	−0.0124	−0.0024	0.154	0.003
			±0.0012	±0.0019	±0.0016	±0.0017	±0.004	±0.002
	B	0.862	+0.9943	−0.0650	−0.0159	−0.0034	0.142	0.004
			±0.0013	±0.0019	±0.0017	±0.0018	±0.004	±0.002
1983.0	*V*	1.011	+1.0025	−0.0761	−0.0304	−0.0010	0.166	0.001
			±0.0014	±0.0018	±0.0020	±0.0022	±0.004	±0.002
	B	0.822	+1.0019	−0.0636	−0.0294	−0.0023	0.138	0.002
			±0.0015	±0.0019	±0.0020	±0.0023	±0.004	±0.002

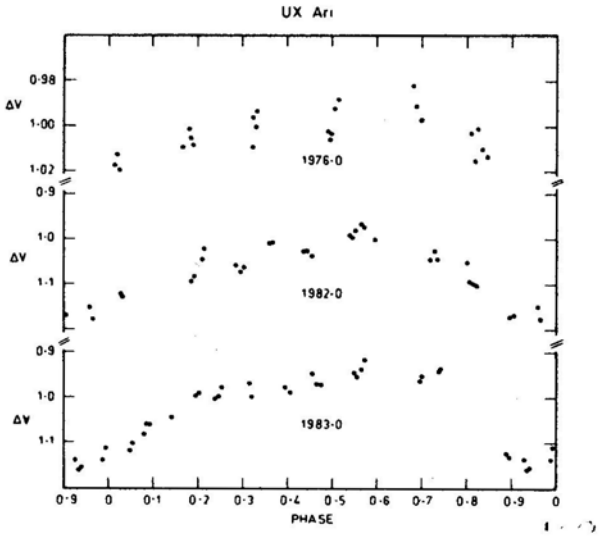


Figure 1. UX Ari: Plot of Δm (normal points) versus phase in V band.

$2(A_1^2 + B_1^2)^{1/2}$ in magnitude units and phase of light minimum, are also given in the table.

3. Discussion

The data available in the literature on the epoch, phase of light minimum, full amplitude, ΔV_{max}, ΔV_{min} and V_{mean} for UX Ari are listed in Table 5.

Table 5. UX Arietis: Photometric properties.

Observing Epoch (year)	Phase of light minimum	Amplitude V (mag)	ΔV_{max} (mag)	ΔV_{min} (mag)	V_{mean} (mag)	References
1972.2	0.14	0.11	+0.865	+0.975	+6.46	1
1972.8	0.12	0.14	0.870	1.010	6.48	1
1974.9	0.99	0.05	1.015	1.065	6.58	1
1976.0	0.006	0.02	0.990	1.010	6.54	2
1976.8	0.80	0.05	1.005	1.055	6.57	1
1977.1	0.68	0.08	0.990	1.070	6.57	1
1979.2	0.03	0.09	0.945	1.035	6.53	1
1979.9	0.87	0.04	0.970	1.010	6.53	1
1980.0	0.90	0.06	0.970	1.030	6.54	1
1980.8	0.95	0.17	0.950	1.120	6.575	1
1981.0	0.97	0.16	0.960	1.120	6.58	1
1981.8	0.97	0.21	0.940	1.150	6.585	3
1982.0	0.004	0.15	1.000	1.160	6.62	2
1983.0	0.002	0.15	0.950	1.150	6.59	2

References:
1. Guinan *et al.* (1981)
2. Present study
3. Zeilik *et al.* (1982)

3.1 *Period of Wave Migration*

Zeilik *et al.* (1982) have reported that the phase of light minimum and amplitude of UX Ari showed an erratic behaviour during the past ten years. However, it is evident from Fig. 2 of SPA that the migration of wave minimum on the light curve was first direct while it was later moving towards decreasing orbital phases. A complete cycle is estimated to last about 5 to 6 years. At the end of the 1975–80 cycle, it appears that there has been a break in the smooth retrograde motion and a new cycle with a direct motion appears to have started around 1981.

SPA have grouped the spots formed before 1975, between 1975–80 and 1981 onwards into three different cycles. Applying the analogy of the sunspots, it is assumed that spots first form on the star at a latitude higher than the co-rotating latitude. During their migration towards the equator the spots will have a direct motion before crossing the co-rotating latitude and a retrograde motion after crossing it. The spots of one cycle disappear after reaching the vicinity of the equator and the spots of the next cycle form above the co-rotating latitude. If the assumption of the solar analogy is applicable, two factors regarding the spots have to be considered. First, if the spots are migrating towards the equator, the rate of wave migration should increase along the cycle; secondly, as the spot cycle nears its end, its activity should reduce resulting in the decrease of the wave amplitude, provided there is no overlapping of spot cycles. We estimate from the data given in Table 5, for the cycle of 1975–80, that the wave migration has a rate of about 0.12 phase per year during its direct motion (1974.9–1975.3) and about 0.23 phase per year during its retrograde motion (1979.2–1979.9), in accordance with the idea that the spots are migrating from higher to lower latitudes. From Fig. 2 of SPA it appears that there is not much variation in the amplitude of the wave during the cycle 1975–80, suggesting that the nature of the spot or spots has not varied much in this cycle. If the solar analogy holds for UX Ari, this constancy in amplitude may be attributed to the overlap of spot cycles. In the absence of more direct evidence—such as photometric complications—indicating the coexistence of different spot cycles, we cannot be positive that the analogy of sunspot cycle holds for the spot cycles in UX Ari, as it does in RS CVn itself (Blanco *et al.* 1982). However, further observations over many cycles are necessary for a number of RS CVn systems before the final conclusion on the applicability of solar analogy to star spots is reached.

3.2 *Colour Dependence*

In order to understand the colour changes of UX Ari, a plot of $\Delta (B - V)$ versus phase is shown in Fig. 2 for the years 1976.0, 1982.0 and 1983.0. We infer from the figure that even though there is some scatter, the $\Delta(B - V)$ colour curve has remained flat during the three years. Also, it can be seen from Table 4 that the amplitude in B and V passbands is fairly constant during each year suggesting that the amplitude is colour-independent. Because of this, any information regarding the physical nature of the spots cannot be obtained immediately.

3.3 *Nature of the Spot Activity*

The wave amplitude showed a remarkable change from $\Delta V = 0.04$ mag in 1979.9 to 0.21

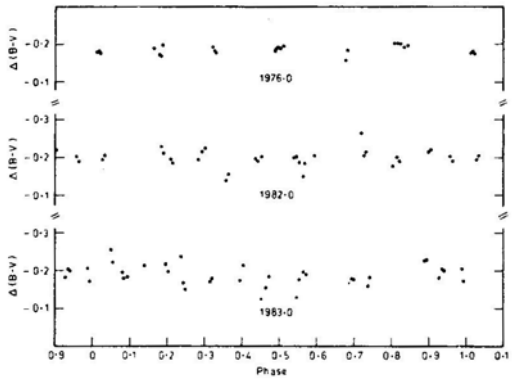


Figure 2. UX Ari: Plot of Δ ($B - V$) versus phase. Note the flatness of the colour curve during the three years of our observations.

mag in 1981.8. The observed changes could have occurred due to the variation of the spot sizes and also due to their distribution in latitude and longitude. In order to understand this aspect, it is necessary to have additional data such as the mean magnitude and the maximum and minimum brightness. These values are given in Table 5. Fig. 3 shows the variation of ΔV at light curve maximum and minimum, with the amplitude. Excluding the two points of 1972–73 observations (amplitudes of 0.11 and 0.14 mag respectively), it appears that ΔV at light maximum has remained almost constant around a value of $\Delta V_{max} = 0.97 \pm 0.03$ mag from late 1974 to early 1983. If we assume that this maximum represents the normal unspotted photospheric brightness of UX Ari, the magnitude of the system will be $V = 6.51 \pm 0.03$ (taking $V = 5.54$ for the comparison star 62 Ari). Assuming a difference of 2.0 mag between the two components (G5 V and K0 IV), the brightness of the individual components of UX Ari are obtained as $V = 6.7$ for K0IVand 8.7 for G5 V. Taking the values of absolute magnitudes at these

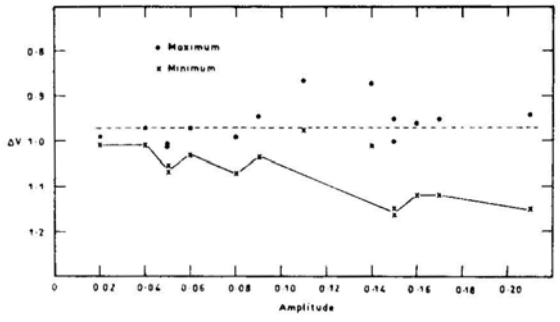


Figure 3. UX Ari: Plot of ΔV versus amplitude. ΔV_{max} has remained almost constant at a value of 0.97 ± 0.03 mag during 1974–83.

spectral types from Allen (1976), one gets a distance of about 50 pc for this system which is in agreement with the value given by Hall (1976). Hence, we feel that a value of $V = 6.51 \pm 0.03$ mag for the photosphere of UX Ari is appropriate.

It is also seen from Fig. 3 that while the light at maximum remained almost constant, the light at minimum decreased with the increasing wave amplitude by an amount which can account for the changes in the wave amplitude. In HR 1099, another RS CVn system, the light at maximum as well as at minimum was found to vary with amplitude (Bartolini *et al.* 1983). It appears that during the period 1975–83 the spot distribution and activity are different for these two RS CVn type stars.

Since in UX Ari ΔV_{max} is found to be almost constant, while ΔV_{min} is changing, the low amplitude of the wave can be attributed to the low spot activity, and the high amplitude to high spot activity and/or concentration of spots at a preferred longitude. We can then infer that during the cycle of 1975–80 (amplitude < 0.1 mag) the spot activity on UX Ari was lower than in the present cycle (amplitude > 0.1 mag). It is necessary to have continuous observations over many cycles to understand more about the spot activity.

3.4 *Intrinsic Variation*

Some RS CVn systems like 39 Cet (Sarma, unpublished), TY Pyx (Vivekananda Rao & Sarma 1981) and UV Psc (Vivekananda Rao & Sarma 1984) are found to be intrinsically variable. UX Ari was also suspected to be an intrinsic variable by Hall (1977). To study the nature of the intrinsic variation of UX Ari, V_{mean} is plotted against time in Fig. 4. V_{max}, V_{min} and amplitude are also plotted in this figure. It is seen that the amplitude, V_{mean} and V_{min} have continuously changed (V_{max} has remained almost constant at a value of 6.51 ± 0.03 mag) during 1974–83 in the sense that the amplitude is smaller when V_{mean} and V_{min} are brighter. We obtain a correlation coefficient $\rho = 0.64$ (excluding the

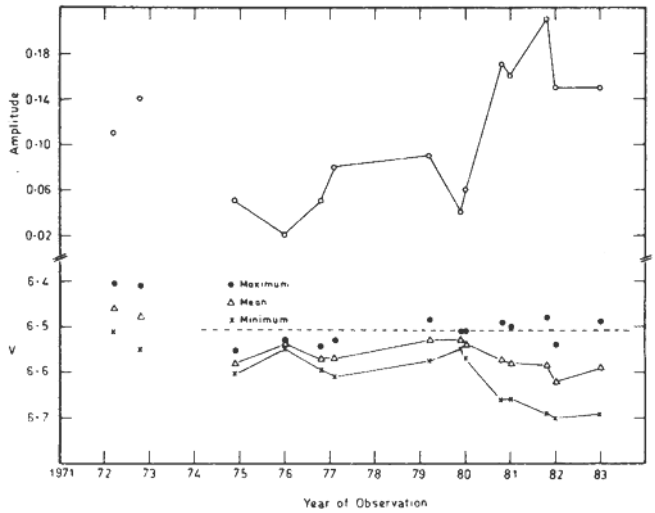


Figure 4. UX Ari: Plot of V magnitude and amplitude versus epoch of observations. It is evident that the V_{mean} curve is following the V_{min} curve.

two points for 1972–73) between V_{mean} and amplitude. The V_{mean} curve has simply followed the V_{min} curve. From this we conclude that the variation of V_{mean} is mainly due to spot activity and not intrinsic variation.

The 1972–73 data of Hall indicate a brightening of the system by ~ 0.1 mag and do not follow the general trend of other observations. In the absence of additional data for the cycles prior to 1975, it is difficult to explain the behaviour of UX Ari during 1972–73. Further observations are needed to ascertain the intrinsic variability of this system.

4. Conclusions

The observations of UX Ari in *B* and *V* bands during the years 1975–76, 1981–82 and 1982–83 yield an average ϕ_{min} of about 0.000 ± 0.003 in phase, and an amplitude of 0.02 mag during 1976.0 and 0.15 mag during 1982.0 and 1983.0. V_{max} is found to be almost constant during 1974–83 and assuming this to be the unspotted photospheric brightness, a magnitude of $V = 6.51 \pm 0.03$ mag is estimated. It is also evident that the amplitude of the wave is directly related to the distribution and activity of the spots. It appears that the contribution to V_{mean} due to intrinsic variation is negligible.

Continuous observations over a long period of time in many passbands are required in order to understand the physical nature of the spot activity, the spot cycle and intrinsic variation of UX Ari.

Acknowledgements

We are grateful to Professor K. D. Abhyankar for his useful suggestions. Thanks are also due to B. D. Ausekar for his help at the telescope. One of the authors (B.V.N.S.P. Rao) wishes to thank the University Grants Commission, New Delhi for awarding a fellowship during the period of work.

References

Allen, C. W. 1976, *Astro physical Quantities,* 3 edn, Athlone Press, London.

Bartolini, C., Blanco, C., Catalano, S., Cerruti-Sola, M., Eaton, J. A., Guarnieri, A., Hall, D. S., Henry, G. W., Hopkins, J. L., Landis, H. J., Louth, H., Marilli, E., Piccioni, A., Renner, T. R., Rodonò, M., Scaltriti, F. 1983, *Astr. Astrophys.,* **117,** 149.

Blanco, C., Catalano, S., Marilli, E., Rodono, M. 1982, *Astr. Astrophys.,* **106,** 311.

Carlos, R. C, Popper, D. M. 1971, *Publ. astr. Soc. Pac.,* **83,** 504.

Guinan, E. F., McCook, G. P., Pragola, J. L., O' Donnell, W. C, Weisenberger, A. C. 1981, *Pub. astr. Soc. Pacific,* **93,** 495.

Hall, D.S.1976, in *IAU Coll. 29: Multiple Periodic Variable Stars: Invited papers* Ed. W. S. Fitch, D. Reidel, Dordrecht, p. 287.

Hall, D. S. 1977, *Acta Astr.,* **27,** 281.

Landis, H. J., Lovell, L. P., Hall, D. S., Henry, G. W., Renner, T. R. 1978, *Astr. J.,* **83,** 176.

Sarma, M. B. K., Ausekar, B. D. 1980, *Acta Astr.,* **30,** 101.

Sarma, M. B. K., Prakasa Rao, B. V. N. S., Ausekar, B. D. 1983, *Inf. Bull. Var. Stars,* No. 2357 (SPA).

Vivekananda Rao, P., Sarma, M. B. K. 1981, *Acta Astr.,* **31,** 107.

Vivekananda Rao, P., Sarma, M. B. K. 1984, *Astrophys. Space Sci.* (in press).

Zeilik, M., Elston, R., Henson, G., Smith, P. 1982, *Inf. Bull. Var. Stars,* No. 2168.

J. Astrophys. Astr. (1984) **5,** 169–180

Effects of Partial Frequency Redistribution Functions R_{II}, R_{III} and R_V on Source Functions

D. Mohan Rao, K. E. Rangarajan & A. Peraiah *Indian institute of Astrophysics, Bangalore 560034*

Received 1982 May 20; accepted 1984 March 1

Abstract. The effects of partial frequency redistribution on the formation of spectral lines have been studied. We considered the angle-averaged R_{II}, R_{III} and R_v types of redistribution with isotropic phase function, Transfer equation with plane-parallel geometry is solved in isothermal atmospheres. For an atmosphere with constant thermal sources, the frequency-dependent source function S_L (R_v) lies below S_L (R_{III}) but above S_L (R_{III}) in the line wings.

Key words: radiative transfer—partial frequency redistribution—line source functions—spectral line formation

1. Introduction

Frequency-dependent source functions were studied by Hummer (1969) in semi-infinite and finite isothermal atmospheres. In the wings, large differences were found to exist between the complete redistribution (CRD) and partial redistribution (PRD) source functions. The effects of photon frequency and angular redistribution on line formation using R_I, R_{II} and R_{III} functions and their role in finite and semi-infinite plane-parallel media were studied in a series of papers by Vardavas (1976 a,b,c), using angle-dependent redistribution functions. Vardavas made a comparison with CRD and also with the results of angle-averaged redistribution functions. The differences between the emergent profiles using CRD with a Voigt absorption profile and R_{III} function was found to be negligible (Vardavas 1976 b). Similar conclusion was arrived at by Finn (1967). R_{II} redistribution function, which is strongly coherent in the wings, was shown to lower the line profile outside the Doppler core (Hummer 1969; Vardavas 1976 c). R_I angle-dependent and angle-averaged redistribution functions were studied in spherecally symmetric expanding media by Peraiah (1978). In moving media he obtained P-Cygni type of profiles. Milkey & Mihalas (1973) used a combination of R_{II} and R_{III} redistribution functions in explaining Solar Lyman-α resonance-line profile.

As far as subordinate lines are concerned, Heinzel (1981) derived the correct laboratory frame redistribution function (LFR) for the scattering of radiation assuming both the atomic levels to be radiatively broadened. This LFR denoted as R_v (x', $\tilde{n}'$: x, $\tilde{n}$) is based on quantum-mechanical results of Omont, Smith & Cooper (1972). R_v can be applied in low-density media like chromospheres, gaseous nebulae *etc.,* where collisions are few. In a subsequent paper, Heinzel & Hubený (1982) extended their LFR to include collisional broadening of both the levels. Some transfer effects of R_v have been discussed by Hubený & Heinzel (1984).

The aim of this paper is to study the influence of angle-averaged R_V redistribution function in radiative transfer calculations. To make a comparative study, we also evaluated the source functions using R_{II} and R_{III}. In Section 2 we have briefly discussed the functions R_{II}, R_{III} and R_v. We have described the basic equations and the computational procedure in Section 3. Results are discussed in Section 4.

2. Redistribution function

When the lower level is sharp and the upper level is radiatively broadened (assuming isotropic phase function), the redistribution function is given by (Hummer 1962)

$$R_{II}(x', x) = \frac{1}{\pi^{3/2}} \int_{\left|\frac{x'-x}{2}\right|}^{\infty} e^{-u^2} \left[\tan^{-1}\left(\frac{\underline{x}+u}{a_j}\right) - \tan^{-1}\left(\frac{\bar{x}-u}{a_j}\right)\right] du, \tag{1}$$

where $\bar{x}$ = max ($|x|$, $|x'|$) and $\underline{x}$ = min ($|x|$, $|x'|$) x' is the frequency of incoming photon measured in Doppler width units, x is the frequency of emitted photon and α_j is the damping constant for the upper level. We chose $\alpha_j = 2 \times 10^{-3}$ and evaluated the above function. The profile function $\phi(x)$ is given by

$$\phi(x) = \int_{-\infty}^{\infty} R_{II}(x', x)\, dx' = H(a_j, x), \tag{2}$$

where $H(\alpha_j, x)$ is the normalized Voigt function. We have plotted $R_{II}(x', x)/\phi(x')$ in Fig. 1 (a). This quantity is the probability of emission at frequency x per absorption, when the absorption is at frequency x'. From this figure we see the coherency for wing photons and that they have the least probability of being emitted at the line centre. In the Doppler core, R_{II} behaves like other redistribution functions.

Radiative and collisional broadening of the upper level with isotropic phase function gives rise to a redistribution function of the form

$$R_{III}(x', x) = \frac{1}{\pi^{5/2}} \int_0^{\infty} e^{-u^2} \left[\tan^{-1}\left(\frac{x'+u}{a_j}\right) - \tan^{-1}\left(\frac{x'-u}{a_j}\right)\right] \times \left[\tan^{-1}\left(\frac{x+u}{a_j}\right) - \tan^{-1}\left(\frac{x-u}{a_j}\right)\right] du. \tag{3}$$

The absorption profile function in this case is defined in a similar way as above. We have plotted $R_{III}(x', x)/\phi(x')$ for $\alpha_j = 2 \times 10^{-3}$ and 10^{-3} in Fig. l(b) and (c) respectively. $R_{II,III}$ are generated using Simpson's rule with small intervals. From these figures we see that the wing photons get completely redistributed and they have a high probability of being emitted at the centre.

When the lower and upper levels are broadened by radiative damping, the angle-dependent LFR is given by (Heinzel 1981)

$$R_V(x', \tilde{n}'; x, \tilde{n}) = \frac{1}{4\pi^2 \sin\theta} \left[H\left(a_j \sec\frac{\theta}{2}, \frac{x+x'}{2}\sec\frac{\theta}{2}\right) H\left(a_j \csc\frac{\theta}{2}, \frac{x-x'}{2}\csc\frac{\theta}{2}\right) + E_V(x', x, \theta)\right], \tag{4}$$

where θ is the angle between incoming ($\tilde{n}'$) and outgoing ($\tilde{n}$) photon directions. The

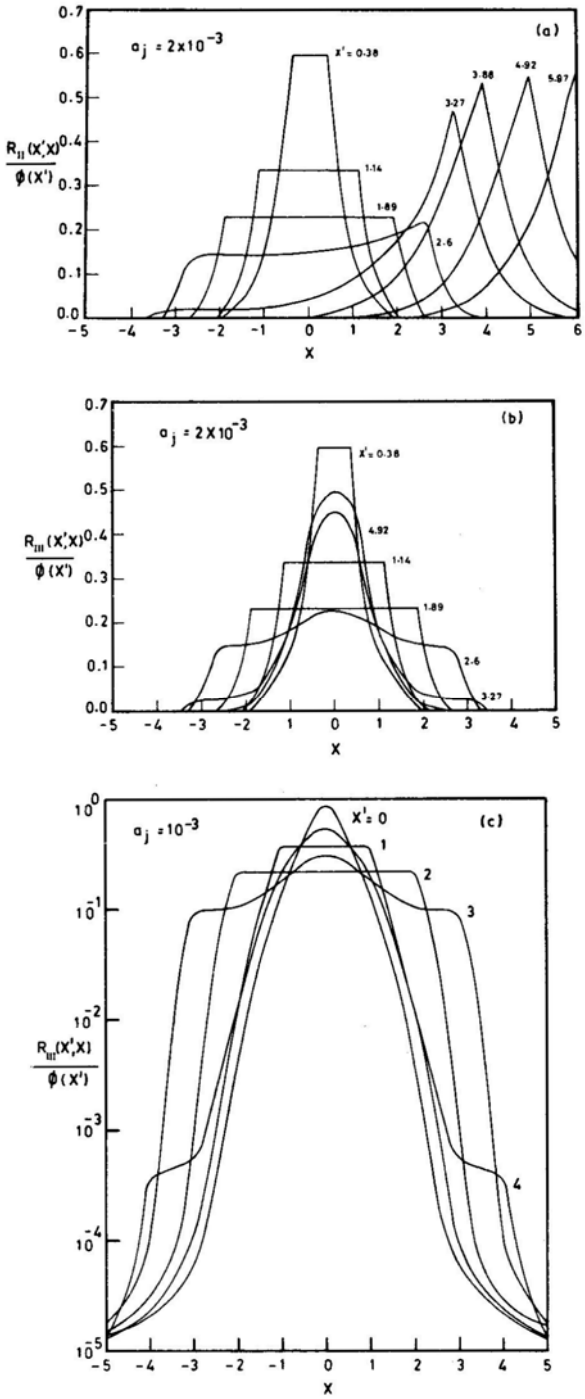


Figure 1. The probability of emission $R(x', x)/\phi(x')$ at frequency x per absorption when the absorption is at frequency x' is plotted for (a) R_{II} with $a_j = 2 \times 10^{-3}$, (b) R_{III} with $a_j = 2 \times 10^{-3}$, and (c) R_{III} with $a_j = 10^{-3}$.

function E_ν is given by

$$E_\nu(x', x, \theta) = \frac{\sin(\theta/2)}{\pi^{1/2}} \mathrm{Re} \int_0^\infty e^{-t^2} (e^{-2wt} + e^{-2w't}) \Delta(t) dt, \quad (5)$$

Where

$$\Delta(t) = D\left(z + t\cos\frac{\theta}{2} + a_i \sec\frac{\theta}{2}\right) - D\left(z + t\cos\frac{\theta}{2}\right),$$

$$z = \sec\frac{\theta}{2}\left(a_j - i\frac{x + x'}{2}\right),$$

$$w = a_i + a_j - ix,$$

$$w' = a_i + a_j - ix',$$

$$D(\omega) = H(p, q) + iK(p, q),$$

$$\omega = p - iq,$$

a_i and a_j being the damping constants for lower and upper levels respectively. The common Voigt functions $H(p, q)$ and $K(p, q)$ were computed using the method due to Matta & Reichel (1971).

The angle-averaged expression can be obtained by

$$R_\nu(x', x) = 8\pi^2 \int_0^\pi R_\nu(x', x, \theta) \sin\theta \, d\theta. \quad (6)$$

The corresponding absorption profile is

$$\phi(x) = \int_{-\infty}^{\infty} R_\nu(x', x) dx = H(a_i + a_j, x). \quad (6a)$$

We have used the method employed by Heinzel (1981) to evaluate $E_\nu(x', x, \theta)$. We have also included the second-order terms in computing the E-function following Heinzel & Hubený (1983). To evaluate the angle-averaged function R_ν, we have adopted the procedure described in detail by Heinzel & Hubený (1983). The atomic frame redistribution (AFR), r_ν, has maxima at $\xi = \xi'$ and $\xi = \xi_0$ (ξ', ξ and ξ_0 being the absorption, emission and line-centre frequencies in the atomic frame). The underlying physics is discussed for example by Mihalas (1978). We have plotted $R_\nu(x', x)/\phi(x')$ in Fig. 2. From this figure we see that a photon, when absorbed in the wings, has a high probability of being emitted in the wing as well as at the centre. The wing emission is like that of R_{II} function and the emission at the centre is like that of R_{III}.

3. Basic equations and the computational procedure

The equation of transfer for a two-level atom with plane-parallel geometry is given by,

$$\mu \frac{d}{dz} I(x, \mu, z) = K_L(z)[\beta + \phi(x)][S(x, z) - I(x, \mu, z)] \quad (7)$$

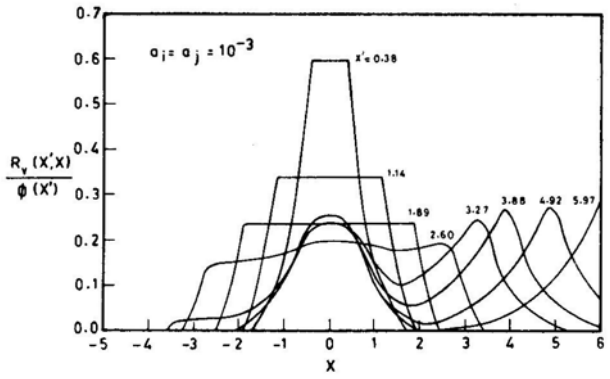


Figure 2. The probability $R_\nu(x', x)/\phi(x')$ for different x' for $a_i = a_j = 10^{-3}$.

and for the oppositely directed beam

$$-\mu\frac{\mathrm{d}}{\mathrm{d}z}I(x, -\mu, z) = K_L(z)[\beta + \phi(x)][S(x, z) - I(x, -\mu, z)] \tag{8}$$

where I (x, μ, z) is the specific intensity at angle $\cos^{-1}\mu$ $[\mu \in (0, 1)]$ at the geometrical point z, and frequency $x = (\nu - \nu_0)/\Delta s$, Δs being some standard frequency interval. The source function S (x, z) is given by

$$S(x, z) = \frac{\phi(x)S_L(x, z) + \beta S_C}{\phi(x) + \beta}, \tag{9}$$

where S_L and S_c refer to the source function in the line and continuum respectively The line source function is given by

$$S_L(x, z) = \frac{(1-\varepsilon)}{\phi(x)}\int_{-\infty}^{\infty} R(x', x)\, J\,(x')\,\mathrm{d}x' + \varepsilon B \tag{10}$$

where ε is the probability per scatter that a photon is destroyed by collisional de-excitation. B is the Planck function. We have set $S_c = B = 1$. β is the ratio of continuous opacity per Doppler width to the line opacity.

We have solved the above equations within the framework of discrete-space-theory technique (Grant & Peraiah 1972). The complete procedure with the computer code is given by Peraiah (1978). Gaussian quadrature points were used for frequency and angular mesh. 24 frequency points and two angles were chosen. Since the equations admit a solution which is symmetric with respect to the line centre, we considered only the positive frequency grid. Then for the evaluation of the scattering integral we adopted the technique described by Adams, Hummer & Rybicki (1971).

4. Results and discussion

Fig. 3 gives the emergent intensity as a function of frequency for a purely-scattering

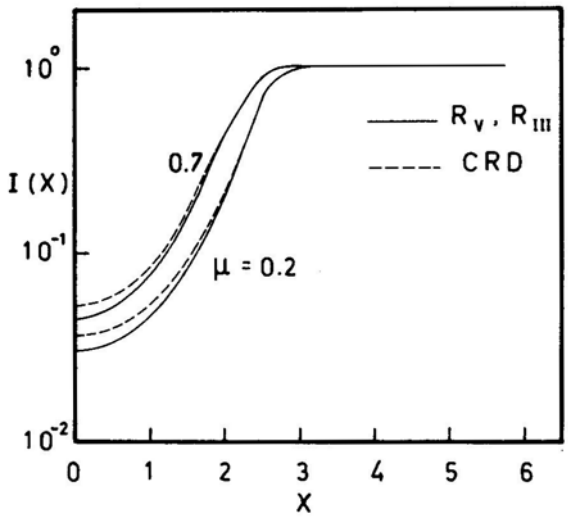


Figure 3. Emergent intensities for R_v and R_{III} are compared with CRD at μ = 0.2 and 0.7 for the case $\varepsilon = \beta = 0$.

atmosphere. The CRD case with Voigt absorption profile and damping parameter $a = 2 \times 10^{-3}$ is also plotted for comparison purposes. Boundary conditions considered are

$$I(x, \mu, \tau = T) = 1; \quad I(x, -\mu, \tau = 0) = 0. \tag{11}$$

The total optical depth considered throughout was T = 156. Since the wings are optically thin, the photons escape in the wings freely and the emergent intensity is nearly the same as the incident intensity. The intensity profiles due to R_{III} and R_v are nearly the same. The source function is plotted as a function of frequency at various optical depths for a purely-scattering medium in Fig. 4. We see that the emergent source function differs from CRD by an order of magnitude in the wings for R_{II} and R_V .For a purely-scattering medium, there is a substantial contribution from radiation in the wings to the scattering integral. This contribution is enhanced by the fact that R_{II} and R_v are coherent in the wings. Thus the R_{II} and R_v emergent source functions are higher in the wings compared to CRD. R_{II} source function lies higher than R_V, since R_{II} is much more coherent in the wings as seen from Fig. 4(a). Similar result was also obtained by Heinzel (1983) for the case of optically-thin solar prominences.

Deeper in the medium, the radiation in the wings does not differ very much from the core. This is because the incident radiation has not undergone much of absorption in the core. Now we see that the differences between the source function values in the wings are reduced for R_{II} , R_v and CRD, and also that they do not deviate very much from the line centre. This can be seen from Fig. 4(b).

Further, we have considered an atmosphere with constant thermal sources ($\varepsilon = 10^{-3}$, $\beta = 0$, $B = 1$) and with no incident radiation. The source function corresponding to the functions R_{II}, R_{III} and R_v at various optical depths are shown in Fig. 5. Typical ratios are given in Table 1. The absorption in the wings is quite small and the bulk of the absorption takes place in the core. The partial coherency impedes the

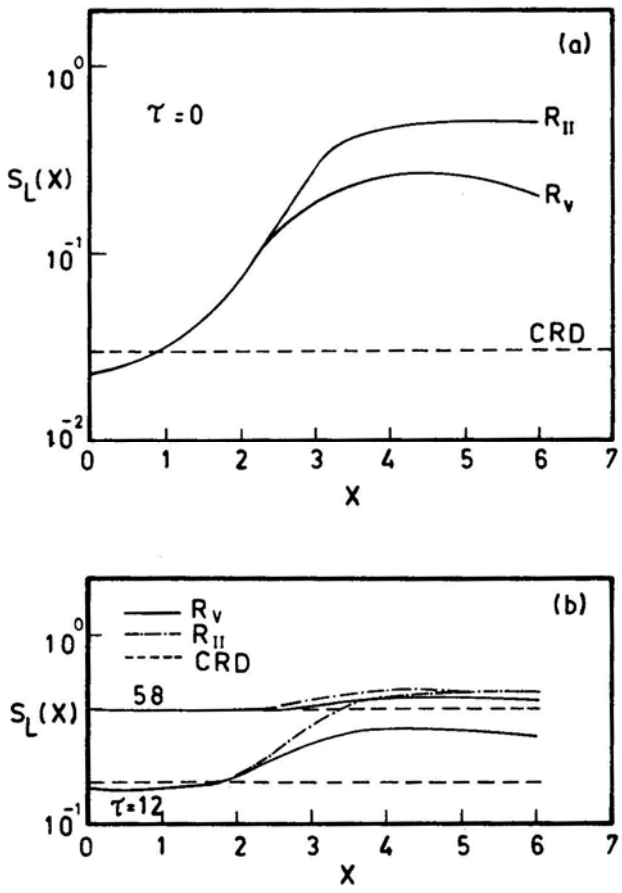


Figure 4. Source functions for R_{II}, R_V and CRD are compared for the case $\varepsilon = \beta = 0$; (a) $\tau = 0$, (b) $\tau = 12$ and 58.

Table 1. Ratios of emergent source functions at different frequencies.

x	$S_L(R_V)/S_L(R_{II})$	$S_L(R_V)/S_L(R_{III})$
0.38	0.99	1.02
1.89	0.98	1.03
3.89	1.80	0.72
5.97	8.18	0.68

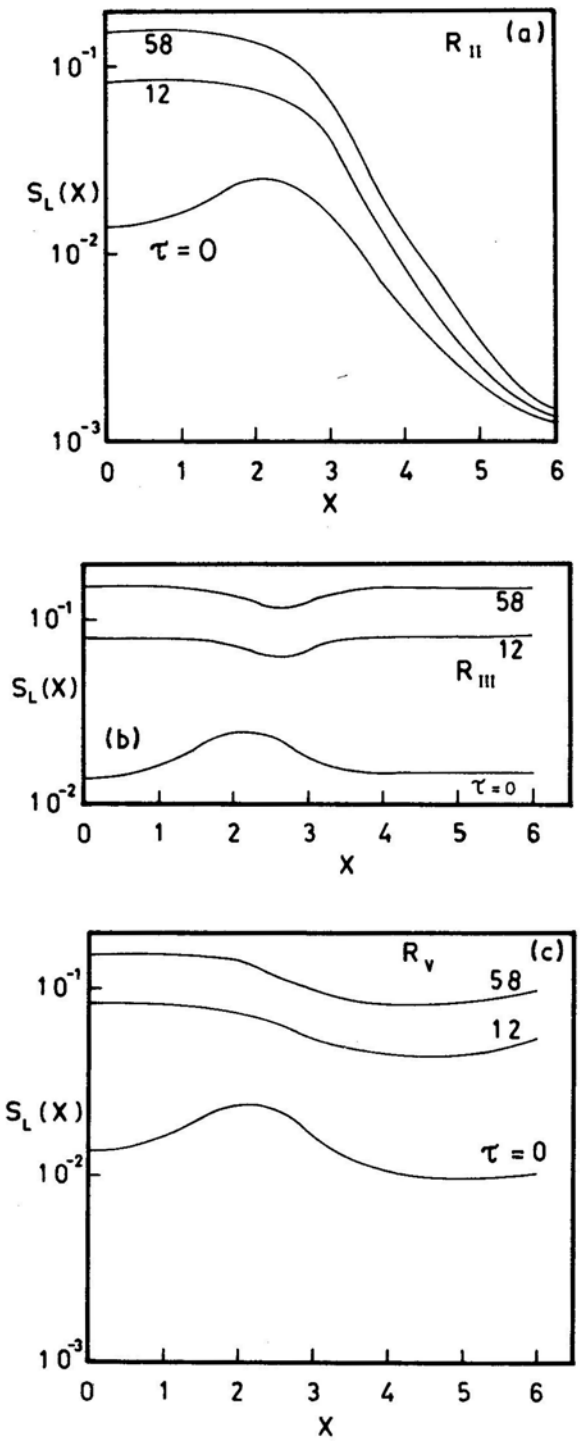


Figure 5. Source functions are plotted against x for the case $\varepsilon = 10^{-3}$, $\beta = 0$ for (a) R_{II} with $a_j = 2 \times 10^{-3}$, (b) R_{III} with $a_j = 2 \times 10^{-3}$, and (c) R_v with $a_i = a_j = 10^{-3}$.

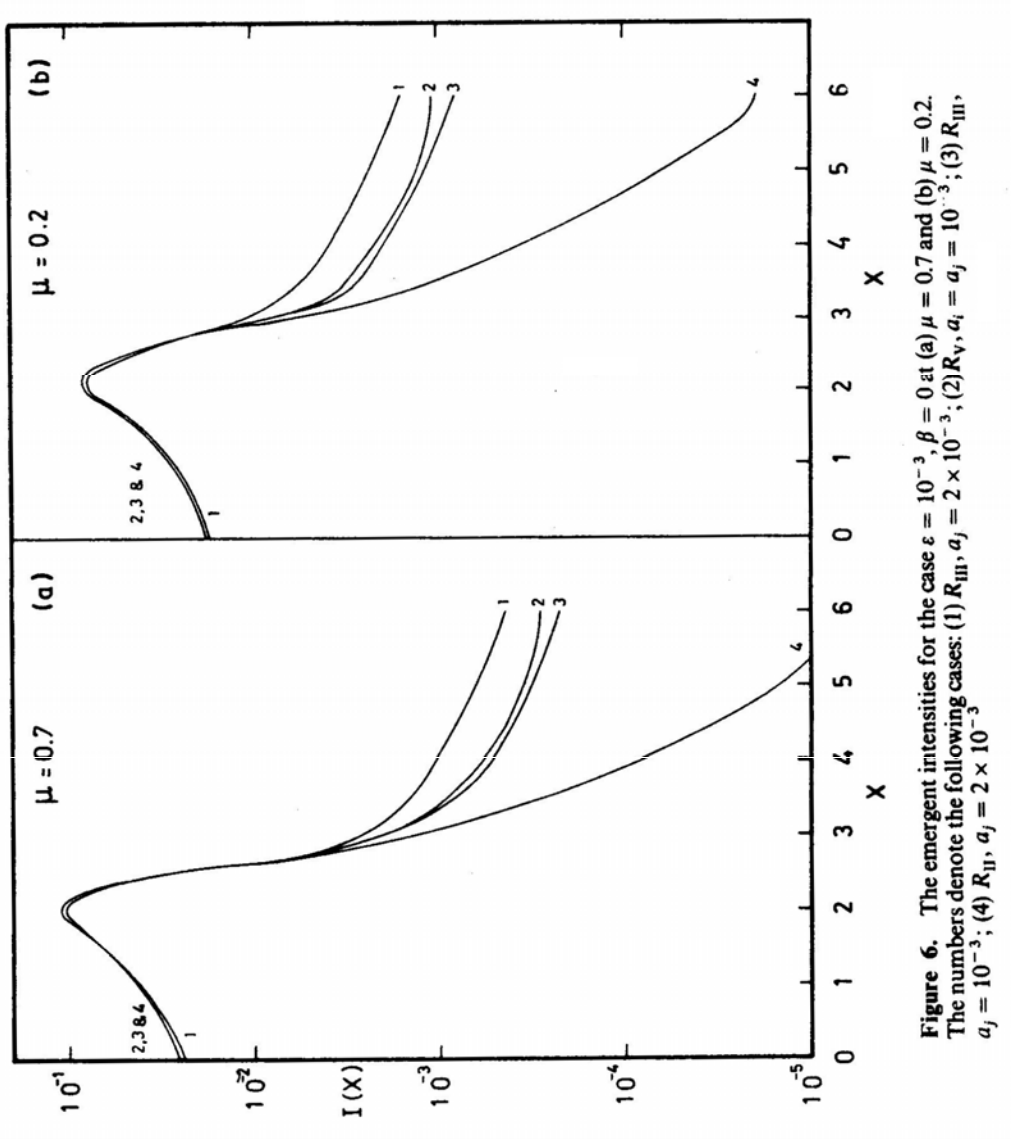


Figure 6. The emergent intensities for the case $\varepsilon = 10^{-3}$, $\beta = 0$ at (a) $\mu = 0.7$ and (b) $\mu = 0.2$. The numbers denote the following cases: (1) R_{III}, $a_j = 2 \times 10^{-3}$; (2) R_V, $a_i = a_j = 10^{-3}$; (3) R_{III}, $a_j = 10^{-3}$; (4) R_{II}, $a_j = 2 \times 10^{-3}$

escape of core photons through the wings. Therefore the efficiency of the transfer of photons to the wings depends on the noncoherency of the redistribution mechanism. R_{III}, being more noncoherent, transfers more photons to the wings.

This trend is very well exhibited by the source functions plotted. The result for R_{II} is in qualitative agreement with that of Hummer (1969) and for R_{III} with that of Vardavas (1976b).

The emergent intensity profiles for the above cases are plotted in Fig. 6. They reflect the behaviour of the source function. Similar emergent profiles have been obtained also by Hubený & Heinzel (1984), but for $T = 10^4$ and $\varepsilon = 10^{-4}$.

To see the effect of continuous absorption on line transfer with R_v redistribution, we considered a case with $\varepsilon = \beta = 10^{-3}$ and $S_c = B = 1$. We have plotted the frequency-dependent source function at various optical depths in Fig. 7(a). The source function

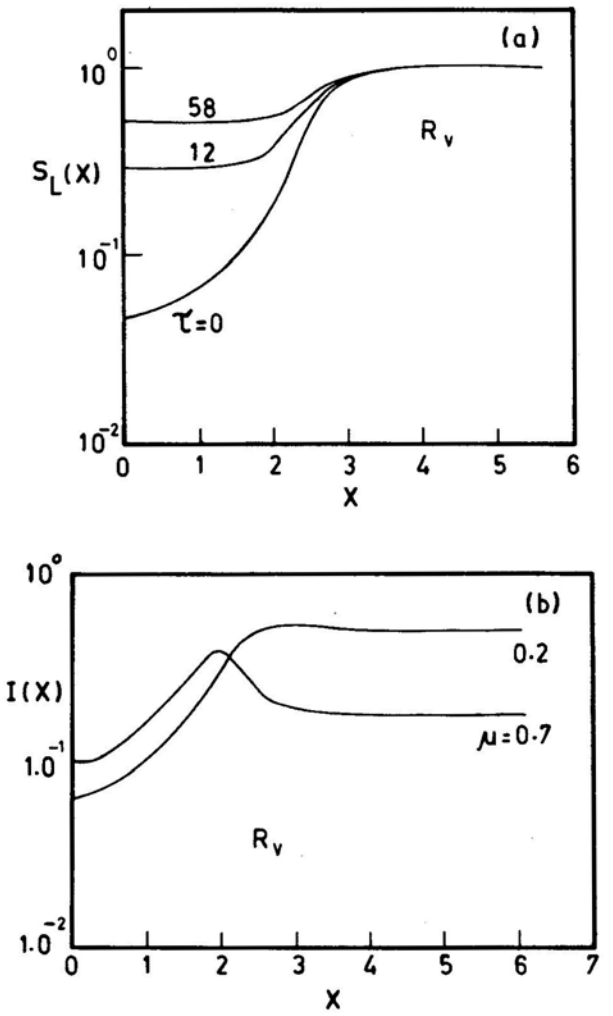


Figure 7. (a) Source function for τ = 0, 12 and 58, and (b) emergent intensity at μ = 0.7 and 0.2, for R_v with $\varepsilon = \beta = 10^{-3}$.

can be defined as

$$S(x) = \frac{1-\xi(x)}{\phi(x)} \int_{-\infty}^{\infty} R(x', x)\, J\,(x')\, \mathrm{d}x' + \xi(x) B, \tag{12}$$

where

$$\xi(x) = \frac{\beta + \varepsilon\phi(x)}{\beta + \phi(x)}. \tag{13}$$

In the far wings $\xi\,(x) \simeq 1$ and therefore, $S(x) \to B$ at all optical depths. In the wings the intensity can be approximated by $I\,(x,\ \mu) \simeq B\beta T/\mu$. These characteristics are reflected in Fig. 7 both in the source functions as well as in the emergent intensity profiles. We see from the figure that in the wings, the line transfer is dominated by the overlying continuum.

5. Conclusions

For a purely scattering medium with frequency-independent incident radiation (white light), we find that S_L (R_v) lies above S_L (R_{III}), since R_v is more coherent. For the same reason, $S_L(R_{II})$ lies above $S_L(R_V)$. For a medium with constant thermal sources and with no incident radiation, we find that the more noncoherent the redistribution, the higher is the source function in the wings. Emergent intensities just follow the source function variation. For the boundary conditions and the parameters considered, if we take $S_L(R_{II})$ and $S_L(R_{III})$ to correspond to two extreme situations, then $S_L(R_V)$ is in between these two, in qualitative agreement with the result of Hubeny & Heinzel (1984). Furthermore, the wings are controlled by the overlying continuum for R_v redistribution when the continuous absorption is present.

The transition from plane-parallel situation to spherical symmetry is being studied by the present authors and the results will be presented in a separate paper.

Acknowledgement

We wish to thank Dr. P. Heinzel for his useful comments and suggestions.

References

Adams, T. F., Hummer, D. G., Rybicki, G. B. 1971, *J. quant., Spectrosc. radiat. Transfer,* **11,** 1365.
Finn, G. D. 1967, *Astrophys. J.,* **147,** 1085.
Grant, I. P., Peraiah, A. 1972, *M on. Not. R. astr. Soc,* **160,** 239.
Heinzel, P. 1981, J. *quant. Spectrosc. radiat. Transfer,* **25,** 483.
Heinzel, P. 1983, *Bull. astr. Inst. Csl.,* **34,** 7.
Heinzel, P., Hubeny, I. 1982, *J. quant. Spectrosc. radiat. Transfer,* **27,** 1.
Heinzel. P., Huben & I. 1983, *J. quant. Spectrosc. radiat. Transfer,* **30,** 77.
Hubeny, L, Heinzel, P. 1984, *J. quant. Spectrosc. radiat. Transfer* (in press).
Hummer, D. G. 1962, *Mon. Not. R. astr. Soc,* **125,** 21.

Hummer, D. G. 1969, *Mon. Not. R. astr. Soc,* **145,** 95.
Matta, F., Reichel, A. 1971, *Math. Comput.,* **25,** 339.
Mihalas, D. 1978, *Stellar Atmospheres,* 2 Edn, Freeman, San Francisco.
Milkey, R. W., Mihalas, D. 1973, *Astrophys. J.*, **185,** 709.
Omont, A., Smith, E. W., Cooper, J. 1972, *Astr.* J., **175,** 185.
Peraiah, A. 1978, *Kodaikanal Obs. Bull. Ser. A,* **2,** 115.
Vardavas, I. M. 1976a, *J. quant. Spectrosc. radial Transfer,* **16,** 1.
Vardavas, I. M. 1976b, *J. quant. Spectrosc. radiat. Transfer,* **16,** 715.
Vardavas, I. M. 1976c, *J. quant. Spectrosc. radiat. Transfer,* **16,** 781.

J. Astrophys. Astr. (1984) **5,** 181–185

Spectroscopic Binaries near the North Galactic Pole
Paper 10: HR 4668

R. F. Griffin *The Observatories, Madingley Road, Cambridge, England CB3 0HA*

Received 1984 February 27; accepted 1984 March 10

Abstract. Photoelectric radial-velocity measurements confirm and refine the preliminary orbit derived for HR 4668 by Christie on the basis of eight Lick spectrograms taken more than 50 years ago.

Key words: radial velocities—spectroscopic binaries—stars, individual

HR 4668 (HD 106760) is one of the only 22 stars within the area covered by the Cambridge radial-velocity survey of the North Galactic Pole field ($b > 75°$) that are as bright as 5.0 mag (Hoffleit 1982). Very accordant determinations of its *UBV* magnitudes have been published by Argue (1963,1966), Ljunggren (1965), Häggkvist & Oja (1966), Eggen (1966, 1969, 1971) and Helfer & Sturch (1970); all their results are very close to $V = 5.00$, $(B - V) = 1.14$, $(U - B) = 1.07$. Notwithstanding that the extreme range in B magnitude of all these determinations is only 0.016 mag, Deridder, Sterken & Vanbeveren (1977) claimed, on the basis of four nights' observations, that HR 4668 shows capricious variations of up to 0.07 mag in B on short timescales. The star appears in the *Two-Micron Sky Survey* (Neugebauer & Leighton 1969) as IRC 30236, with $I = 4.16$, $K = 2.33$. Other infrared photometry, not all on the same system, has been given by Helfer & Sturch (1970), Jacobsen (1970) and Eggen (1971); there is some agreement that the Johnson $(R - I)$ colour is close to 0.57.

The spectral type of HR 4668 has been classified as K0 III: by Keenan (1940), Kl III by Roman (1952) and Schild (1973), K2 III (from objective-prism spectra) by Upgren (1962), and is given in the recent edition of the *Bright Star Catalogue* (Hoffleit 1982) as K0.5 IIIb—a type which has the appearance of a Keenan classification. Highly consistent absolute-magnitude estimates, ranging between extreme values of + 0.5 and + 1.3 mag, have been obtained spectroscopically by Rimmer (1925), Adams *et al.* (1935) and Keenan (1940), photometrically by Helfer & Sturch (1970), Hansen & Kjærgaard (1971) and Boyle & McClure (1975), and from the group parallax which follows from the membership claimed for HR 4668 in the 'Wolf 630' group by Eggen (1969). By virtue of its brightness the object also features in many papers, not cited individually here, concerned with narrow-band spectrophotometry. Taken together they show that it is a very normal star with elemental abundances probably slightly lower than those of the Sun.

Unlike the other stars treated in the present series of papers, HR 4668 has already been the subject of a published orbit. The radial velocity was observed at Lick, starting in 1918, and its variability was announced by Campbell (1922). The final results of the Lick work (Campbell & Moore 1928) contain eight velocities for HR 4668, but the authors evidently did not care to base an orbit on so few data. Christie (1936), who seems

not to have suffered from the same inhibitions, simply took the Lick observations from the literature and calculated the orbit; the propriety of this action is belatedly vindicated below, where a much more firmly based orbit is shown to have elements quite similar to those found by Christie. As a caution to those who may be disposed to share his optimism, however, it may be remarked that Christie (1936) was not uniformly successful in respect of all of the 16 orbits he derived, in the paper that includes HR 4668, from published velocities. Four of the 16 orbits (not counting HR 4668) have

Table 1. Radial-velocity measurements of HR 4668.

Date		MJD	Velocity km s^{-1}	Phase	$(O-C)$ km s^{-1}
1918 Apr	29.33*	21712.33	−38.2	0.968	+0.3
1919 Apr	29.30*	22077.30	−36.6	1.246	+0.7
1920 May	21.25*	22465.25	−44.1	1.541	−1.0
1921 May	27.21*	22836.21	−45.3	1.823	0.0
1922 Mar	8.45*	23121.45	−33.2	2.040	−0.6
1923 Mar	25.44*	23503.44	−40.5	2.331	−1.1
Dec	5.56*	758.56	−42.1	.525	+0.7
1926 May	31.25*	24666.25	−36.3	3.216	+0.2
1966 Apr	13.95	39228.95	−36.7	14.296	+1.9
1969 Mar	5.07	40285.07	−32.0	15.100	+0.9
1972 Apr	8.04	41415.04	−39.5	15.960	−0.1
Nov	23.27	644.27	−33.7	16.134	+0.2
1973 Apr	27.89†	41799.89	−36.9	16.252	+0.6
1977 Mar	31.02	43233.02	−39.4	17.343	+0.3
Apr	25.00	258.00	−39.9	.362	+0.2
May	27.90	290.90	−41.0	.387	−0.5
June	10.92	304.92	−41.1	.398	−0.4
1978 Jan	18.21	43526.21	−42.7	17.566	+0.7
Mar	24.07	591.07	−44.5	.615	−0.5
May	23.20‡	651.20	−45.2	.661	−0.6
June	18.91	677.91	−43.7	.681	+1.1
Nov	16.23	828.23	−46.3	.796	−0.9
1979 Jan	3.20	43876.20	−45.6	17.832	−0.4
Feb	25.13	929.13	−45.3	.873	−0.8
Mar	8.08	940.08	−44.3	.881	0.0
Apr	25.02	988.02	−42.4	.917	+0.3
May	13.93	44006.93	−41.1	.932	+0.7
June	22.92	046.92	−39.5	.962	−0.4
Nov	28.27	205.27	−32.6	18.083	−0.1
Dec	25.19	232.19	−33.0	.103	0.0
1980 Jan	24.24	44262.24	−34.4	18.126	−0.8
Feb	12.14	281.14	−34.1	.140	0.0
May	3.95	362.95	−37.1	.203	−1.0
1981 Feb	2.12	44637.12	−40.2	18.411	+0.8
Apr	27.96	721.96	−42.4	.476	−0.3
May	24.97	748.97	−42.7	.496	−0.3
July	4.90	789.90	−43.1	.528	−0.2
1982 Jan	10.17	44979.17	−44.8	18.672	−0.1
Mar	2.08	45030.08	−44.9	.710	+0.1
Apr	14.95	073.95	−43.9	.744	+1.3
May	5.00	094.00	−45.0	.759	+0.3
June	29.93	149.93	−45.4	.801	0.0

Table 1. Continued.

Date		MJD	Velocity km s^{-1}	Phase	$(O-C)$ km s^{-1}
1983 Jan	19.14	45353.14	−39.5	18.956	+0.2
Feb	3.50§	368.50	−38.8	.968	−0.3
	18.49§	383.49	−37.7	.979	−0.4
	28.07	393.07	−36.5	.986	+0.1
Mar	7.06	400.06	−36.0	.992	0.0
	15.03	408.03	−34.8	.998	+0.6
Apr	15.90	439.90	−34.0	19.022	−0.5
May	9.90	463.90	−32.4	.040	+0.2
June	6.90	491.90	−32.0	.062	+0.3
July	2.92	517.92	−32.3	.081	+0.2
Dec	11.20	679.20	−36.0	.204	+0.1
1984 Jan	2.22	45701.22	−36.8	19.221	−0.2
Feb	9.10	739.10	−38.2	.250	−0.8

* Lick photographic observation (Campbell & Moore 1928).
† Observed by Dr G. A. Radford with the Cambridge telescope.
‡ Observed, in collaboration with Dr J. E. Gunn, with the 200-inch telescope (Griffin & Gunn 1974).
§ Observed with the Dominion Astrophysical Observatory 48-inch telescope (Fletcher *et al.* 1982).

subsequently been redetermined with the aid of additional radial-velocity data: two (those of τ Per and 31 Cyg) have been broadly confirmed, one (η Gem) has been shown to be entirely mistaken, and one (δ Sge) was seriously vitiated by an error in one of the published lists of velocities upon which Christie relied.

A minor enigma surrounds three Mount Wilson radial velocities which Christie claimed in his paper to exist for HR 4668. Christie worked at Mount Wilson himself, so he was in a position to know what velocities were available; but he neither published the three measurements to which he referred, nor did he plot them on his orbit. They are not traceable through the *Bibliography of Stellar Radial Velocities* (Abt & Biggs 1972), and their absence from Abt's (1973) catalogue of Mount Wilson radial velocities implies that they do not even appear in the original card files at the Mount Wilson Observatory.

The radial velocity of HR 4668 was first measured from Cambridge in 1966, right at the start of the photoelectric work: it was one of the bright stars observed in the initial tests of the system, but was omitted from the discussion of the test results (Griffin 1967) not only on account of its known variability but because the discussion was in any case restricted to those objects for which at least three observations had been made. A few velocities were subsequently measured in the course of routine work on the Galactic Pole field; not until 1977 did the present author realize the very preliminary nature of the published orbit and institute a systematic programme of measurements of HR 4668. It is a helpful circumstance that the orbital period is close to a half-integral number of years: the inevitable seasonal gaps in phase coverage of one cycle of the orbit are automatically made good in the following cycle, so now that two periods have elapsed since systematic observations began the orbit is ripe for discussion. The total number of photoelectric radial velocities is 47; they are listed in Table 1 together with the eight Lick velocities, to which an adjustment of + 0.8 km s^{-1} has been made in the interest of homogeneity (Griffin & Herbig 1981).

A preliminary solution using the photoelectric measurements alone showed the

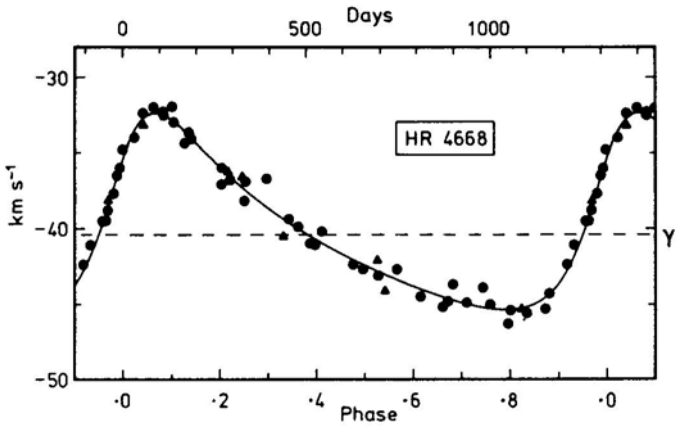


Figure 1. The computed radial-velocity curve for HR 4668, with the measured radial velocities plotted. Photoelectric observations are represented by circles; the early Lick photographic velocities are plotted as triangles.

Table 2. Orbital elements for HR 4668.

Element	Christie (1936)	This paper
P(days)	1300	1314.3 ± 0.4
γ (km s^{-1})	-40.7*	-40.39 ± 0.09
K (km s^{-1})	6.8	6.54 ± 0.13
e	0.3	0.426 ± 0.016
ω (degrees)	300	303.6 ± 2.8
T (MJD)†	21749.5	41468 ± 7
$a_1 \sin i$ (Gm)	116	107.0 ± 2.3
$f(m)$ ($M_\odot$)	0.0368	0.0283 ± 0.0018

* Adjusted by + 0.8 km s^{-1} (Griffin & Herbig 1981) for direct comparability with the velocity found in this paper.

† The two epochs refer, of course, to different periastron passages. The epoch and period found here, extrapolated back to Christie's epoch, give MJD 21753 ± 9

period to be l311 ± 3 days. The Lick velocities were then included, with the same weight as the others, in a final solution which refined the period to 1314.3 ± 0.4 days. The elements are listed in Table 2, with those of Christie (1936) added for comparison. It will be seen that Christie's elements provided a very fair approximation to those derived now; on the other hand the sum of the squares of the 55 residuals is reduced from 1333 to 17.9 (km s^{-1})^2 by the new elements, which therefore clearly furnish a much more accurate representation of the true orbit than those they supersede.

References

Abt, H. A. 1973, *Astrophys. J. Suppl.,* **26,** 365.

Abt, H. A., Biggs, E. S. 1972, *Bibliography of Stellar Radial Velocities,* Kitt Peak National Observatory, Tucson.

Adams, W. S., Joy, A. H., Humason, M. L., Brayton, A. M. 1935, *Astrophys. J.,* **81,** 187.
Argue, A. N. 1963, *Mon. Not. R. astr. Soc,* **125,** 557.
Argue, A. N. 1966, *Mon. Not. R. astr. Soc,* **133,** 475.
Boyle, R. J., McClure, R. D. 1975, *Publ. astr. Soc. Pacific,* **87,** 17.
Campbell, W. W. 1922, *Publ. astr. Soc. Pacific,* **34,** 167.
Campbell, W. W., Moore, J. H. 1928, *Publ. Lick Obs.,* **16,** 181.
Christie, W. H. 1936, *Astrophys. J.,* **83,** 433.
Deridder, G., Sterken, C, Vanbeveren, D. 1977, *Inf. Bull. Var. Stars,* No. 1295.
Eggen, O. J. 1966, *R. Obs. Bull.,* No. 125.
Eggen, O. J. 1969·, *Publ. astr. Soc, Pacific,* **81,** 553.
Eggen, O. J. 1971, *Astrophys. J.,* **165,** 317.
Fletcher, J. M., Harris, H. G, McClure, R. D., Scarfe, C. D. 1982, *Publ. astr. Soc. Pacific,* **94,** 1017.
Griffin, R. F. 1967, *Astrophys. J.,* **148,** 465.
Griffin, R. F., Gunn, J. E. 1974, *Astrophys. J.,* **191,** 545.
Griffin, R. F., Herbig, G. H. 1981, *Mon. Not. R. astr. Soc,* **196,** 33.
Häggkvist, L., Oja, T. 1966, *Ark. Astr.,* **4,** 137.
Hansen, L., Kjærgaard, P. 1·971, *Astr. Astrophys.,* **15,** 123.
Helfer, H. L., Sturch, C. 1970, *Astr. J.,* **75,** 971.
Hoffleit, D. 1982, *Bright Star Catalogue,* Yale University Observatory, New Haven.
Jacobsen, P. -U. 1970, *Astr. Astrophys.,* **4,** 302.
Keenan, P. C. 1940, *Astrophys. J.,* **91,** 506.
Ljunggren, B. 1965. *Ark. Astr.,* **3,** 535.
Neugebauer, G., Leighton, R. B. 1969, *Two-Micron Sky Survey* (NASA SP–3047), NASA, Washington, D.C.
Rimmer, W. B. 1925, *Mem. R. astr. Soc,* **64,** 1.
Roman, N. G. 1952, *Astrophys. J.,* **116,** 122.
Schild, R. E. 1973, *Astr. J.,* **78,** 37.
Upgren, A. R. 1962, *Astr. J.,* **67,** 37.

J. Astrophys. Astr. (1984) **5,** 187–207

Millisecond Pulsars

D. C. Backer *Radio Astronomy Laboratory and Astronomy Department, University of California, Berkeley CA 94720, USA*

(*Invited article*)

Abstract. In 1982 we discovered a pulsar with the phenomenal rotation rate of 642 Hz, 20 times faster than the spin rate of the Crab pulsar. The absence of supernova debris in the vicinity of the pulsar at any wavelength indicates an age of the neutron star greater than 10^5 yr. The miniscule spindown rate of 1.1×10^{-19} confirms the old age and indicates a surface magnetic field of 10^9 G. A second millisecond pulsar was discovered by Boriakoff, Buccheri & Fauci (1983) in a 120-day orbit. These fast pulsars may have been spun-up by mass transfer in a close binary evolutionary stage. Arrival-time observations of the 642-Hz pulsar display remarkably low residuals over the first 14 months. The stability implied by these observations, 3×10^{-14}, suggests that millisecond pulsars will provide the most accurate basis for terrestrial dynamical time. If so, the pulsar data will lead to improvements in the planetary ephemeris and to new searches for light-year scale gravitational waves. Many new searches for fast pulsars are under way since previous sky surveys excluded pulsars with spins above 60 Hz.

Key words: pulsars: 1937 + 21, 1953 + 29—pulsar surveys—time—gravitational waves

1. Introduction

"If the neutron star hypothesis of the origin of Supernovae can be proved, it will be possible to subject the general theory of relativity to tests which according to the considerations presented in this paper deal with effects which in order of magnitude are large compared with the tests so far available. The general theory of relativity, although profound and exceedingly satisfactory in its epistemological aspects, has so far practically not lent itself to any very obvious and generally impressive applications. This unfortunate discrepancy between the formal beauty of the general theory of relativity and the meagerness of its practical applications make it particularly desirable to search for phenomena which cannot be understood without the help of the general theory of relativity."

F. Zwicky (1939)

These prescient comments of Fritz Zwicky are remarkable in the light of recent discoveries of degenerate neutron stars in a heretofore unimagined variety of configurations. Two key developments opened the door to neutron-star investigations. Exploration of the sub-second domain of radio source variability led to the discovery of

a 1.377-s pulsar (Hewish *et al.* 1968) which was subsequently identified as a neutron star (Gold 1968). The opening of the X-ray spectrum with rocket flights and satellite launches led to the detection of neutron stars accreting mass in close binary systems (Giacconi *et al.* 1971). Shortly after the pulsar discovery a 33-ms pulsar was discovered coincident with the peculiar 16-mag star noted by Baade (1942) and Minkowski (1942) in the centre of the Crab nebula supernova remnant. Seven years later Hulse & Taylor (1975) discovered a pulsar orbiting a second neutron star. The slow decay of the orbital period of this system has provided the first evidence for gravitational radiation predicted in the general theory of relativity; Zwicky's dream has been realized.

In 1982 the observable range of stellar rotation was extended by a factor of 20 with the discovery of a 1.558-ms pulsar in Vulpecula, not far from the location of the first pulsar (Backer *et al.* 1982b). Observations subsequent to the discovery soon resolved an apparent discrepancy between the pulsar's rapid spin, which indicated a youthful object, and the absence of supernova debris at any wavelength, which indicated a minimum age of 10^5 yr. The pulsar's spin, $\Omega = 2\pi/P$, was decaying extremely slowly, $\Omega/\dot{\Omega} = P/\dot{P} \sim 5 \times 10^8$ yr. This new member of the cosmic menagerie placed the pre-discovery speculations of Radhakrishnan (1982) and others concerning anomalous pulsars in a P–$\dot{P}$ diagram on firmer ground. Radhakrishnan had identified moderately fast pulsars with low spin-decay rates as neutron stars that had been 'recycled' to their present fast spins by angular momentum transfer from an evolving secondary.

In the following pages I will recount the path which led to the discovery and then summarize three areas of inquiry which have been stimulated by this discovery: (1) further observations and investigations of neutron star astrophysics; (2) pulse arrival-time measurements and gravitational wave physics; (3) search for new fast pulsars that were excluded in past surveys of the galaxy owing to computational limitations and to a prejudice concerning the period distribution. Scenarios for the origin of the fast pulsars are discussed by van den Heuvel (1984) elsewhere in this issue.

2. Discovery—persistence pays off

In 1977–78 Stuart Vogel and I were investigating with VLBI the radio sources in the Cygnus region of the sky to assess the prevalence of interstellar scattering at low galactic latitudes. The source 4C 21.53 came to our attention since it displayed strong interplanetary scintillations (IPS) despite its low galactic latitude (Duffett-Smith & Readhead 1976). Pulsar observations had indicated that at low galactic latitudes interstellar scattering (ISS) would suppress the IPS modulation. The steep intensity spectrum of the object added to its peculiar nature. While identification of 4C 21.53 with a pulsar would have explained its peculiar properties, no known pulsar was within the errors of the 4C 21.53 position.

While searching the literature for references to 4C 21.53 in January 1979, I found a source, 1937 + 215, located 30 s west of the 4C 21.53 in several published catalogues. This source could be identified with 4C 21.53 if the 4C position was in error by one lobe; lobe errors in the 4C Catalogue occur with a frequency of about 3 per cent(Backer *et al.* 1970). The difficulty with connecting 1937 + 215 and the IPS source was that the spectra were very different. Furthermore the large size, 60 arcsec, for 1937 + 215 in the 5-GHz catalogue of Altenhoff *et al.* (1979) confounded the lobe-error hypothesis since the IPS object was necessarily smaller than 1 arcsec. My initial synthesis of these observations

was that the steep-spectrum, IPS object was a pulsar co-located with a faint (1 Jy), extended supernova remnant, 1937 + 215. Curiously no pulsar had been detected in this region in the very sensitive Arecibo survey (Hulse & Taylor 1974). While interstellar scattering could have smeared the pulsation of a pulsar so that it was not detectable in the 430-MHz survey, the IPS observations at 81 MHz discussed above had already indicated that ISS was not severe for this object. A simple calculation demonstrated that ISS could only smear pulsations for periods of order 10 ms or less. The Arecibo survey, and most others, were not sensitive to periods below 60 ms. A report on the hypothesis that 1937 + 215 was a young pulsar-supernova pair similar to the Crab nebula and its pulsar—published prior to the discovery of pulsars (Hewish & Okoye 1965)—was returned from a journal with the referee's comment as 'too speculative'.

Several new pieces were added to the puzzle early in 1979. Mike Davis at the Arecibo Observatory measured the spectrum of 1937 + 215 between 430 MHz and 2.4 GHz and conducted pulse searches. The Arecibo pulse searches were doomed to failure owing to the imprecision of the position at this time. Rickard & Cronyn (1979) briefly discussed the peculiar properties of 4C 21.53 in a paper which proposed a new class of compact, low-latitude objects which they called 'scintars'. David Cudaback, Stuart Vogel, John Middleditch and I conducted a pulsar search with the 90-foot antennas at Owens Valley Radio Observatory in 1979 March at 600 MHz. These data had a Nyquist frequency of only 50 Hz and were corrupted by extensive interference. I then learned about the independent investigations by Erickson (1979, personal communication) with the VLA in September 1978 which showed the presence of a second source 30s east of 4C 21.53. The steep spectrum and compactness of the eastern component suggested that it was the IPS object. The pulsar-supernova hypothesis now seemed unwarranted.

In 1981 my interest in the western source, 1937 + 215, was rekindled by a report from Erickson (1980) that 34-MHz data from the Clark Lake Radio Observatory indicated IPS sources at the positions of both the eastern and the western components of 4C 21.53. The pulsar-supernova-remnant hypothesis was resurrected. I recalled a discrepancy between the catalogued positions of 1937 + 215 at cm wavelengths and those at metre wavelengths, *viz.,* the 365-MHz position of Douglas *et al.* (1980) and the 81-MHz position of Slee (1977). I reasoned that these measurements indicated a southward shift of the steep spectrum IPS source from the extended source. In 1982 March, Miller Goss and I obtained a 610-MHz image of 1937 + 215 with the Westerbork Synthesis Radio Telescope which confirmed the suspected offset. This was the first hard evidence for the location of the IPS source in the western component. The intensity, 0.13 Jy, confirmed the steep spectrum seen at decametric wavelengths. A brief VLA measurement in 1982 May further isolated the IPS object from the extended source to the north. When the Westerbork data became available, Val Boriakoff at the Arecibo Observatory conducted a pulse search which was sensitive to periods as short as 4 ms. Again no pulsar was detected.

In 1982 August, lively discussions at the Patras IAU meeting about the riddle of the western components of 4C 21.53 gave further investigations high priority. I still suspected a fast pulsar was hidden in 1937 + 215. In Berkeley Shri Kulkarni and I planned a pulse search experiment at the Arecibo Observatory that would be sensitive to a wide range of pulsar periods and dispersion measures. The report from Goss that the compact source in 1937 + 215 was highly polarized at 610 MHz strengthened the fast pulsar hypothesis. In late September, Kulkarni and Mike Davis recorded data at

1400 MHz with 0.5 ms sampling. The average 1024-point FFT from a portion of this data is shown in Fig. 1. The peaks in channels 329 and 347 are the fundamental and the aliased second harmonic of the pulsar. The signal was detected only in 'on source' observations which suggested a celestial origin rather than terrestrial interference. However the signal was present only for a portion of the observation and was undetected a few days later at both 1400 or 2300 MHz. We were left with no firm conclusion about the nature of the signal.

In November a team of radio astronomers—Kulkarni, Davis, Carl Heiles and myself—conducted a series of experiments at Arecibo on the candidate pulsar. My list of proposed observations included a 1400-MHz search for ISS modulation of the object, which would indicate an extremely small angular size, and a period analysis down to periods of 100 μs. Our first observations immediately showed ISS modulation. I then realized that the intermittent nature of the signal seen in September was a result of ISS. The next observing run confirmed the 1.558-ms periodicity. By the fourth night of observations we devised a signal averager display of the pulse waveform which showed the peaks of the fastest pulsar rising out of the noise after 10 seconds of integration. This dramatic display was enhanced by the appearance of two peaks in each period, a pulse and an interpulse, spaced by nearly a half period in striking similarity to the Crab pulsar (Fig. 2).

An apparent period change between the September and November observations led to the announcement of a large apparent spindown in the discovery telegram (Backer *et al.* 1982a). The change was later traced to a sampling error in September and the spindown was revised (Backer, Kulkarni & Heiles 1982). During November we also measured a recombination line from the extended object north of the pulsar and the dispersion measure of the pulsar. The recombination line detection and its velocity implied that the northern source was probably an HII region, either in the solar vicinity

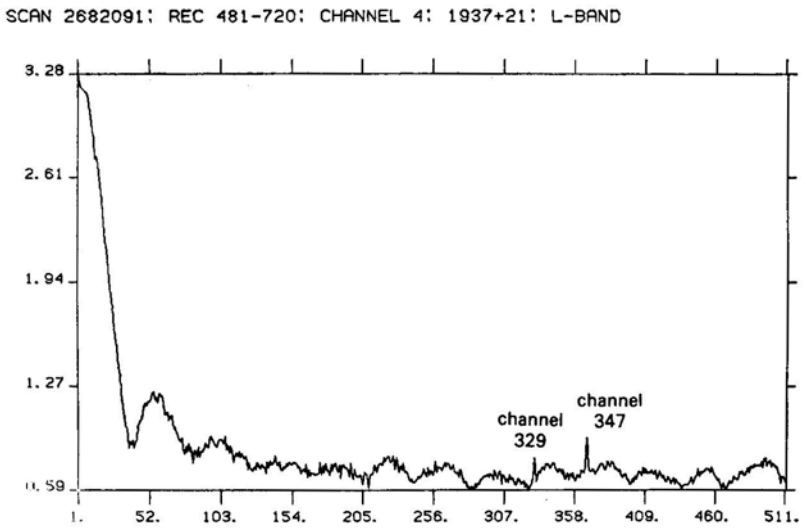


Figure 1. FFT power spectrum of PSR 1937 + 21 from the observations at Arecibo Observatory on 1982 September 28. The peaks in channels 329 and 347 are from the fundamental (642 Hz) and the alias of the second harmonic (1284 Hz) about the Nyquist frequency (1000 Hz). The undulations in the spectrum are from a sampling error.

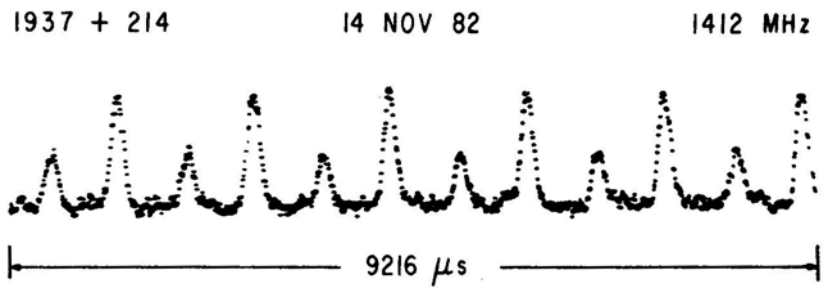


Figure 2. Signal averager oscilloscope display of PSR 1937 + 21 recorded at the Arecibo Observatory. Six periods are displayed with a sample spacing of 6 μs. Each period contains a main pulse followed by an interpulse nearly half a period later. The pulse width is dominated by instrumental effects (Backer *et al.* 1982b; © *Nature*).

or on the other side of the galaxy. The dispersion measure indicated a distance of 2.5 kpc for the pulsar. Despite the distance discrepancy we suggested a possible relation between the pulsar and the HII region owing to their close proximity on the sky (Backer *et al.* 1982b).

Assessment of the distance to PSR 1937 + 21 was particularly important to issues other than the potential relation of the pulsar to the HII region: detection of high-frequency radiation and the direct detection of gravitational waves. The initial distance estimate of 2.5 kpc cast doubt on the pulsar–HII region association since the radial velocity of the HII region indicated its distance was either near 0 kpc or near 10 kpc. Subsequent measurements of the HI absorption spectra for the pulsar and the HII region confirmed the non-association hypothesis (Heiles *et al.* 1983). Heiles *et al.* developed a novel phase-switching scheme for autocorrelation spectroscopy of fast pulsars. The spectra for the HII region clearly favoured the distance of 10 kpc, while the spectra for the pulsar indicated a distance of 5 kpc. The mean electron density resulting from the greater distance, 0.015 e cm^{-3}, is consistent with previous measurements (Weisberg 1978) for nearby paths through the interarm region of the galaxy as discussed by Ables & Manchester (1976).

Early in 1983 a second fast pulsar, PSR 1953 + 29, with a period of 6.133 was discovered by Boriakoff and his colleagues (Boriakoff, Buccheri & Fauci 1983). The 'pointer' to this pulsar was one of the unidentified point sources in the COS B survey (Swanenburg *et al.* 1981). Their search observations were distributed over three years owing to the large uncertainties of the gamma-ray source positions and the small size of the radio antenna beam. Association of the 6.1-ms pulsar with 2CG 065 + 00 is uncertain for the same reason. However Boriakoff *et al.* point out that the present limit on the spindown rate, $\dot{P} \sim 5.8 \times 10^{-16}$ s s^{-1}, leads to a rotational energy loss rate comparable to that of the gamma-ray source luminosity. The remarkable property of PSR 1953+29 is that it is in a binary orbit with an invisible companion. The orbit is nearly circular with a period of 120 days. Several authors conclude that the companion is a 0.3 $M_\odot$ white dwarf (Joss & Rappaport 1983; Paczyński 1983; Savonije 1983; Helfand, Ruderman & Shaham 1983). The distance of this system, 7 kpc using the mean electron density discussed above, suggests that detection of the companion will be difficult. Unlike PSR 1937 + 21, this pulsar emits a broad beam of radiation filling nearly a third of the period (Fig. 3).

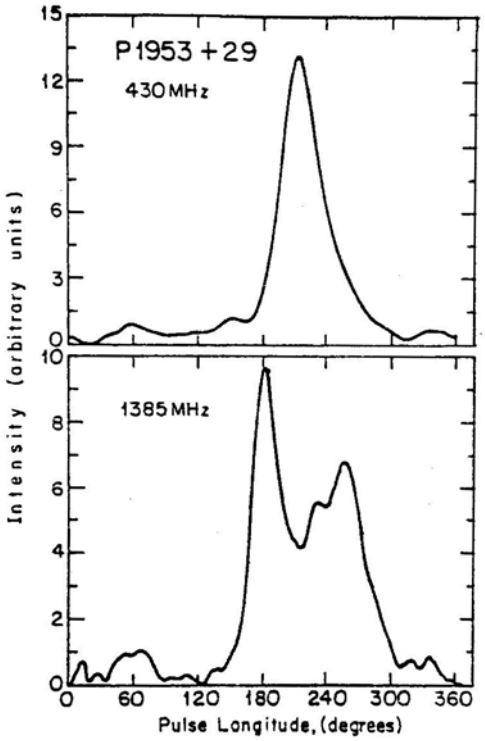


Figure 3. Average pulse profiles of PSR 1953 + 29 at 430 MHz and 1385 MHz from observations by Boriakoff, Buccheri & Fauci (1983) at the Arecibo Observatory. Full scale in pulse longitude corresponds to the period of 6.1 ms (© *Nature*).

3. Further observations—astrophysical implications

3.1 *Beaming*

The sharpness of the radio pulses in most pulsars and their systematic linear polarization variations within the pulses are commonly explained by a hollow cone of emission centred on a magnetic dipole axis and inclined to the rotation axis (Radhakrishnan &Cooke 1969). The axis of a dipole is the likely site for radio emission owing to the plausibility of particle acceleration along field lines which connect to the interstellar medium beyond the pulsar. The hollow-cone model is purely kinematic since the details of particle acceleration and radio photon generation are certainly more complex (Arons 1981). The magnetosphere is treated as closed and corotating for dipolar field lines with maximum radii less than the light-cylinder radius. Backer (1976) and, more recently, Rankin (1983) have shown that many pulsars emit a narrow pencil beam near the centre of the hollow cone. An observer sees one narrow pulse peak when the cone grazes our line of sight, and two or three peaks when the central region of the cone passes through our line of sight. If the pulsar dipole axis is nearly perpendicular to the rotation axis the observer may see pulses coming from both poles spaced by half a period. The width of the Double and Triple pulses in slow pulsars, 15°, is much larger than the dimension of the open field lines at the surface of a neutron star for a

magnetosphere which closes at the light cylinder, 1.7–2.5° $P^{-0.5}$. The breadth of these pulses in dipole-axis models is attributed to both high-altitude emission and to closed magnetosphere dimensions significantly smaller than the light-cylinder radius.

The millisecond pulsars provide dramatic evidence that, despite their fast rotation speed, they can produce beamed radio emission which in many ways resembles that of their slower cousins. I propose below a morphological scheme which ties the new pulsars closely to the slower objects by a simple period scaling. The physical significance of this isomorphism will be the subject of future work.

The angular widths of pulsars that are clearly identifiable as Double, or Triple, display a dependence on period of 15° $P^{-0.35}$ (Fig. 4). PSR 1953 + 29 is included since at 1400 MHz its profile is Triple (Fig. 3). The agreement with a power law is remarkable since any randomness of the dipole axis-spin axis obliquity would produce scatter. The angular widths of Single pulsars follow a parallel power law of 5° $P^{-0.35}$ (Fig. 4). The precursor component of the Crab pulsar and the Vela radio pulse are in the Single category. Pulsars classified as Single are probably a mixture of those where the observer grazes the edge of the hollow cone, and those where the central beam is dominant (Rankin 1983). The Double/Triple line plotted in Fig. 4 gives the opening angle of a dipole field at a period dependent radius of 1500 km $P^{+0.3}$ for a magnetosphere which extends to the light cylinder. This period scaling was first noted by Roberts & Sturrock (1972) in a much more limited data set. Both Roberts & Sturrock and, more recently,

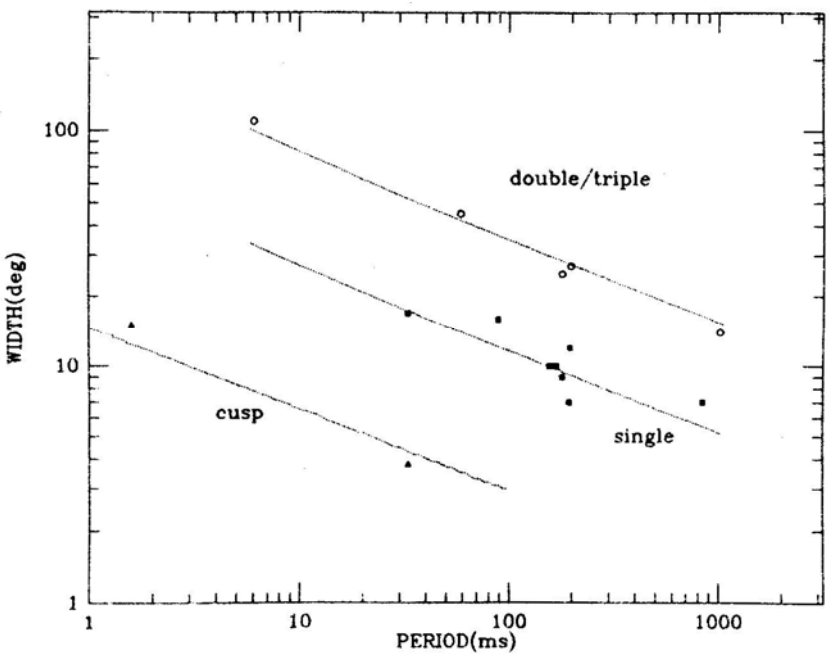


Figure 4. Pulse width-period relation for three types of pulse morphology. Double and Triple pulses have two and three distinct peaks in the average profile(O);Single pulses have only one peak (■). Cusp denotes the sharp main pulse and interpulse in the average profiles of the Crab pulsar and PSR 1937 + 21 (▲). Data points near period of 1000 ms are averages for pulsars with periods greater than 200 ms from observations in Backer (1976); other points represent individual objects. Lines with slope – 0.35 appear to represent the period scaling of pulse widths for all three types of pulse morphology.

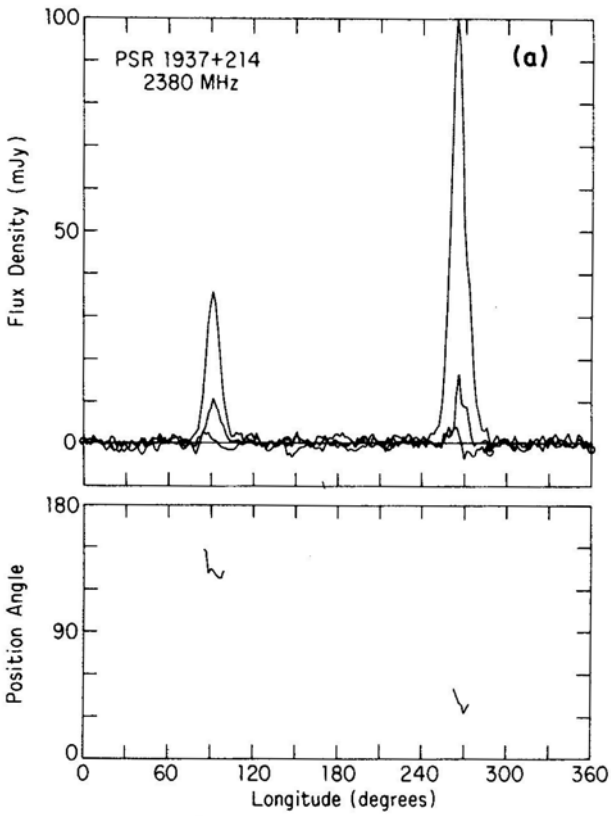


Figure 5. High-resolution pulse profiles for PSR 1937 + 21 recorded at 2380 Hz (Stinebring 1983) and at 430 MHz (Stinebring & Cordes 1983). The increased pulse width at 430 MHz in comparison with 2380 MHz is an effect seen in other pulsars. At 2380 MHz there is evidence for structure in the main pulse. There is a clear rotation of the polarization angle across the 430-MHz main pulse (© *Nature*).

Lyne & Smith (1979) did not select data according to the specific pulse morphology.

The mainpulse and interpulse components in the Crab pulsar have widths of only 3.8° in sharp contrast to the beams discussed above since they are less than one half of the opening angle of the smallest possible open magnetosphere region at the neutron star surface. This radio emission is also notable for its high-energy counterparts in the optical through gamma-ray bands. The beaming in the unseen latitude coordinate must be much broader than 3.8° so that the probability of the beams sweeping by the Earth is reasonable for this unique supernova remnant. The sharp cusp in the optical pulse is also expected to be extended in latitude. The mainpulse and interpulse emission of the Crab pulsar does not fit in with the dipole-axis radiation discussed above. The pulse-interpulse morphology of PSR 1937 + 21 (Figs 2 and 5) shows strong resemblance to the two sharp peaks of the Crab. The pulse widths, 15°, are also narrower than the opening angle of the closed field lines, as in the Crab. The widths of the PSR 1937 + 21 and Crab pulses also scale with a low power of the period. In Fig. 4 I relate the sharp pulse-interpulse morphology in the Crab and PSR 1937 + 21 with a line, $2.0° \; P^{-0.35}$, marked Cusp for the sharp peaks evident in the Crab optical pulse. Alternatively the

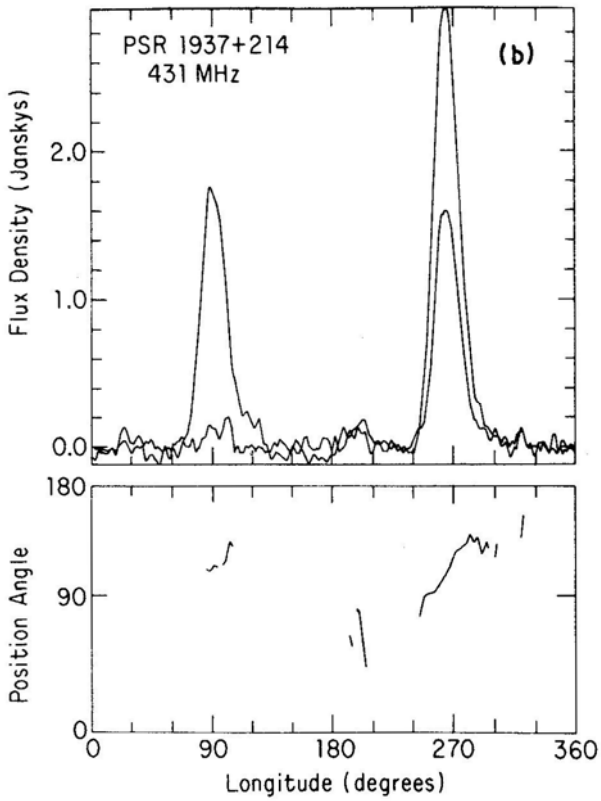


Figure 5. Continued.

Crab and PSR 1937 + 21 sharp pulses could be treated as the periphery of a very wide hollow cone (Manchester & Lyne 1977; Arons 1981), and plotted above the Double/Triple line in Fig. 4. The relationship of the pulse-width data to accurate models of radio emission in pulsar magnetospheres will be treated elsewhere.

3.2 *Other Observations*

The radio images in the discovery paper and in Becker & Helfand (1983) are consistent with the hypothesis that the pulsar is not associated with an extended continuum source. Recent observations at the VLA have provided further confirmation of this hypothesis (Goss, Fomalont & Backer 1984). Extended in this context means 2 arcsec to 60 arcsec, and nondetection is a brightness temperature limit of 1600 K. This observation places a limit of $10^{30}\ \theta_1^2$ erg s^{-1}, with θ in units of 10 arcsec, on the radio luminosity from a light year scale region around the pulsar. On a scale of 50 pc Sieber & Seiradakis (1984) find a plateau of emission surrounding the pulsar and nearby sources with a radius of 19 arcmin and an integrated intensity of 0.6 Jy. At our 5-kpc distance this plateau source would have a luminosity of $\sim 10^{32}$ erg s^{-1}. While these luminosities are well below the energy loss indicated by the spin decay, 10^{36} erg s^{-1}, the radio continuum radiation expected from synchrotron radiation of the high-energy particle

spectrum produced in the pulsar is several orders of magnitude below either of the above luminosities (Arons 1983).

The most likely band for the detection of the continuum from the particle spectrum produced by the pulsar is the hard X-rays around 100 keV (Arons 1983). A complete determination of the spectrum of this radiation is possible with future X-ray observatories and would provide a direct measure of the contribution of particle acceleration to the spin decay. Synchrotron nebulae have been detected for several other pulsars with the Einstein Observatory (Cheng & Helfand 1983). Arons' model for subsonic expansion of a bubble blown by the pulsar relativistic-particle flux predicts a 2–10 keV X-ray flux of 10^{-11} erg cm^{-2} s^{-1} from a region 10 arcsec in diameter. Observations with the image proportional counter of the Einstein satellite show the presence of a source of low significance that is consistent with Arons' estimate when corrected for interstellar absorption and the narrow bandwidth of the observation (Marshall & Grindlay 1983). No pulse search is possible since there are at most 20 photons in the observation. An image of the pulsar field was also recorded with the high-resolution imager (0.1–3.5 keV) on board the Einstein Observatory by Becker & Helfand (1983). Their intensity limit is consistent with the IPC possible detection when corrected for interstellar absorption over a 5-kpc path. Becker & Helfand note that their limit places an upper bound to the surface temperature of the neutron star of 2×10^6 K.

Many attempts have been made to detect a signal from PSR 1937 + 21 at high frequencies. Djorgovski (1982) proposed an identification with a 20-mag star. More recent astrometry and spectroscopy place Djorgovski's star 2 arcsec from the pulsar and identify it as a K giant (Lebofsky & Rieke 1983). Coincidentally its distance is estimated as 5 kpc, similar to that of the pulsar. Middleditch *et al.* (1983) place limits of 23 mag (red) and 16 mag (2 μm) on pulsed power in low harmonics of the pulsar period. The limits in Middleditch *et al.* are consistent with the tentative detection of pulsed power by Manchester, Peterson & Wallace (1983). More recent work has not confirmed this detection (R. N. Manchester 1983, personal communication).No continuum object brighter than 23 red magnitude is found at the pulsar position on deep CCD images taken by Djorgovski & Spinrad (1983).

A search of the COS-B source list (Swanenburg *et al.* 1981) and a pulse search by Thompson *et al.* (1983) reveals no gamma-ray source associated with the pulsar.

These investigations have revealed no convincing detection of either pulsed or continuum radiation from 1937+21 and its environs above 10^{10} Hz. The large radio source found by Sieber & Seiradakis (1984) and the X-ray point source seen by Marshall & Grindlay (1983) require further investigation with more sensitivity. These observations are summarized in Fig. 6.

Two other experiments have been conducted to search for signals from PSR 1937 +21. Hough *et al.* (1983), encouraged by the initial report of a large spindown rate, searched for gravitational waves with a detector tuned to the second harmonic of the pulsar frequency. Their amplitude limit, $h \sim 10^{-20}$, is seven orders of magnitude above the amplitude expected if all the spindown energy loss, 10^{36} $ergs^{-1}$, were in gravitational waves at 2 Ω. More recently higher sensitivity gravitational wave experiments have been initiated with more sensitive detectors operated for longer intervals (M. Herald 1983, personal communication). A second class of experiments has been conducted to search for direct magnetic multipole radiation from the pulsar by Fourier analysis of Viking satellite magnetometer data (D. Morris & R. Muller 1983, personal communication).

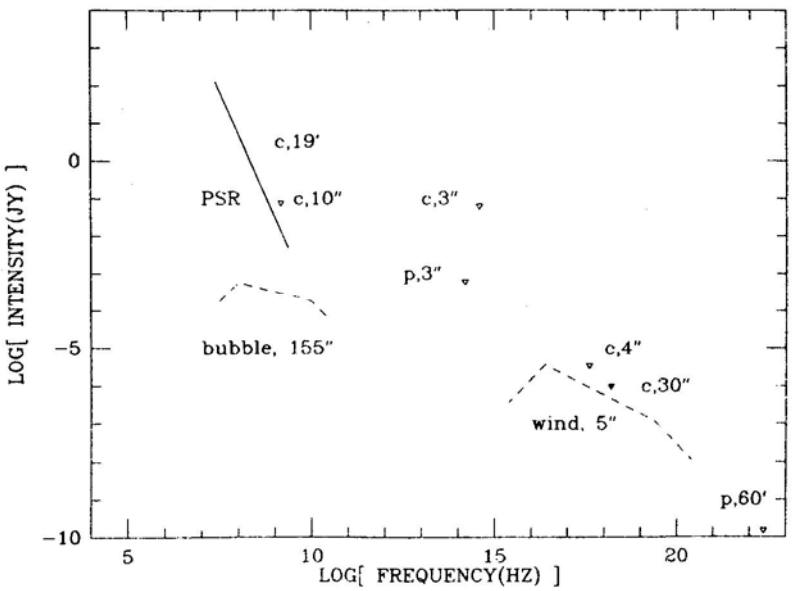


Figure 6. Electromagnetic spectrum of PSR 1937 + 21. The radio pulsar spectrum is shown with a solid line marked PSR. Limits on pulsed intensity at various frequencies are shown with the symbol (∇) and notation p, x, where x is the detector aperture size. Detections or limits of continuum intensity associated with the pulsar are noted by c, x, Arons' (1983) predictions for continuum emission resulting from particles and magnetic field emanating from the pulsar are shown with dashed lines.

Free-free absorption of these waves along the 5-kpc path in the interstellar medium will be severe for the first 80 harmonics of PSR 1937 + 21 (Lipunov 1983).

3.3 *Formation scenarios*

Two classes of models have been proposed to explain the puzzling combination of three properties of PSR 1937 + 21: the fast spin of 642 Hz, the low inferred surface field of 10^9 G perpendicular to the rotation axis, and the large inferred age of 10^5 yr. In the favoured class of models the neutron star begins its life in a binary system and with a strong surface magnetic field of 10^{12} G. The assumption of an initially strong field is supported both by the observed narrow range of $P\dot{P}$ in most pulsars, by the observed cyclotron line in spectra of X-ray binaries, and by the idea that heat flux from the cores of cooling neutron stars will always set up currents that produce surface fields of order 10^{12} G independent of the magnetic flux of the pre-supernova stellar core (Blandford, Applegate & Hernquist 1983). If the secondary leaves the main sequence after the companion star's field has decayed, then angular momentum transfer from the swelling secondary can spin the neutron star up to millisecond periods. Bright X-ray sources in the galactic bulge are presumed to be low-mass binaries in this stage of evolution (Alpar *et al.* 1982; Fabian *et al.* 1983). When the secondary evolves to a degenerate state the binary may or may not be disrupted. The result is either an isolated pulsar such as PSR 1937 + 21 or a degenerate binary such as PSR 1953 + 29. This scenario was first suggested to explain the isolated location of several pulsars in a P–$\dot{P}$ diagram

(Radhakrishnan 1982; Backus, Taylor & Damashek 1982). Alternatively Helfand, Ruderman & Shaham (1983) propose a binary evolution model wherein the primary evolves first into a white dwarf and later into a neutron star after mass accretion from the evolving secondary coaxes it over the Chandrasekhar mass limit for white dwarfs.

Ruderman & Shaham (1983) have refined the binary model to account for the difficulties associated with the disruption of a low-mass binary system required to explain PSR 1937 + 21 (Arons 1983). In their model the secondary is nearly consumed by the accreting neutron star. They suggest that an asteroidal mass may remain in orbit around the 1.5-ms pulsar. Pulse timing accuracy is insensitive to the small effects of such a mass. The birth and evolution of degenerate binary systems is considered by van den Heuvel (1984).

One consequence of the binary evolution scenario is that most neutron stars are spun down during their lifetime with their main sequence companion. During this time they are not seen as radio pulsars. If they escape the binary before their field decays and mass accretion leads to spin-up, then they appear as radio pulsars with slow periods. This injection of pulsars into the P-$\dot{P}$ diagram at slow periods has been suggested recently by Vivekanand & Narayan (1981). However, the amount of injection they require exceeds the estimated values for this effect (de Loore, de Greve & de Cuyper 1975). Instead, Vivekanand & Narayan suggest a throttle on the radio emission that depends on period.

While the binary hypothesis is strongly favoured by the existence of low magnetic moment pulsars in binaries and by accreting low-mass X-ray binaries, PSR 1937 + 21 is not in a binary and thus could have formed in the collapse of an isolated star. If so, a mechanism is required to limit the surface magnetic field to the observed value despite the currents associated with the heat flux in the rapid cooling phase of the neutron star. One consequence of this model which contrasts sharply with the binary models is the large expected population of millisecond pulsars (Brecher & Chanmugam 1983). The new fast pulsar searches discussed below are essential to deciding between alternative formation models.

3.4 *Neutron-Star Astrophysics*

The rapid rotation of PSR 1937 + 21 places it near the stability edge for neutron stars. The ratio of kinetic to potential energy, $T/|W| = \Omega^2/2\pi G\rho$, is 0.08 for a mean density of 5×10^{14} g cm^{-3}. Neutron-star models make a transition from stable, axisymmetric configurations to unstable, non-axisymmetric configurations at values of T ranging from 0.08 to 0.22 (Datta & Ray 1983; Friedman 1983; Harding 1983). A limit to the axisymmetry of PSR 1937 + 21 can be obtained by equating the observed spindown energy loss to gravitational wave radiation from a weakly triaxial mass distribution. The ellipticity for the equatorial moments of inertiá must be less than $10^{-6.5}$; in other words, the equatorial surface is circular with an accuracy of 38 μm!

The rotation rate 642 Hz does not lead to new limits on the equation of state of neutron-star matter (Shapiro,Teukolsky &Wasserman 1983). If the age of the pulsar is less than 10^7 yr, then even the original spin is consistent with existing neutron-star models. However, the proximity of 642 Hz to the stability limit and the spin-up models discussed above suggest that PSR 1937 + 21 was probably marginally stable at birth. If so, then this system and others like it are a likely source of gravitational waves (Wagoner 1984).

4. Pulse arrival-time events—a classical clock

Analysis of accurate arrival-time data from PSR 1937+21 at the Arecibo Observatory has demonstrated that the times of pulse arrival 'events' can be modelled with a precision of 1 μs (Backer, Kulkarni & Taylor 1983; Fig. 7). The fractional error of 1 μs for one year, 3×10^{-14}, implies that observations of PSR 1937 +21 may provide the most stable basis for reckoning dynamical time over long timescales. The source of this stability is the rapid and near frictionless rotation of a stellar mass; its mechanical Q, $2\pi/\dot{P}$ is 6×10^{19}. While the structure of a neutron star and the beam-production process require a knowledge of modern physics, the rotation of the star is governed by the classical physics of a rotating, magnetized body (Landau & Lifshitz 1962; Pacini 1967). Sensitive timing observations of PSR 1937 + 21 and other fast pulsars are likely to provide new data both for the construction of precise models of solar system dynamics and for the search for gravitational waves.

Following the discovery of PSR 1937 + 21 we began accurate arrival-time measurements with the system developed for timing the binary pulsar PSR 1913 + 16 (Taylor & Weisberg 1982). The pulse-broadening effects of interstellar dispersion were removed with a digital de-disperser (Boriakoff 1973). This equipment provides an average pulse every two minutes. The time at the peak of this average for a pulse in the middle of the two minutes is the observatory, or topocentric, arrival time for that average. Topocentric arrival times are then compared to a model which includes parameters of the pulsar's spin, the equatorial coordinates of the object, the electron column density between the earth and the pulsar, the time scale at the observatory, and the location of the observatory in an inertial reference frame. The model has been upgraded

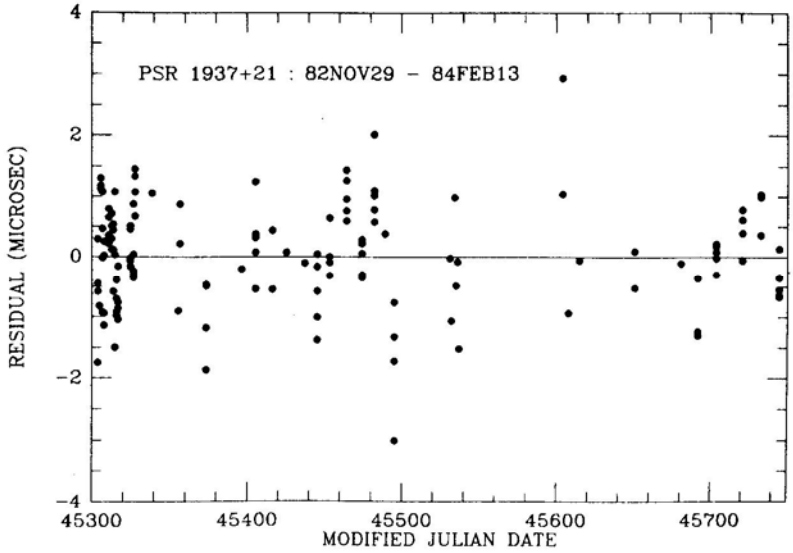


Figure 7. Residuals from an arrival-time model fit to the first 14 months of data from PSR 1937 + 21 recorded at 1400 MHz at the Arecibo Observatory (Davis *et al.* 1984). Each point represents 30 minutes of integration. The errors are dominated by clock errors of the observatory clock with respect to the USNO standard.

Substantially over the past year to match the precision of the new data. Parameters from a recent fit are given in Table 1.A summary of the modelled parameters and estimates of unmodelled effects is given in Table 2 with explanations below. Column 1 gives the effect for groups of model parameters given in column 2. Columns 3 and 4 contain contributions to the arrival time in microseconds for our model and for estimates of errors. A more complete discussion of the first year of timing is being prepared for publication elsewhere (Davis, *et al.* 1984).

The 'spin' properties of the neutron star are modelled by an initial phase, a frequency, Ω, and a frequency derivative, $\dot{\Omega}$. The second derivative of the spin frequency is small for a spin decay by electromagnetic torques; *i.e.*, $\dot{\Omega} \sim \Omega^n$ with $n \sim 3$. Free precession could affect the arrival times, but has never been detected in any pulsar observation (*e.g.,* Pines & Shaham 1974). Starquakes and other unstable 'activity' are correlated with $\dot{P}$ (Cordes & Helfand 1980). We can only guess that the extremely low rate of energy loss for PSR 1937+21 will lead to smooth changes, or rare, abrupt changes as the star adjusts to new equilibrium figures.

The pulsar signal passes through four 'propagation' media with non-unity refractive index. Both the interstellar plasma and the interplanetary plasma may have time-variable propagation path-lengths with observable effects. The interstellar path will vary if the power-law spectrum indicated by interstellar scintillation studies at wavelengths of 10^{10-12} cm continues to rise at much longer spatial scales (Armstrong 1984). The interplanetary path will vary with solar cycle (Muhleman & Anderson 1981). These plasma effects may be suppressed by careful multifrequency observations (Cordes & Stinebring 1984).

Comparison of arrival-time data from different Observatories' will require corrections to a fiducial point given by the observatory coordinates. These corrections include both a free-space term and cable/hardware propagation. While these are listed in Table 2 as error terms, they are nearly constant and do not affect estimation of modelled parameters.

The 'time' of a given pulse arrival event, such as the pulse peak, is initially recorded on the observatory UTC time scale. This scale is generated by an atomic clock, and is referred to UTC at Jupiter, Florida, UTC(JUP), by reception of Loran C transmissions at 100 kHz. The stability of this ground wave propagation path can be maintained, with care, to 0.2 μs. UTC(JUP) is referred to UTC at the U.S. Naval Observatory, UTC(USNO), by a USNO monitoring programme. The UTC scale does not run uniformly. Approximately once a year an extra second is added to keep UTC close to solar time defined by the rotation of the Earth. The next step in assessing the pulse arrival-time on a uniformly running clock is a coordination offset between UTC(USNO) and uniformly running atomic time, TAI(USNO), which is based on a

Table 1. Parameters for PSR 1937 + 21.

Declination (1950.0)	21° 28′ 01″.447
Period	0.001 557 806 448 858 s
Period derivative	1.0490×10^{-19} s s^{-1}
Epoch	JD 2445303.2634439
Dispersion measure	71.078 e pc cm^{-3}
RMS residual for 2 min averages	2.2 μs
Interval reduced	1982 November–1984 February

Table 2. Contributions to the arrival time model.

Effect Parameter	Model* (μs)	Estimate of error* (μs)
Star		
spin phase	$8.6 \times 10^{10}\ t$	...
spin decay	$2.5 \times 10^{-1}\ t^2$	...
spin deceleration ($n = 3$)	...	$1.5 \times 10^{-12}\ t^3$
precession, θ	...	$1.6 \times 10^{3}\ \theta \sin i$
star quakes	??	
Propagation (1.4 GHz)		
interstellar	1.5×10^{5}	$2.2 \times 10^{-3}\ t^{0.8}$
interplanetary	...	0.17–0.44
ionosphere	...	0.01–0.04
troposphere	...	0.006–0.010
Observatory		
fiducial point	...	1
cable	...	2
disperser	...	15
Time		
AO-UTC(JUP)	50818	0.2
UTC:JUP-USNO	< 1	0.1
TAI-UTC	21.034	$1.0 \times 10^{-3}\ t$
TDT-TAI	32.184	0
TDB-TDT	1500	2
Inertial Frame		
Earth centre	21000	0.01
Earth tides	...	0.003
Polar motion/UTI	...	0.010
Barycentre	3.5×10^{8}	$1.4 \times 10^{-3}\ t$
Other		
Proper motion ($v = 100\,\mathrm{km\,s^{-1}}$)	...	$1.8 \times 10^{-2}\ t$
Parallax ($d = 5$ kpc)	...	0–0.25
Shklovskii term	...	$8.3 \times 10^{-4}\ t^2$
Satellite ($m = 10^{21}$ gm; $P = 10^{4}$ s)	...	1.3×10^{-6}
GR delay of Sun	0.2–20	0.02
GR delay of star	...	$640 + 2.6 \times 10^{-4}\ t$
$\dot{G}/G = 10^{-18}$	...	$3.7 \times 10^{-3}\ t^2$
Galactic acceleration	...	$2.1 \times 10^{-3}\ t^2$
Gravitational waves	??	

* t in the units of days

bank of 35 cesium clocks. The absolute stability of TAI is thought to be 10^{-14}. A further coordination offset is required to refer the TAI arrival time to the Terrestrial Dynamical Time scale, TDT, for computation of the Earth's location in inertial space.

Both TAI and TDT are proper times for an observer on Earth. However the pulse

arrival times will be periodic only for an observer moving uniformly with respect to the pulsar in a fixed gravitational potential. Consequently, the Earth proper time scales, TAI or TDT, need to be corrected for the integrated effects of transverse Doppler and gravitational redshift. Moyer (198la, b) has given formulae for correction of Earth-based proper time to coordinate time for a clock in a potential equal to the average at the Earth's orbit and moving with the average speed of the Earth. The submicrosecond precision required for timing PSR 1937 + 21 exceeds that possible with Moyer's closed-form expressions (R. W. Hellings 1983, personal communication). A numerical integration of the ratio of coordinate and proper time scales is required. The definition of coordinate time, or Barycentric Dynamical Time (TDB), depends on the method used to remove the effects of transverse Doppler and gravitational redshift; current definitions of time are discussed in the 1984 American Ephemeris and Nautical Almanac. A dynamical time scale based on pulsar observations was proposed at the Patras IAU meeting (Il'in *et al.* 1982). Comparison of the PSR 1937 + 21 data to an accurate model constitutes a test of the Strong Equivalence Principle over long timescales and on a solar system length-scale with a precision of $\sim 10^{-4}$; this precision is comparable to that attained in a short-term, near-Earth test performed with hydrogen masers (Vessot *et al.* 1980; see also Canuto, Goldman & Shapiro 1984).

The topocentric pulse arrival time on the TDB scale at the observatory must next be corrected for time of flight to a fixed point in an 'inertial reference frame'. We correct first from the observatory fiducial point to the Earth centre. Next we correct from the Earth centre to the solar system barycentre using an ephemeris for the Earth's orbit based on recent radar observations as well as a long record of optical observations (Ash, Shapiro & Smith 1967; Downs & Reichley 1983). These corrections require knowledge of the pulsar position with an accuracy of 300 μarcsec. The position is derived by least-squares fit from the data since the accuracy of independent radio interferometry is only 50 milliarcsec (Goss, Fomalont, & Backer 1984). Comparison of the position derived from timing and from interferometry may be used to tie the two independent coordinate frames together at a new level of accuracy (Fomalont *et al.* 1984).

Several 'other' effects are, or may be, important for arrival time analysis. With several years of observations we will be able to solve for a proper motion. The magnitude and orientation of this motion, $v_\perp/D$, with respect to the galactic plane will be important information to apply to discussions of the origin of PSR 1937+21. The ISS parameters of PSR 1937+21 lead to a proper motion estimate of 100 kms^{-1} using the method of Lyne & Smith (1982). The observations may also provide the first solution for a trigonometric parallax by pulse-timing technique. Solution for the proper motion will also allow computation of the contribution of the Shklovskiĭ term to $\dot{\Omega}$, — $v_\perp^2\Omega/2D\mathrm{c}$ (Shklovskiĭ 1969).

The refinement of the binary model by Ruderman & Shaham (1983) suggests the possibility of a remnant satellite in orbit around the pulsar with an asteroidal mass. With the microsecond residuals presented above we are unable to discern the effects of companion object less than 10^{27} g for periods of 10^4 s. The new millisecond pulsars add to the list of pulsars at 19 hours with low $\dot{P}/P$ ratios used by Harrison (1977) to suggest the presence of a dark companion to the Sun. Cowling (1983) has shown that Harrison's effect is not significant with a larger data sample.

There are several effects on the photons from the pulsar involving nature's most puzzling force, gravity. The passage of photons through the weak solar gravitational field results in a delay which varies from 0.2 to 20 μs (Shapiro 1964). This delay can be

predicted with a precision of 10^{-3} based on recent radar measurements of the PPN coefficient (Reasenberg *et al.* 1979). The star which was originally identified with PSR 1937 + 21 will also produce a gravitational delay if it is closer to the Earth than the pulsar. Variation of the delay from differential proper motion will produce an unmeasurable shift of 10^{-14} in the true pulse frequency. Gravitational waves from a binary system near the line of sight may be detectable (Sazhin 1978).

Dirac cosmologies predict spontaneous creation of matter accompanied by a change in the gravitational constant, G, by an amount $\dot{G}/G \sim 3 \times 10^{-18}\ \mathrm{s}^{-1}$ (Dirac 1973). A change in G and in mass will be observed in timing PSR 1937+21 owing to its effect on the moment of inertia. The contribution of this effect to the observed spindown is less than 10 per cent. We expect that the spindown is dominated by energy losses related to the generation and acceleration of electron-positron plasma which is required in most models to produce the observed radiation. Finally the residuals of the arrival times of PSR 1937+21 from our best model of the above a priori effects can be inspected for the presence of gravitational waves from either discrete objects (Adams *et al.* 1982) or from a stochastic background (Romani & Taylor 1983; Hellings & Downs 1982). This new data will be the best detector for parsec-scale gravitational radiation; the sensitivity to a stochastic background will be improved by 10^4.

5. New searches—needle(s) in a haystack

The discovery of a pulsar spinning twenty times faster than the cutoff rate in various all-sky surveys has led to great interest in defining the galactic population of similar fast pulsars. The entire sky is reopened to new pulsar searches. During 1983 one simlarly fast pulsar was confirmed in a search area of 4 square degrees (Boriakoff, Buccheri & Fauci 1983). There are two results which will come from the current round of pulsar investigations. First, we will assess the formation rate of fast pulsars with respect to the slow pulsars; is this one in 10, or 100, or less? Second, deep searches over a large volume of the galaxy may find a solitary object whose properties are even more exotic than PSR 1937 + 21; for example, a shorter period which would lead to a revision of current notions concerning the maximum spin rate for neutron stars, or a fast spin and a short binary period which would lead to general relativity tests envisioned by Zwicky with precision exceeding that possible in the PSR 1913 + 16 system.

Previous pulsar search surveys have been limited in sensitivity by the use of narrow bandwidths to prevent dispersion smearing and short integration time to allow coverage of a large field of view. Computational limitations and the 'known' period distribution led to typical sampling intervals of 20 ms. Searches for faster pulsars require a two-fold increase in the search complexity if the detection level is maintained since both the filter width and the sample time must be reduced.

The discovery story related above suggests several paths toward pulsar detection which do not initially require pulsed signal searches. Scintars are compact radio sources found at low galactic latitudes which show strong IPS at metre wavelengths (Rickard & Cronyn 1979). The steep spectrum of scintar 4C 21.53 added to its peculiarity. S. R. Kulkarni, C. Heiles, A. Purvis, J. Baldwin, P. Werner, W. M. Goss & J. van Gorkom (1984, personal communication) have investigated a new list of steep-spectrum scintars in the northern sky from recent Cambridge studies. They have done high-resolution imaging of candidates for the scintars at 1400 MHz and have done pulse searches

without positive results at the present time. W. E. Erickson (1983, personal communication) has compiled a similar list of compact, steep-spectrum objects in the southern sky. One conclusion from the scintar studies is that IPS can be seen for extragalactic objects at low galactic latitudes. Therefore the material responsible for interstellar scattering is not uniformly distributed.

A second approach to pulsar candidate surveying uses interstellar scintillation (ISS). At metre wavelengths ISS requires a source size of 10^{-6} arcsec or less. A 10 mJy radio source would need to have a brightess temperature of 10^{18} K to show ISS at 1-m wavelength. Only pulsars are known to have a coherent radiation mechanism that allows such brightness temperatures. ISS is distinguished from IPS by its narrow modulation bandwidth, $d\nu/\nu < 10^{-3}$, and its slow time scale, $t > 10$s. The first sky survey for ISS objects has been initiated by M. A. Stevens, C. Heiles & D. C. Backer on the Green Bank 300-ft telescope. A portion of this survey covering 200 square degrees is shown in Fig. 8. The detection level at present is near 100 mJy as shown by the response to PSR 0329 + 54 in the beam adjacent to its position on the sky; the full intensity of PSR 0329 + 54 is about 1 Jy. Interference is responsible for most of the other peaks.

A third approach to pulsar searching was suggested by the pre-pulsar detection of linear polarization of the compact object in 4C 21.53. High-resolution, polarized-intensity maps of the galactic plane would be expected to show up further undetected

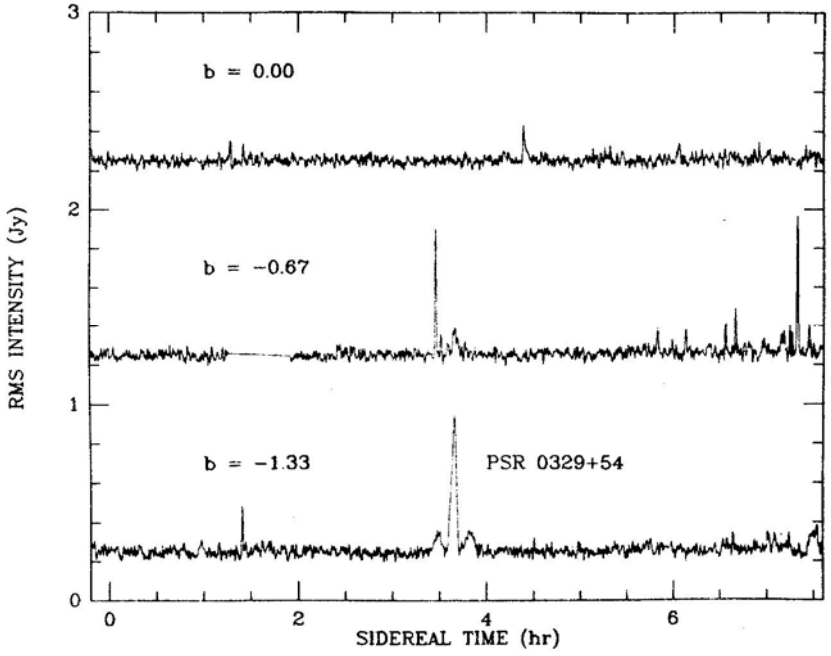


Figure 8. A sample of the interstellar scintillation survey conducted by M. A. Stevens, C. Heiles & D. C. Backer at 360 MHz with the NRAO 300-foot antenna in Green Bank. Each tracing contains the rms intensity over a 750-kHz spectrum which was integrated for 40s. A reference spectrum is removed using preceding and following scans. The three tracings cover approximately 200 square degrees of the galactic plane. The detection of PSR 0329 + 54 in the $b = -0.67°$ scan indicates a sensitivity of ~ 100 mJy for this survey technique. Most other peaks can be traced to narrow-band interference.

pulsars. High sensitivity and wide-field coverage necessitate a survey at metre wavelengths. An individual field of several square degrees can be mapped in several hours to a detection level of ~ 10 mJy in the polarized intensity; for 30 per cent polarization fraction this gives a sensitivity of ~ 30 mJy. Groups at Westerbork (M. A. Stevens, C. Heiles, D. C. Backer, U. J. Schwarz, A. Purvis & W. M. Goss), at Jodrell Bank (A. G. Lyne 1983, personal communication), and at the VLA (J. Dickey 1984 and D. Stinebring 1983; personal communications) have initiated this type of research.

Finally several groups are undertaking the demanding task of high sensitivity, fast pulsar searches. Boriakoff has searched the fields of a number of unidentified gamma-ray sources with 4 ms sampling (Boriakoff, Buccheri & Fauci 1983). Groups at Jodrell Bank and at Molonglo are taking data at specified targets such as plerions (R. N. Manchester 1983 and A. G. Lyne 1983; personal communications). Taylor and colleagues have begun a new all-sky survey with period sensitivity increased to 4 ms (J. H. Taylor 1984, personal communication). At Berkeley we are developing a 'search machine' capable of a real-time search for pulse periods down to 0.5 ms.

6. Conclusion

The discovery of fast pulsars has led to fresh insights into the astrophysics of neutron stars. This new chapter in pulsar astronomy will remain unfinished until searches have defined the galactic population of fast pulsars. The diversity of the pulsars with spins in excess of 10 Hz makes these searches compelling. Even a few additional fast pulsars will reveal new clues about magnetospheric structure, about the formation of neutron stars, and, quite likely, about the fundamental physics of gravitational waves. The ensemble of fast pulsars may lead to the Earth's most precise time scale for improving the solar system dynamical model and for detection of minute effects of general relativity. The discovery of PSR 1937 + 21 in an unexplored domain of parameter space is a reminder to all scientists that many of nature's best secrets remain to be discovered.

Acknowledgements

I thank my colleagues S. Kulkarni, J. Arons and J. Barnard for reading this manuscript and for discussions.

References

Ables, J. G., Manchester, R. N. 1976, *Astr. Astrophys.*, **50,** 177.
Adams, P. J., Hellings, R. W., Zimmerman, R. L., Farhoosh, H., Levine, D. I., Zeldich, S. 1982, *Astrophys. J.*, **253,** 1.
Alpar, M. A., Cheng, A. F., Ruderman, M. A., Shaham, J. 1982, *Nature*, **300,** 728.
Altenhoff, W. J., Downes, D., Pauls, T., Schraml, J. 1979, *Astr. Astrophys. Suppl. Ser.*, **35,** 23.
Armstrong, J. 1984, *Nature*, **307,** 527.
Arons, J. 1981, in *IAU Symp. 95: Pulsars*, Eds W. Sieber & R. Wielebinski, D. Reidel, Dordrecht, p. 69.
Arons, J. 1983, *Nature*, **302,** 301.
Ash, M. E., Shapiro, I. I., Smith, W. B. 1967, *Astr. J.*, **72,** 338.
Baade, W. 1942, *Astrophys. J.*, **96,** 188.

Backer, D., Kulkarni, S., Heiles, C. 1982, *IAU Circ.* No. 3746.
Backer, D., Kulkarni, S., Heiles, C., Davis, M., Goss, M. 1982a, *I AU Circ.* No.3743.
Backer, D. C. 1976, *Astrophys. J.*, **209,** 895.
Backer, D. C, Hazard, C., Jauncey, D. L., Sutton, J. 1970, *Astr. J.*, **75,** 529.
Backer, D. C., Kulkarni, S. R., Heiles, C, Davis, M. M., Goss, W. M. 1982b, *Nature*, **300,** 615.
Backer, D. C., Kulkarni, S. R., Taylor, J. H. 1983, *Nature*, **301,** 314.
Backus, P. R., Taylor, J. H., Damashek, M. 1982, *Astrophys. J.*, **255,** L63.
Becker, R. H., Helfand, D. J. 1983, *Nature*, **302,** 688.
Blandford, R., Applegate, J., Hernquist, L. 1983, *Mon. Not. R. astr. Soc*, **204,** 1025.
Boriakoff, V. 1973, *Ph D thesis*, Cornell University, Ithaca.
Boriakoff, V., Buccheri, R., Fauci, F. 1983, *Nature*, **304,** 417.
Brecher, K., Chanmugan, G. 1983, *Nature*, **302,** 124.
Canuto, V. M., Goldman, I, Shapiro, I. I. 1984, *Astrophys. J.*, **276,** 1.
Cheng, A. F., Helfand, D. J. 1983, *Astrophys. J.*, **271,** 271.
Cordes, J. M., Helfand, D. J. 1980, *Astrophys. J.*, **239,** 640.
Cordes, J. M., Stinebring, D. 1984, *Astrophys. J.*, **277,** L53.
Cowling, S. A. 1983, *Mon. Not. R. astr. Soc.*, **204,** 1237.
Datta, B., Ray, A. 1983, *Mon. Not. R. astr. Soc.*, **204,** 75p.
Davis, M. M., Taylor, J. H., Weisberg, J., Backer, D. C. 1984, in preparation.
de Loore, C., de Greve, J. P., de Cuyper, J. P. 1975, *Astrophys. Space Sci.*, **36,** 219.
Dirac, P. A. M. 1973, *Proc. Roy. Soc. London*, **A333,** 403.
Djorgovski, S. 1982, *Nature*, **300,** 618.
Djorgovski, S., Spinrad, H. 1983, *Nature*, **306,** 569.
Douglas, J. N., Bash, F. N., Torrence, G. W., Wolf., C. 1980, *Univ. of Texas Publ. Astr.*, No. 17.
Downs, G. S., Reichley, P. E. 1983, *Astrophys. J. Suppl. Ser.*, **53,** 169.
Duffett-Smith, P. J., Readhead, A. C. S. 1976, *Mon. Not. R. astr. Soc.*, **174,** 7.
Erickson, W. E. 1980, *Bull. am. astr. Soc.*, **11,** 685.
Fabian, A. C, Pringle, J. E., Verbunt, F., Wade, R. A. 1983, *Nature*, **301,** 222.
Fomalont, E. B., Goss, W. M., Lyne, A. G., Manchester, R. N. 1984, *Mon. Not. R. astr. Soc.*, (in press).
Friedman, J. L. 1983, *Phys. Rev. Lett.*, **51,** 11.
Giacconi, R., Gursky, H., Kellogg, E., Schreier, E., Tananbaum, H. 1971, *Astrophys. J.*, **167,** L67.
Gold, T. 1968, *Nature*, **218,** 731.
Goss, W. M., Fomalont, E. B., Backer, D. C. 1984, in preparation.
Harding, A. K. 1983, *Nature*, **303,** 683.
Harrison, E. R. 1977, *Nature*, **270,** 324.
Heiles, C, Kulkarni, S. R., Stevens, M. A., Backer, D. C., Davis, M. M., Goss, W. M. 1983, *Astrophys. J.*, **273,** L75.
Helfand, D. J., Ruderman, M. A., Shaham, J. 1983, *Nature*, **304,** 423.
Hellings, R. W., Downs, G. S. 1982, *Astrophys. J.*, **265,** L39.
Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F., Collins, R. A. 1968, *Nature*, **217,** 709.
Hewish, A., Okoye, S. 1965, *Nature*, **207,** 59.
Hough, J., Drever, R. W. P., Ward, H., Munley, A. J., Newton, G. P., Meers, B. J., Hoggan, S., Kerr, G. A. 1983, *Nature*, **303,** 216.
Hulse, R., Taylor, J. H. 1974, *Astrophys. J.*, **191,** L59.
Hulse, R., Taylor, J. H. 1975, *Astrophys. J.*, **195,** L51.
Il'in, V. G., Ilyasov, Ya. P., Kuzmin, A. D., Pushkin, S. B., Palij, G. N. Shabanova, T. Y., Shitov, Yu. P. 1982, *Trans. I AU*, **18B,** 241.
Joss, P. C., Rappaport, S. A. 1983, *Nature* **304,** 419.
Landau, L. D., Lifshitz, E. M. 1962, *The Classical Theory of Fields*, *2* edn, Pergamon, New York.
Lebofsky, M. J., Rieke, G. H. 1983, *IAU Circ.* No.3809.
Lipunov V. M. 1983, *Astr. Astrophys.*, **127,** L1.
Lyne, A. G., Smith, F. G. 1979, *Mon. Not. R. astr. Soc.*, **188,** 675.
Lyne, A. G., Smith, F. G. 1982, *Nature*, **298,** 825.
Manchester, R. N., Lyne, A. G. 1977, *Mon. Not. R. astr. Soc.*, **181,** 761.
Manchester, R. N., Peterson, B. A., Wallace, P. T. 1983, *IAU Circ.* No. 3795.
Marshall, F., Grindlay, J. E. 1983, presented at the APS meeting, Baltimore, 18–21 April.

Middleditch, J., Cudaback, D., Oliver, B., Pennypacker, G, Lebofsky, M. J., Rieke, G. H., McGraw, J. T., Dearborn, D., Wisniewski, W., Chini, R. 1983, *Nature*, **306,** 163.
Minkowski, R. 1942, *Astrophys. J.*, **96,** 199.
Moyer, T. D., 1981a, *Cel. Mech.*, **23,** 33.
Moyer, T. D. 1981b, *Cel. Mech.*, **23,** 57.
Muhleman, D. O., Anderson, J. D. 1981, *Astrophys. J.*, **247,** 1093.
Pacini, F. 1967, *Nature*, **216,** 567.
Paczyński, B., 1983, *Nature*, **304,** 421.
Pines, D., Shaham, J. 1974, *Comments Astrophys.*, **6,** 37.
Radhakrishnan, V. 1982, *Contemp. Phys.*, **23,** 207.
Radhakrishnan, V., Cooke, D. J. 1969, *Astrophys. Lett.*, **3,** 225.
Rankin, J. M. 1983, *Astrophys. J.*, **274,** 333.
Reasenberg, R. D., Shapiro, I. I., MacNeil, P. E., Goldstein, R. B., Breidenthal, J. C, Brenkle, J. P., Cain, D. L., Kaufman, T. M., Komarek, T. A., Zygielbaum, A. I. 1979, *Astrophys. J.*, **234,** L219.
Rickard, J. J., Cronyn, W. 1979, *Astrophys. J.*, **228,** 755.
Roberts, D. H., Sturrock, P. A. 1972, *Astrophys. J.*, **172,** 435.
Romani, R. W., Taylor, J. H. 1983, *Astrophys. J.*, **265,** L35.
Ruderman, M. A., Shaham, J. 1983, *Nature*, **304,** 425.
Savonije, G. J. 1983, *Nature*, **304,** 422.
Sazhin, M. V. 1978, *Soviet Astr.*, **22,** 36.
Shapiro, I. I. 1964, *Phys. Rev. Lett.*, **13,** 789.
Shapiro, S. L., Teukolsky, S. A., Wasserman, I. 1983, *Astrophys. J.*, **272,** 702.
Shklovskiĭ, I. S. 1969, *Soviet Astr.*, **13,** 562.
Sieber, W., Seiradakis, J. H. 1984, *Astr. Astrophys*., **130,** 257.
Slee, O. B. 1977, *Austr. J. Phys. Suppl*, **36,** 1.
Stinebring, D. R. 1983, *Nature*, **302,** 690.
Stinebring, D. R. Cordes, J. M. 1983, *Nature*, **306,** 349.
Swanenburg, B. N., Bennett, K., Bignami, G. F., Buccheri, R., Caraveo, P., Hermsen, W., Kanbach, G., Lichti, G. G., Masnou, J. L., Mayer–Hasselwander, H. A., Paul, J. A., Sacco, B., Scarsi, L., Wills, R. D. 1981, *Astrophys. J.*, **243,** L69.
Taylor, J. H., Weisberg, J. 1982, *Astrophys. J.*, **253,** 908.
Thompson, D. J., Bertsch, D. L., Hartman, R. C, Hunter, S. D. 1983, *Astr. Astrophys.*, **127,** 220.
van den Heuvel, E. P. J. 1984, *J. Astrophys. Astr.*, **5,** 209.
Vessot, R. F. C, Levine, M. W., Mattison, E. M., Blomberg, E. L., Hoffman, T. E., Nystrom, G. U., Farrel, B. F., Decher, R., Eby, P. B., Baugher, C. R., Watts, J. W., Teuber, D. L., Wills, F. D. 1980, *Phys. Rev. Lett.* **45,** 2081.
Vivekanand, M., Narayan, R. 1981, *J. Astrophys. Astr.*, **2,** 315.
Wagoner, R. V. 1984, *Astrophys. J.*, **278,** 345.
Weisberg, J. M. 1978, *Ph D thesis*, University of Iowa.
Zwicky, F. 1939, *Phys. Rev.*, **55,** 726.

J. Astrophys. Astr. (1984) **5,** 209–233

Models for the Formation of Binary and Millisecond Radio Pulsars

Edward P. J. Van den Heuvel *Astronomical Institute, University of Amsterdam, Roetersstraat 15, 1018WB Amsterdam, The Netherlands*

(*Invited article*)

Abstract. The peculiar combination of a relatively short pulse period and a relatively weak surface dipole magnetic field strength of binary radio pulsars finds a consistent explanation in terms of (i) decay of the surface dipole component of neutron-star magnetic fields on a timescale of (2–5) × 10^6 yr, in combination with (ii) spin-up of the rotation of the neutron star during a subsequent mass-transfer phase.

The four known binary radio pulsars appear to fall into two different categories. Two of them, PSR 0655 + 64 and PSR 1913 + 16, have short orbital periods (< 25 h) and high mass functions, indicating companion masses $\gtrsim 0.7\ M_\odot$ (~ 1 (± 0.3) $M_\odot$ and 1.4 $M_\odot$, respectively). The other two, PSR 0820 + 02 and PSR 1953 + 29, have long orbital periods ($\gtrsim 117$d), nearly circular orbits, and low, almost identical mass functions of about $3 \times 10^{-3}\ M_\odot$, suggesting companion masses of about 0.3 $M_\odot$. It is pointed out that these two classes of systems are expected to be formed by the later evolution of binaries consisting of a neutron star and a normal companion star, in which the companion was (considerably) more massive than the neutron star, or less massive than the neutron star, respectively. In the first case the companion of the neutron star in the final system will be a massive white dwarf, in a circular orbit, or a neutron star in an eccentric orbit. In the second case the final companion to the neutron star will be a low-mass (~ 0.3 $M_\odot$) helium white dwarf in a wide and nearly circular orbit.

In systems of the second type the neutron star was most probably formed by the accretion-induced collapse of a white dwarf. This explains in a natural way why PSR 1953 + 29 has a millisecond rotation period and PSR 0820 + 02 has not.

Among the binary models proposed for the formation of the 1.5-millisecond pulsar, the only ones that appear to be viable are those in which the companion disappeared by coalescence with the neutron star. In such models the companion may have been a red dwarf of mass $\gtrsim 0.03\ M_\odot$, a neutron star, or a massive (> 0.7 $M_\odot$) white dwarf. Only in the last-mentioned case is a position of the pulsar close to the galactic plane a natural consequence. In the first-mentioned case the progenitor system most probably was a cataclysmic-variable binary in which the white dwarf collapsed by accretion.

Key words: millisecond pulsars—binaries—spin-up—accretion—magnetic field decay

1. Introduction

Four binary radio pulsars and one single millisecond pulsar are known, as listed in Table 1. One of the binary radio pulsars, PSR 1953 + 29 is also a millisecond pulsar. Three of the binary radio pulsars as well as the single millisecond pulsar differ from the majority of the radio pulsars in having an unusually short pulse period P in combination with a very low spindown rate $\dot{p}$.

This indicates that they have an unusually weak surface dipole magnetic field, as the strength B_s of this field is expected to be proportional to $(\dot{P}P)^{1/2}$ (see Section 2.1). As a result they occupy peculiar positions in the B_s *vs* P diagram—and its equivalent: the $\dot{p}$ *vs* P diagram—of radio pulsars, as can be seen in Fig. 1. This figure shows that these four pulsars have $B_s \leqslant 8 \times 10^{10}$ G, whereas the bulk of the radio pulsars have $B_S = 3 \times 10^{11}$–3×10^{13} G. The one remaining binary pulsar, PSR 0820 + 02 also has a B_s=value considerably below the average although it is still situated just within the region occupied by the majority of the radio pulsars. The combination of a short rotation period and a weak surface magnetic field is, on theoretical grounds, expected to be excluded for newborn neutron stars (Flowers & Ruderman 1977; Srinivasan & van den Heuvel 1982). We will therefore not consider models which produce this combination by direct core collapse of a single star (Arons 1983; Brecher & Chanmugam 1983; Pacini 1983).

A model in which the peculiar (P, B_S) combination of PSR 1913 + 16 is linked to the evolutionary history of its binary system was proposed by Smarr & Blandford (1976) and further worked out by Srinivasan & van den Heuvel (1978, 1982), Radhakrishnan & Srinivasan (1981), Sutantyo (1981) and Alpar *et al.* (1982). In this model the neutron star is already fairly old ($\geqslant 10^7$ yr), but later in its life it underwent spin-up by accretion, when its companion evolved to the giant stage. At this moment the surface magnetic field of the old neutron star had already partly decayed and the result of the accretion was a rapidly rotating neutron star with an unusually weak surface dipole magnetic field. After the end of the nuclear evolution of the companion, this 'recycled' neutron star became observable as a radio pulsar in a binary. This explanation for the origin of PSR 1913 + 16 was considerably strengthened by the discovery of the second binary radio pulsar PSR 0655 + 64 which exhibits a similarly abnormal (P, B_S) combination (Damashek *et al.* 1982; Taylor 1981; Blandford & DeCampli 1981; van den Heuvel 1981). This led Radhakrishnan & Srinivasan (1981) to suggest that there is an entire class of 'recycled' radio pulsars (see also Damashek *et al.* 1982).They suggested that in the case of single radio pulsars with a peculiar (P, B_S) combination, the systems had been disrupted in the second supernova explosion. After the discovery of the 1.5-ms pulsar the existence of a class of 'recycled' pulsars was independently suggested by Alpar *et al* (1982). These authors, as well as Radhakrishnan & Srinivasan (1982), Fabian *et al.* (1983) and Henrichs & van den Heuvel (1983) have suggested a variety of ways in which a companion in a binary might disappear, such that a single millisecond pulsar with a weak B_s may remain. The emphasis in this paper will be on a critical discussion of the binary recycling models for the origin of the binary and millisecond pulsars. First, in Section 2, the relevant aspects of the evolution of radio pulsars and of neutron stars in binaries are summarized. In Section 3 we discuss the evolutionary histories of the four binary radio pulsars. In Section 4 we discuss the various binary models proposed for the formation of the 1.5-ms pulsar. (For an earlier review, see Ruderman & Shaham 1983b).

2. The spin-history of single and binary neutron stars

2.1 *Relation between Spindown Rate and Surface Magnetic Field Strength of Radio Pulsars*

In all models for radio pulsars the energy loss rate of a rotating magnetized neutron star is of order

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \left(\frac{2R^6}{3c^2}\right) B_s^2 \Omega^4, \tag{1}$$

where B_S is the dipole strength of the magnetic field at the stellar surface, Ω is the angular velocity of rotation and R is the stellar radius (see, for example, Manchester & Taylor 1977). The expression for $\mathrm{d}E/\mathrm{d}t$ depends only weakly on the inclination i between the magnetic axis and the rotation axis. It holds for unipolar-inductor models with $i = 0°$ (Goldreich & Julian 1969) as well as for models based on emission of magnetic dipole radiation, for $i = 90°$ (Pacini 1968; Gunn & Ostriker 1970).

The rate of rotational energy loss of the neutron star is, on the other hand, equal to

$$\frac{\mathrm{d}E}{\mathrm{d}t} = I\Omega\dot{\Omega}, \tag{2}$$

where I is the moment of inertia of the neutron star. The equality of the expressions (1) and (2) yields

$$B_s = \left(\frac{3Ic^3}{8\pi^2 R^6}\right)^{1/2} (P\dot{P})^{1/2}. \tag{3}$$

The B_s-values in Fig. 1 (adapted from Radhakrishnan 1982 and Radhakrishnan & Srinivasan 1981) were derived from the observed P and $\dot{P}$ values of radio pulsars (Manchester & Taylor 1981) under the assumption that all neutron stars have the same moment of inertia $I = 10^{4.5}$ gm cm ^{2}, and the same radius, $R = 10$ km. Although in practice these assumptions will not be exactly true, Fig. 1 is still expected to give a good impression of the mean magnetic field strengths of radio pulsars. The figure shows that over 90 per cent of the radio pulsars have surface dipole magnetic field strengths between $10^{11.5}$ and $10^{13.5}$ G.

2.2 *Observational Indications of Magnetic Field Decay of Pulsars*

Proper-motion measurements for several dozens of radio pulsars indicate that pulsars tend to be born with high space velocities, on the average ~ 200 km s $^{-1}$ (Lyne 1981). The observed proper motions tend to be directed away from the galactic plane, indicating that most pulsars are born close to this plane, presumably as evolutionary products of massive stars. The proper motion together with the distance to the galactic plane yield the 'kinematic age' τ_k of the pulsar, which will be a good measure of the true age of the pulsar. It appears that (*cf.* Manchester & Taylor 1977; Lyne 1981) (a) practically all pulsars have $\tau_k < 10^7$ yr, indicating that pulsars turn off on about this timescale; (b) for kinematic ages $\leqslant 5 \times 10^6$ yr the spin-down age $\tau_{sd} = P/2\dot{P}$ is similar to the (real) kinematic age; for $\tau_k > 5 \times 10^6$ yr, the spin-down ages become systematically (much) larger than the kinematic ages.

The latter observations suggest that for true ages > 5×10^6 yr, the spin-down rates $\dot{P}$ decrease much faster than would be expected, according to Equation (3), if the surface dipole magnetic field strength of the pulsar remained constant. The most straightforward interpretation for the rapid drop in P for ages > 5×10^6 yr is that the B_s-values of pulsars decay on a timescale of order (2–5) $\times$ 10^6 yr (Lyne 1981). The alternative suggestion that the decrease of $\dot{P}$ is due to alignment of the rotation axis and the magnetic axis is less likely for a variety of theoretical and observational reasons. (One of the main reasons is that aligned dipoles are expected, according to the Goldreich–Julian model, to lose rotational energy at practically the same rate as perpendicular ones; further, see Taam & van den Heuvel 1984.)That the timescale of field-decay is of the order of (2–4) $\times$ 10^6 yr can also be inferred from Fig. 1 where evolutionary tracks of pulsars with an exponentially decaying surface dipole field are drawn, for $\tau_d \sim 2 \times 10^6$ yr, and Fig. 2 where similar tracks are also given for $\tau_d \sim 4 \times 10_6$ yr. The tracks were calculated under the assumption that radio pulsars are born in the upper left-hand part

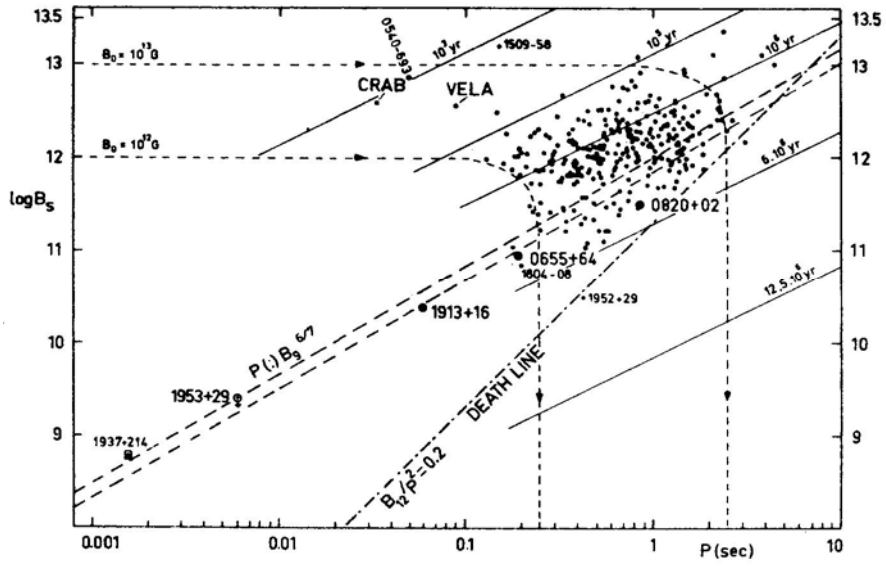


Figure 1. Surface dipole magnetic field strength B_s *vs* pulse period P for over 300 radio pulsars, as derived from their observed P values. Data for single pulsars are from Manchester & Taylor (1981), Seward, Harnden & Helfand (1984) and Backer (1984). Data for the binary pulsars (circled dots) are from the references listed in Table 1. The B_s-value indicated for PSR 1953 + 29 is the theoretical upper limit (*cf.* Helfand, Ruderman & Shaham 1983). The 1.5-ms pulsar is indicated by a square.

Evolutionary tracks of radio pulsars for two initial B_s-values are indicated (small-dash lines) for an assumed exponential-decay timescale of the field strength of 2×10^6 yr. Corresponding timelines indicate—for this decay timescale—the pulsar positions expected at various ages. The dash-dotted line is the Ruderman-Sutherland (1975) deathline beyond which the pulsar-emission presumably ceases.

Heavily dashed lines indicate the shortest possible rotation periods that can be reached during spin-up by accretion in a binary; upper line corresponds to $\dot{M} = \dot{M}_{Edd}$, and lower one to $\dot{M} = 0.5\ \dot{M}_{Edd}$. Radio pulsars that underwent spin-up by accretion are expected to be found in the wedge-shaped region between these lines and the deathline (based on Figs 12 and 13 of Radhakrishnan 1982; see also Radhakrishnan & Srinivasan 1981 and Alpar *et al.* 1982).

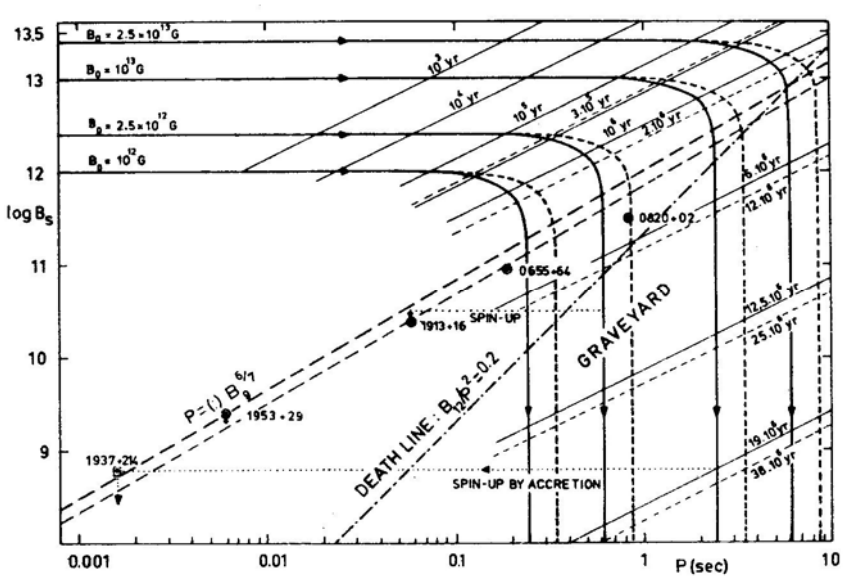


Figure 2. Theoretical evolutionary tracks and timelines of radio pulsars in the B_s *vs* P diagram for magnetic field decay timescales of 2×10^6 yr (fully drawn curves) and 4×10^6 yr (small-dash curves). Dotted lines indicate two possible spin-up tracks which old neutron stars from the 'graveyard' may follow during heavy accretion in a binary system. The meaning of the other lines is the same as in Fig. 1. The positions of the binary pulsars strongly suggest that they are old neutron stars that were 'recycled' by accretion in their binary systems.

of the figure, with a short period and a relatively strong field. When they are young, their large energy loss rate implied by Equation (1) causes them to move rapidly towards the right along a horizontal track. When their age exceeds a few million years, the ensuing field decay causes the tracks to curve downwards, becoming nearly vertical in the end.

For pulsars which were born at the same time in the upper left-hand part of the diagram the 'timelines' of 10^3 yr, 10^4 yr, *etc.* indicate the positions at these respective ages.

2.3 *Turn-off of Radio Pulsars: The* '*Graveyard*'

Fig. 1 shows that at the right-hand side of a certain inclined 'death line' in the B_s *vs* P diagram no radio pulsars are found. This in combination with the fact that hardly any pulsars have true ages $> 10^7$ yr, suggests that when a pulsar on its evolutionary track crosses the 'death-line' the emission of its pulsed radiation turns off. Various pulsar theories give various equations for the inclination of the death-line. The dash-dotted line in Figs 1 and 2 is the Ruderman–Sutherland (1975) deathline, which appears to give a fair representation of the observations.

Fig. 1 suggests that with the exception of a few peculiar objects such as the binary and millisecond radio pulsars, all radio pulsars are born with a relatively short period and a relatively strong surface dipole magnetic field, in the range $10^{12-13\,5}$ G. Notice that (i) the pulsars with the highest field strengths, of order $10^{13\cdot5}$ G (i.e., PSR 1509 – 58) pass the deathline already within (1–2) $\times$ 10^6 yr after they were born, (ii) those born with a field

strength ~ 10^{12} G may take ~ 10^7 yr to reach the deathline, and (iii) many neutron stars in the graveyard may have appreciable magnetic field strengths, of order 10^{11}–10^{13} G.

2.4 *Spin-up by Accretion in a Binary System: Predicted and Observed Pulse Periods of 'Recycled' Old Neutron Stars.*

A neutron star which is born in a binary system with a normal non-degenerate star as a companion is not expected to be observable as a radio pulsar, as even a very tenuous wind or corona will disperse the pulsed signal beyond detectability. In a later stage of its evolution, when the neutron star is already in the 'graveyard', the evolutionary expansion of the companion may lead to mass transfer, causing the system to become observable as a pulsating X-ray binary.

In many of the pulsating X-ray binaries the pulse period is observed to be continuously decreasing on a relatively short timescale, of order 10^3–10^5 yr (*cf.* Rappaport & Joss 1983). This holds especially for systems in which there is clear evidence for the presence of an accretion disc, *e.g.* Her X–1, Cen X–3 and SMC X–l (van Paradijs 1983). This spin-up moves the pulsar back from the 'graveyard' into the region of 'living' pulsars in the B_s *vs* P diagram. The two dotted horizontal lines in Fig. 2 depict examples of such spin-up tracks for neutron stars that underwent considerable field decay before the onset of the X-ray phase. That spin-up by accretion may indeed produce very short spin periods is demonstrated by the systems of SMC X–l and A 0538 — 66 which presently have spin periods of 0.71 s and 0.069 s, respectively, which are still decreasing.

For a given accretion rate $\dot{M}$ the spin-up will end when the neutron star reaches its so-called 'equilibrium' spin period P_{eq} given by (*cf.* van den Heuvel 1977; Henrichs 1983)

$$P_{eq} = (2.4\ \mathrm{ms})\ (B_9)^{6/7}\, M^{-5/7}\, (\dot{M}/\dot{M}_{Edd})^{-3/7}\, R_6^{15/7}, \tag{4}$$

where B_9 , M and R_6 are the surface dipole magnetic field strength of the neutron star in units of 10^9 G, its mass in solar masses, and its radius in units of 10^6 cm, respectively. $\dot{M}_{Edd}$ is the maximum possible 'Eddington-limit' accretion rate ~ 10^{-8} $M_\odot$ yr^{-1}. (For $P < P_{eq,}$ matter cannot enter the magnetosphere as it would be swung out again by centrifugal forces.) Equation (4) shows, that for a 'standard' neutron star with M= 1, R_6= 1, the shortest possible spin-period P_{min} that can be reached—for $\dot{M}$=$\dot{M}_{Edd}$—depends only on the value of B_9 , as $P_{min} \sim B_9^{6/7}$. This relation is indicated in Figs 1 and 2 by the upper of the two heavily dashed lines (the lower one corresponds to $\dot{M} = 0.5\ \dot{M}_{Edd}$).

All radio pulsars that originated from spin-up in binary systems are, in these figures, expected to be found in the wedge-shaped region between this line and the 'deathline', as was pointed out by Radhakrishnan & Srinivasan (1981; see also Radhakrishnan 1982) and by Alpar *et al.* (1982).

It is significant that the 1.5-ms radio pulsar as well as the three binary radio pulsars with well-determined $\dot{P}$ values are indeed situated precisely in this predicted region. This gives strong support to the idea that they are old neutron stars that obtained their short pulse periods by spin-up during a preceding mass-transfer phase (Alpar *et al.* 1982). For the one remaining binary pulsar PSR 1953 + 29 presently only an upper limit

to $\dot{P}$ is known, leading to an upper limit on the B_s-value. The binary-recycling model predicts that its B_s is smaller than 2.5×10^9 G, corresponding to a $\dot{P} \leqslant 10^{-17}$ (Helfand, Ruderman & Shaham 1983). This prediction can be tested in the near future.

2.5 *Conclusions*

We conclude the following from the foregoing sections.

(1) The observed positions of the binary radio pulsars in the B_s *vs* P diagram (or the P *vs* $\dot{P}$ diagram) strongly support the idea that they are relatively old neutron stars that have been spun up by accretion at a time when their surface dipole magnetic fields had already undergone substantial decay.

(2) During the preceding evolution the mass-transfer rate reached a value close to (or possibly larger than) the Eddington limit.

(3) The companions to these pulsars must already be highly evolved objects near the end of their nuclear evolution (as they are already beyond the stage of overflowing their Roche lobes, and presently are not producing any dispersion of the pulsar signals). This means that they are likely to be white dwarfs or neutron stars (or possibly, black holes).

(4) The position of the 1.5-ms pulsar in the B_s *vs* P diagram suggests that it may also have undergone spin-up in a binary system.

3. Evolutionary models for the binary radio pulsars

3.1 *The Two Types of Binary Radio Pulsars*

The four binary radio pulsars seem to fall into two different categories (see Table 1 and Fig. 3). Two of them, PSR 0655 + 64 and PSR 1913 + 16, have short orbital periods (< 25 h) and high mass functions, indicating companion masses $> 0.7\ M_\odot$ (respectively, $1(\pm 0.3)\ M_\odot$ and $1.4\ M_\odot$: Damashek *et al.* 1982; Taylor & Weisberg 1982). The other two, PSR 0820 + 02 and PSR 1953 + 29, have long orbital periods (117 d), nearly circular orbits, and low, almost identical mass functions of about $3 \times 10^{-3}\ M_\odot$, suggesting companion masses around $0.2\ M_\odot$ to $0.4\ M_\odot$ (Taylor 1981; Manchester *et al.* 1983; Boriakoff, Buccheri & Fauci 1983). It was pointed out by van den Heuvel & Taam (1984) that these two classes of systems are expected to be formed by the later evolution of binaries consisting of a neutron star and a normal companion star, in which the companion was (considerably) more massive than the neutron star, or less massive than the neutron star, respectively. We will consider these two cases separately.

3.2 *Evolution of a Binary Consisting of a Neutron Star and a More Massive Normal Companion: the Origin of PSR 0655 + 64 and PSR 1913 + 16*

When the companion has a mass several times that of the neutron star *(i.e.* $\geqslant 3$–$4\ M_\odot$, see below) and the system is relatively wide *(i.e.* orbital period $\geqslant$ a few weeks) so that at the onset of the mass transfer the companion is a giant with a deep convective envelope, runaway mass transfer is unavoidable. This is because (i) mass transfer from the more massive to the less massive star leads to shrinking of the Roche-lobe of the more

Table 1. Some important properties of the four binary radio pulsars and the single millisecond pulsar, together with estimates of their surface magnetic field strengths and of the masses of the companions in the binary systems.

Name	P_{orb} (d)	e	Mass function ($M_\odot$)	Most likely companion mass ($M_\odot$)	P_{pulse} (s)	B_s (G)	Ref.
PSR 1913+16	0.32	0.617	0.1322	1.40 ± 0.05	0.059	2×10^{10}	(1)
PSR 0655+64	1.03	0.000	0.0712	1.00 ± 0.30	0.196	8.6×10^{10}	(2)
PSR 0820+02	1232	0.012	0.00301	0.2–0.4	0.865	3.3×10^{11}	(3)
PSR 1953+29	~117	<0.01	0.00272	0.2–0.4	0.0061	2.5×10^{9} (predicted)	(4, 5, 6, 7)
PSR 1937+214					0.00155	5×10^{8}	(8)

References:

(1) Taylor & Weisberg (1982)
(2) Damashek *et al.* (1982)
(3) Manchester *et al.* (1983)
(4) Boriakoff, Buccheri & Fauci (1983)
(5) Buccheri (1983, personal communication)
(6) Helfand, Ruderman & Shaham (1983)
(7) This paper, Section 3
(8) Backer (1984)

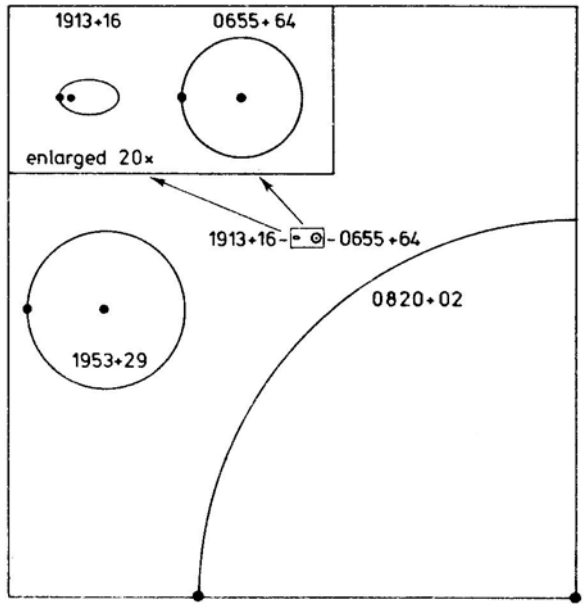


Figure 3. Relative orbital dimensions of the four radio pulsar binaries suggest that there may be two classes of systems: PSR 0655 + 64 and PSR 1913 + 16 have orbital dimensions that are over 25 times smaller than those of PSR 1953 + 29 and PSR 0820 + 02. On the other hand, the former two pulsars have relatively massive companions (~ 0.7–1.4 $M_\odot$), whereas the latter two have low-mass companions (0.2–0.4 $M_\odot$).

massive star, while (ii) a convective envelope tends to expand when the star loses mass (Paczyński & Sienkiewicz 1971). This will, on a timescale of ~ 10^3 yr lead to the formation of an extended convective common envelope in which the neutron star and the evolved core of the companion will spiral towards each other as a consequence of the large frictional drag (Paczyński 1976; Webbink 1979; Taam, Bodenheimer & Ostriker 1978; Meyer & Meyer-Hofmeister 1979). The duration of this spiral-in process, in which finally the envelope is lost, is expected to be short, *i.e.* ~ 10^3–10^5 yr. The final result is expected to be a system consisting of the neutron star and the evolved core of the companion with an orbital period ranging from about one hour to a few days (in analogy to the case of cataclysmic variables which are products of a similar type of spiral-in evolution: Paczynski 1976, Meyer & Meyer-Hofmeister 1979). In order to have deep spiral-in the two components should differ sufficiently in mass, preferably by more than a factor of two, since otherwise soon after the onset of the mass transfer the mass ratio will be reversed which would lead to subsequent spiral-out. (For systems with mass ratios in the range 0.85 to 2 the outcome of the evolution is more complex and, for the sake of argument, will not be considered here.) Fig. 4 depicts as an example the anticipated evolution of a binary initially consisting of a $M \simeq 5M_\odot$ star and a neutron star, with an orbital period of $\geqslant 80$d. The 5 $M_\odot$ star overflows its Roche lobe when it has exhausted helium in its core [i.e. when it climbs up along the asymptotic giant branch (AGB)]. At helium exhaustion it has a 0.95 $M_\odot$ degenerate CO core (Paczyński 1970) surrounded by He- and H-burning shells. During its ascent of the AGB this core mass increases to 1.39 $M_\odot$. We assume that the star engulfs its

companion when $M_{\rm core} \simeq 1.0\ M_\odot$, implying a binary period of about 100 d (see Table 2). The system after spiral-in will consist of a 1.0 $M_\odot$ CO white dwarf and a neutron star with a narrow and circular orbit, *i.e.* closely resembling the PSR 0655 + 64 system (see Table 1). The type of binary evolution considered here is so-called case C (Plavec 1968; Paczyński 1971). [In shorter-period systems—case B—the evolution will be somewhat more complex and may lead to several stages of mass transfer (Habets 1984, Delgado & Thomas 1981). The outcome will, however, in most cases be roughly similar.] Table 2 lists the maximum and minimum orbital periods for case C evolution for companion masses in the range 3 to 7 $M_\odot$, derived from Paczyński's evolutionary tracks. The corresponding CO-core masses are also listed.

Table 2 shows that an evolution similar to the one depicted in Fig. 4 is expected in all wide (case C) neutron-star binaries in which the mass of the companion is in the range ~ 5–8 $M_\odot$ as on the AGB all such stars develop degenerate CO cores with a mass in the range 0.95–1.39 $M_\odot$ (Paczynski 1970). For companion masses 3–5 $M_\odot$ the same holds if the binary periods are $\geqslant$ 80 d, such that after helium exhaustion the core still can grow to $\geqslant 0.90\ M_\odot$ before the star overflows its critical lobe.

Companions more massive than ~ 8–10 $M_\odot$ will develop evolved cores too massive to terminate as white dwarfs (van den Heuvel 1981; Habets 1984). Following the spiral-in and subsequent nuclear evolution such a core is expected to collapse to a neutron star. The mass ejection in this second supernova in the system may induce a considerable orbital eccentricity or may even disrupt the system. PSR 1913 + 16 is expected to have been formed in this way (Smarr & Blandford 1976, Srinivasan & van den Heuvel 1982). (In the case of disruption, two runaway neutron stars will be formed, one young and one old.)

At the onset of the spiral-in the neutron star will often be older than 10^7 yr (companion stars of 5–15 $M_\odot$ need some (1–5) × 10^7 yr to leave the main sequence) such that its surface magnetic field strength may have decayed by several orders of magnitude.

During the spiral-in the accretion onto the neutron star will take place near the Eddington-limited rate $\dot{M}_{\rm Edd}$, implying that its rotation will be spun up to a minimum period that according to Equation (4) depends practically only on the surface dipole magnetic field strength B_9 . After the loss of the common envelope these neutron stars will become observable as radio pulsars with a relatively low surface dipole magnetic

Table 2. Minimum and maximum orbital periods for neutron star binaries in which the companion overflows its critical lobe after it has exhausted helium in its core ('case C' evolution). A neutron star mass of 1.40 $M_\odot$ was assumed and the corresponding masses $M_{\rm core}$ of the degenerate CO-core of the companion are indicated (evolutionary tracks of core Paczynski 1970 were used). The first value of $P_{\rm max}$ corresponds to the moment at which the orbital separation equals the radius of the companion; the value within parenthesis corresponds to the moment at which the companion fills its Roche lobe.

Companion mass ($M_\odot$)	$P_{\rm min}$ (d)	$M_{\rm core}$ (He exhaustion) ($M_\odot$)	$P_{\rm max}$ (yr)	$M_{\rm core}$ (max) ($M_\odot$)
3	32	0.51	5.5 (19)	1.39
5	76	0.95	4.5 (13)	1.39
7	173	1.02	3.6 (10)	1.39

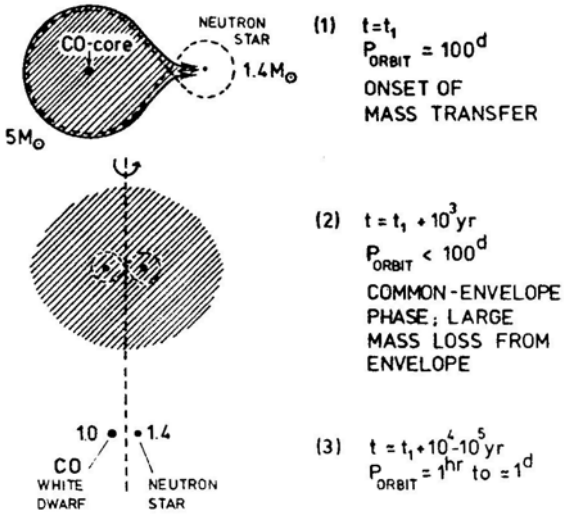


Figure 4. Anticipated evolution of a relatively wide binary consisting of a 5 $M_\odot$ star with a neutron star companion. At the onset of the mass transfer the 5 $M_\odot$ star is a giant with a 1 $M_\odot$ degenerate CO core and a gradually expanding envelope (1). Rapidly, a common convective envelope forms (2) in which the CO-core and the neutron star spiral-in. During this spiral-in the envelope is ejected due to frictional heating and a very close system remains (3) consisting of a neutron star and a massive CO white dwarf in a circular orbit. All wide (case C) neutron star binaries with companions in the range 3 $M_\odot$ to 8–10 $M_\odot$ are expected to evolve in this way. For companions more massive than 8–10 $M_\odot$ the evolved core after spiral-in collapses to a neutron star, and a close neutron star binary with an eccentric orbit (or two runaway pulsars) will be formed. (© Nature).

field strength. For the observed B_s-values of PSR 0655 + 64 and PSR 1913 + 16 (Table 1), and $R_6 = 1$, $M = 1.4\ M_\odot$, Equation (4) with $\dot{M}_{\rm Edd}$ yields $P = 0.12$ and 0.04 s, respectively, in fair agreement with their observed pulse periods (allowing for some spin-down since their formation).

3.3 *Reason for the Absence of Radio Pulses from the Companion of PSR 1913 + 16; Types of Progenitors*

The absence of radio pulses from the second (younger) neutron star in the PSR 1913 + 16 system may either be due to beaming effects, or to the relatively short spindown timescale of a newborn strong-magnetic-field neutron star. If the second supernova in the system took place a few million years ago, the new neutron star may already have reached the turn-off period of a few seconds (see Fig. 1) whereas the 'recycled' old one, due to its weaker surface dipole field, will remain observable as a pulsar for a much longer time (van den Heuvel & Taam 1984). The best-known progenitor candidates for the above-described type of evolution seem to be the B-emission X-ray binaries, which have orbital periods ranging from 15 days to several years (Rappaport & van den Heuvel 1982), and companion masses ranging from ~ 7 $M_\odot$ (spectral type B3Ve) to ~ 20 $M_\odot$ (spectral type O9Ve). The companion of PSR 0655 + 64 should have had a mass below the limit for exploding as a supernova, *i.e.* $\leqslant$ 8–10 $M_\odot$ (in case C; in case B: $\leqslant$10–12 $M_\odot$) whereas that of PSR 1913 + 16 had a mass above this limit.

3.4 *Evolution of a Binary Consisting of a Neutron Star and a Less Massive Normal Companion: Origin of PSR 1953 + 29 and PSR 0820 + 02*

Consider a system in which the main-sequence companion star is less massive than the neutron star. When in this system (mass ratio $\leqslant 0.85$) the companion has evolved to the giant stage and begins to overflow its Roche lobe, the ensuing mass transfer will be self-stabilizing as it leads to expansion of the orbit. Webbink, Rappaport & Savonije (1983) and Taam (1983) have shown that in this way, starting with companions with $M \leqslant 1.2\ M_\odot$ and orbital periods upwards from about 1 d, one obtains a long-lasting ($\sim 10_6$–10_8 yr) stage of relatively high mass transfer ($\dot{M} \geqslant 10^{-8}\ M_\odot\ \mathrm{yr}^{-1}$). This explains the existence of several wide low-mass X-ray binaries such as Cygnus X-2 (P_{orb} = 9.8d), 2S 0921 – 63 (P_{orb} = 9.0d) and GX 1+4 (P_{orb} at least several months, see below). During the mass-transfer phase the companion ascends the giant branch and has a degenerate helium core with a mass in the range 0.2–0.45 $M_\odot$, and generates energy by burning hydrogen in a shell around this core. The mass transfer, and the associated expansion of the orbit, are driven by the gradual growth of the core mass. The duration of the mass-transfer phase depends mainly on the initial mass of the companion and the initial orbital period. In the end, only the degenerate helium core of the companion is left behind, as a helium white dwarf, and the final orbit will always be wide, since the mass and orbital angular momentum of the system are, in first approximation, expected to be conserved. Fig. 5 depicts as an example, the evolution of a system that started out with a 1.0 $M_\odot$ star together with a 1.3 $M_\odot$ neutron star companion with an initial orbital period P_0 = 12.5 d, as calculated by Joss & Rappaport (1983).

Savonije (1983a), Paczyński (1983), and Joss & Rappaport (1983) have independently pointed out that the final systems resulting from this type of evolution show a striking resemblance to that of the 6-ms binary radio pulsar PSR 1953 + 29, since (i) the final companion mass will be low ($\sim 0.3\ M_\odot$),and (ii) the orbit will be (almost) circular as a consequence of the long-lasting, preceding mass-transfer stage. (When the companion is a red giant which is filling its Roche lobe, tidal dissipation in its convective envelope will be very efficient (Zahn 1978; Savonije 1983b). Fig. 5 shows that the most likely initial orbital period of PSR 1953 + 29 was around 12.5 d; similarly Joss & Rappaport (1983) showed that the initial orbital period of PSR 0820 + 02 was about 1 yr. The duration of the mass-transfer phase in these systems was (3.5–7.7) $\times 10^7$ yr and 4×10^6 yr, respectively.

3.5 *Arguments for Accretion-Induced Collapse in Wide Radio Pulsar Binaries*

Helfand, Rudermann & Shaham (1983) have suggested that PSR 1953 + 29 was formed by the accretion-induced collapse of a massive white dwarf, during a mass-transfer phase in its progenitor system. Their main argument was that if the neutron star had been formed by direct core collapse in a binary with a low-mass companion, the system would have received a runaway velocity $\geqslant 100$ km s $^{-1}$, which during its lifetime would have carried it to a distance of several kiloparsecs from the galactic plane. In this case the position of the system close to the galactic plane cannot be understood. On the other hand, accretion-induced collapse need not impart a runaway velocity $\geqslant 10$ km s^{-1} to the system (see below).

Van den Heuvel & Taam (1984) have put forward the following strong additional arguments for formation by accretion-induced collapse of the neutron stars in both of the wide radiopulsar binaries. The evolutionary model for these wide systems outlined in the foregoing section appears to explain their observed orbital characteristics, including their mass functions; but this model at the same time implies that the systems must be very old (at least several times 10^9 yr) since the companion stars should have started out with masses $\leqslant 1.2\ M_\odot$ and had already completed their main-sequence evolution before the onset of the mass transfer. Since surface dipole components of magnetic fields of neutron stars appear to decay on a timescale $\leqslant 10^7$ yr (see above) one would therefore not expect any detectable surface dipole fields to be left in them. However, PSR 0820 + 02 and PSR 1953 + 29 still have a surface dipole magnetic field strength 3×10^{11} G and $> 10^9$ G, respectively (Table 1). The same problem exists in the GX 1 + 4 system which is a 120-s X-ray pulsar, and must be a very wide system as the companion star is an M6IIIe red giant (Bradt & McClintock 1983). Such a system seems an excellent progenitor candidate for a system like PSR 0820 + 02 (van den Heuvel 1981). From its spin-up rate, in combination with its X-ray luminosity of $\sim 10^{38}$ ergs^{-1} (Bradt & McClintock 1983), one derives a lower limit to its magnetic field strength of $\sim 3 \times 10^{11}$ G (Henrichs 1983) *i.e.* similar to that of PSR 0820 + 02. Since an M6IIIe giant cannot be more massive than about 3 $M_\odot$, the age of this sytem is at least $\geqslant 2 \times 10^8$ yr (the main-sequence lifetime of a 3 $M_\odot$ star). Moreover, the position close to the galactic centre and at high z suggests an old disc population with an age of $(5–10) \times 10^{10}$ yr.

The only consistent way in which one may explain why the neutron stars in these old (wide) systems still can have such fairly strong surface magnetic fields seems to be that they were formed only recently, *i.e.*, during the mass-transfer stage itself, by the accretion-induced collapse of a (relatively massive) white dwarf. (It seems reasonable to assume that, regardless of the formation mechanism, all neutron stars are formed with a 'canonical' surface dipole magnetic field strength of $10^{12–13\,5}$ G, due to dynamo action during the collapse process; *cf.*, Flowers & Ruderman 1977). Such a model gives an excellent explanation for the fact that PSR 1953 + 29 has a 6 ms rotation period and PSR 0820 + 02 has not. Because, in the progenitor of the. PSR 1953 + 29 system the mass transfer lasted $\sim 8 \times 10^7$ yr (see Fig. 5) such that the collapse can have occurred several times 10^7 yr before the end of the mass-transfer phase. Consequently, the magnetic field had time to decay to $< 2 \times 10^9$ G and, according to Equation (4) spin-up to ~ 5 ms was possible. On the other hand, in view of the much shorter (4×10^6 yr) mass-transfer timescale in the progenitor of the PSR 0820 + 02 system, such a field decay and spin-up were not possible here.

That PSR 0820 + 02 indeed still has the strongest magnetic field of the two and the largest pulse period is therefore in excellent agreement with the prediction of this accretion-induced collapse model. In addition, the conditions for achieving accretion-induced collapse are expected to be favourable just in systems with a low-mass ($\leqslant 1.2\ M_\odot$) giant component for the following reasons. In order to reach collapse, the white dwarf (consisting of CO or O–Ne–Mg) should be able to grow substantially by accretion—as the *a priori* probability that it was born with a mass close to the Chandrasekhar limit will be very low. Just at accretion rates produced by low-mass giant companions (*i.e.* in the range $10^{-8}\ M_\odot$ yr^{-1} to $2 \times 10^{-7}\ M_\odot$ yr^{-1}, see above) a considerable growth of the white dwarf is indeed possible since the accreted hydrogen burns in shell flashes that are too weak to cause mass ejection (Sion, Acierno &

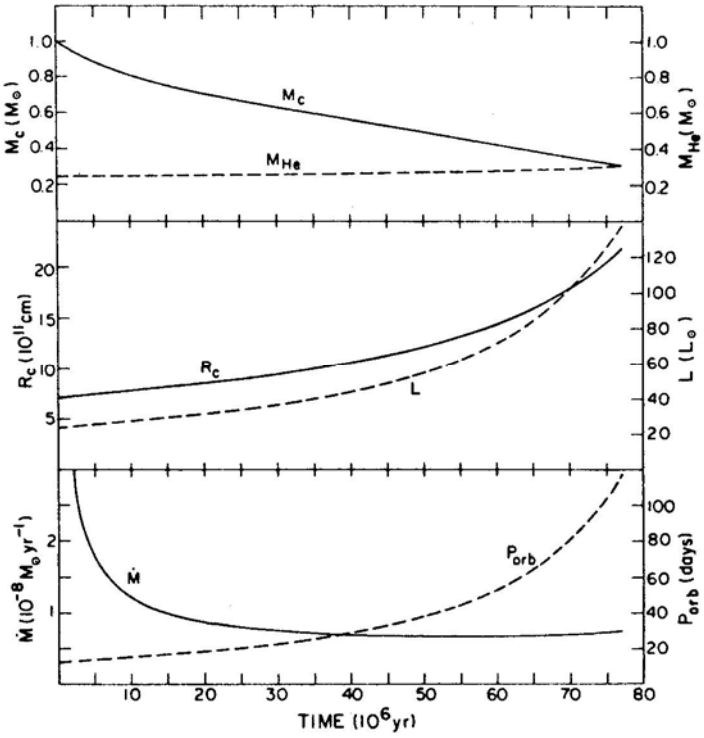


Figure 5. Evolution of a binary with a lower giant-branch secondary component of initial mass 1.0 $M_{\odot}$ and surface composition $X = 0.70$, $Z = 0.02$, as calculated by Joss & Rappaport (1983). Plots as functions of time: upper, total mass (M_c) and core mass of the secondary; middle, radius and intrinsic bolometric luminosity of the secondary; lower, mass accretion rate M onto the neutron star, and orbital period. Conservative mass transfer has been assumed. The initial mass of the neutron star was taken to be 1.3 $M_{\odot}$. (© Nature).

Tomsczyk 1979; Nomoto 1980, 1982; see also, Iben & Tutukov 1984). At lower accretion rates, $< 10^{-8}\ M_{\odot}\ \mathrm{yr}^{-1}$, hydrogen burns with very strong nova-like flashes in which part (and possibly all) of the accreted matter is ejected such that a substantial growth of the white dwarf is more difficult to achieve. In addition, at the age of these systems ($\geqslant 10^9$ yr) the oxygen in a carbon-oxygen white dwarf may have separated from the carbon and have settled in the core (Stevenson 1980) creating a situation favourable for the occurrence of accretion induced collapse (Bravo *et al.* 1983).

Since, at any time during the accretion phase, the envelope mass of the white dwarf is small (more than a few hundredths of a solar mass of hydrogen would not fit inside its Roche lobe) and since the mass loss of the white dwarf during the collapse need not exceed the equivalent of the binding energy of the neutron star ($\sim 0.1\ M_{\odot}$), the orbital eccentricity and runaway velocity induced by the mass loss in the collapse need not exceed 0.05 and $\sim 10\ \mathrm{km\ s}^{-1}$, respectively (Flannery & van den Heuvel 1975) and the orbit may be subsequently circularized by the continued mass transfer from the companion and by tidal forces. The above arguments lead to the conclusion that the existence of the two binary radio pulsars with wide circular orbits and low mass

functions (and of the GX 1 + 4 system as well), provides for the first time very strong quantitative evidence that neutron stars can be formed in an old stellar population by the accretion-induced collapse of a white dwarf in a relatively wide binary ('Whelan-Iben model' 1973).

4. Binary scenarios for the formation of the 1.5-ms pulsar PSR 1937 + 214

4.1 *Introduction*

Since one of the binary radio pulsars is also a millisecond pulsar, 'recycling' of an old neutron star in a binary is a viable mechanism for producing a millisecond pulsar. In binary models, PSR 1937 + 214 could have lost its companion either (i) by *disruption* of the system due to the supernova (SN) explosion of its companion or (ii) by *coalescence* with the companion after spiral-in due to orbital angular momentum losses, for example by the emission of gravitational waves.

The first possibility appears to be unlikely for a variety of reasons (*cf.* Henrichs & van den Heuvel 1983; Arons 1983):

(1) The disruption would have imparted a runaway velocity of several hundreds of kilometers per second to the pulsar. This is because prior to the SN explosion the system is expected to have been a very close one like the progenitor of the PSR 1913 + 16 system (see Section 3) in which the components have orbital velocities of this order. With such a large runaway velocity, the position of the pulsar at only 30 pc from the galactic plane is hard to explain—unless as a pure coincidence.

(2) In a system that is massive enough to produce a second SN, there cannot have been enough time to accrete the 0.12 $M_\odot$ required to spin up the rotation to 1.5 ms (Alpar *et al.* 1982). This is because, in order to explode, the companion must have been more massive than 8–10 $M_\odot$. With such a companion the combined duration of the mass transfer (X-ray phase) and subsequent spiral-in phase will have been less than a few times 10^5 yr, being the duration of the phase of beginning of the Roche-lobe overflow plus strong stellar wind (Savonije 1983b). As the maximum possible accretion rate was $\sim 10^{-8}$ $M_\odot$ yr^{-1} the total amount accreted was at most about 0.002 $M_\odot$, *i.e.*, much smaller than the minimum amount required to spin a neutron star up to 1.5 ms.

(3) The time required for the surface magnetic field to decay to $5 \times 10_8$ G from its 'canonical' value at birth of $\sim 10^{12-13}$ G is at least 2×10^7 yr, even for the shortest possible exponential decay time. Since only stars with masses < 8 $M_\odot$ live longer than 2×10^7 yr, this excludes in fact companion stars that can have exploded as Supernovae, unless one believes in carbon-deflagration SN models for stars in the mass range 5–10 $M_\odot$. In the latter case presumably the entire companion is disrupted, but still problems (1) and (2) remain.

For these reasons we will further concentrate only on coalescence models. Coalescence models with three kinds of companions have been proposed: (i) a red dwarf or red degenerate star (Alpar *et al.* 1982; Fabian *et al* 1983; Ruderman & Shaham 1983a), (ii) a neutron star (Henrichs & van den Heuvel 1983), and (iii) a massive white dwarf (van den Heuvel & Bonsema 1984). Before discussing these, we consider some general aspects of the evolution of close mass-transfer binaries with one compact component.

4.2 *Stability of Angular-Momentum Loss-Driven-Evolution of Close Binaries*

4.2.1 *Condition for coalescence*

Consider binaries in which the mass-receiving neutron star is the more massive component. When, due to angular momentum losses, the system has shrunk sufficiently for its companion to fill its Roche lobe, mass transfer will ensue. As the mass-losing star is the less massive component the mass transfer leads to the expansion of the orbit, and (for $M_2/M_1 \leqslant 0.85$) to expansion of the Roche-lobe radius R_L (M_2). Whether or not the mass transfer will be self-stabilizing can be examined by comparing the increase of R_L with that of the radius R_s (M_2) of the secondary star when mass is transferred to the neutron star. If

$$\left|\frac{\mathrm{d}\ln R_L}{\mathrm{d}\ln M_2}\right| < \left|\frac{\mathrm{d}\ln R_s}{\mathrm{d}\ln M_2}\right|, \tag{5}$$

the radius of the star increases faster than that of its Roche lobe. In that case the mass transfer will immediately continue and have a runaway character, such that coalescence may ensue. On the other hand, when the left-hand side of inequality (5) is larger than the right-hand side, the mass transfer will be self-stabilizing: when a small amount of mass is transferred the Roche-lobe radius becomes larger than the stellar radius. Subsequently, the orbital angular momentum losses (due to gravitational radiation—GR—and other causes) will after some time have reduced R_L so much that it again equals R_S (M_2), and again transfer of a small amount of mass is sufficient to keep R_L larger than R_s. In this way, the continuous angular momentum loss drives a continuous and self-stabilizing mass transfer from the secondary to the neutron star. Following Faulkner (1971), calculations for such mass transfer in low-mass X-ray binaries were carried out by Rappaport, Joss & Webbink (1982), to which we refer for details. Stable mass transfer can occur, for example, if the companion is a red dwarf or, alternatively, a degenerate star with a mass $\leqslant 0.5\ M_\odot$ (see below).

4.2.2 *Conservative mass transfer* vs *transfer of mass towards a disc*

The two extreme possibilities for the change of the Roche-lobe radius during the mass transfer towards a more massive compact companion are (i) 'conservative' mass transfer, in which the total mass $M_1 + M_2$ and the total orbital angular momentum

$$J_{\mathrm{orb}} = M_1 M_2 \{Ga/(M_1 + M_2)\}^{1/2} \tag{6}$$

of the system are assumed to be conserved during the transfer, where a is the orbital radius and G is the gravitational constant, and (ii) mass transfer into a disc (or ring) surrounding the companion star. In the latter case the orbital angular momentum is transformed into Keplerian angular momentum of the rotating disc, leading to a much smaller expansion rate of the orbit and of the Roche lobe than in case (i).

In case (ii), the change in orbital angular momentum is given by (Kieboom & Verbunt 1981)

$$\frac{\mathrm{d}J_{\mathrm{orb}}}{\mathrm{d}t} = (a - x_L)^2 \omega \frac{\mathrm{d}M_2}{\mathrm{d}t} - x_L^2 \omega \frac{\mathrm{d}M_2}{\mathrm{d}t}, \tag{7}$$

where x_L is the distance between the centre of the mass-losing star and the first Lagrangian point L_1 and ω is the angular velocity of revolution of the system. Expressions for x_L can be found in Kieboom & Verbunt (1981) and allow one to calculate the changes in a and the Roche-lobe radius for this case, if it is assumed that the total mass of the system is conserved and J_{orb} is given by Equation (6) in which M_1 now denotes the combined mass of the compact star and the disc. For the Roche-lobe radius one can use the expression

$$R_L = 0.462\, aq^{1/3} \tag{8}$$

which holds for $q = M_2 / (M_1 + M_2) \leqslant 0.8$(Paczyński 1971).

Fig. 6 depicts as examples the resulting Roche-lobe radii $R_{L,c}$ and $R_{L,d}$ for the conservative and the disc-transfer case, respectively, as a function of M_2 for three

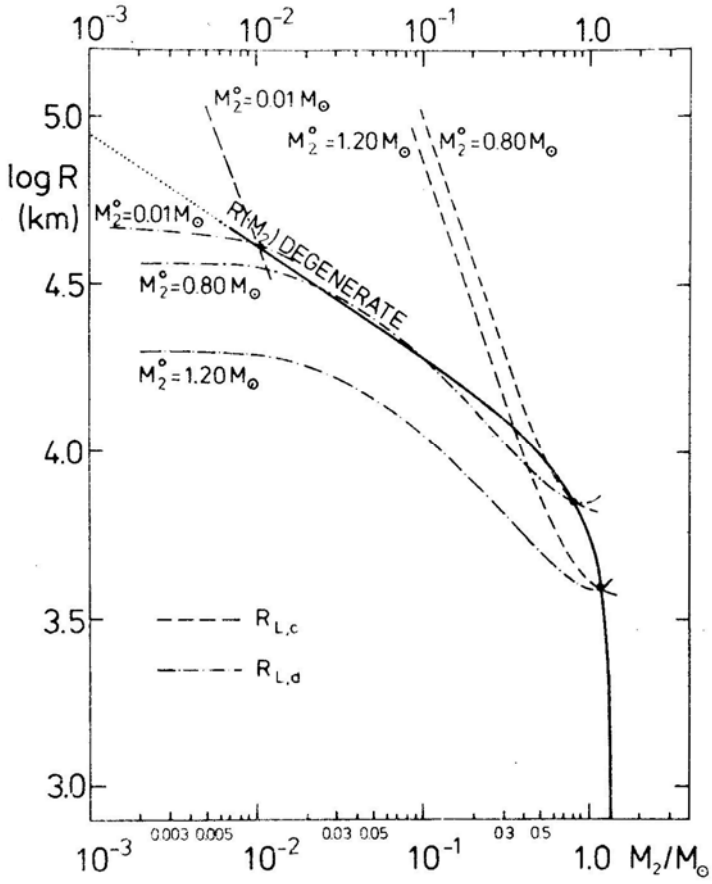


Figure 6. Critical Roche-lobe radii $R_{L,c}$ and $R_{L,d}$ for a secondary star of mass M_2 which is losing mass to a 1.4 $M_\odot$ companion, for three initial values of the secondary mass, M_2^0 = 1.2, 0.8 and 0.01 $M_\odot$. $R_{L,c}$ corresponds to 'conservative' mass transfer, $R_{L,d}$ to mass transfer into a disc around the companion. Also drawn is the mass-radius relation R_s (M_2) for degenerate stars that do not contain hydrogen (R in km). Comparison between the slopes of the curves allows one to judge whether the mass transfer will be stable or unstable (see text).

systems, with initial companion masses M_2^0 = 1.2, 0.8 and 0.01 $M_\odot$, respectively. In all cases the initial mass of the compact star was assumed to be M_1 = 1.4 $M_\odot$. Also drawn in this figure is the mass–radius relation for electron-degenerate stars that do not contain hydrogen (see next section).

4.2.3 *Validity of the 'conservative' assumption*

The assumption of 'conservative' mass transfer is expected to hold during long-lasting steady stages of mass transfer, also if the mass flows towards the compact star through an accretion disc. This is because in a steady state the total amount of mass and angular momentum in the disc will be constant (the transfer time of matter through a disc is relatively short, of the order of weeks; *cf.* Pringle 1981). Therefore, the presence of the disc does not contribute to changes in the orbital parameters. (In such a steady state, presumably the disc angular momentum is steadily fed back into the orbit, by tidal torques and/or some return mass transfer from the disc edge; *cf.* Pringle 1981.) In this case one can simply consider the companion plus disc as one object which is the companion of the mass-losing star. On the other hand, in the case that the mass transfer assumes a runaway character, or when it is just starting—such that the disc is just forming—the conservative approximation will not be valid and the Roche-lobe radius will in first instance follow one of the $R_{L,d}$ curves in Fig. 6.

4.3 *Coalescence Models*

4.3.1 *Coalescence with a red dwarf*

(a) Evolution of low-mass X-ray binaries and cataclysmic variables

Red main-sequence dwarfs with masses $\geqslant 0.1\ M_\odot$ in thermal equilibrium have a mass-radius relation

$$R_s(M_2) = k_1 M_2^n \quad \text{with} \quad 0.5 \leqslant n \leqslant 1 \tag{9}$$

where K_1 is a constant. This ensures according to condition (5) and Fig. 6 that red dwarfs will always be able to achieve stable mass transfer (at a rate $\sim 10^{-10}\ M_\odot\ \mathrm{yr}^{-1}$; *cf.* Rappaport, Joss & Webbink 1982), while the orbit is shrinking due to GR losses on a timescale

$$\tau_{GR} = \frac{(M_1 + M_2)^{1/3}}{2^{1/3} M_1 M_2} \left(\frac{P}{1.6\,\mathrm{h}} \right)^{8/3} (5 \times 10^7)\,\mathrm{yr}. \tag{10}$$

When the red dwarf reaches a mass $\leqslant 0.2\ M_\odot$, τ_{GR} becomes shorter than its thermal timescale and the mass-transfer drives the star out of thermal equilibrium, causing it to finally follow the mass-radius relation for a convective polytrope

$$R_s(M_2) = k_2 M_2^{-1/3} \tag{11}$$

where k_2 is a constant.

Furthermore, if the companion has a mass $\leqslant 0.1\ M_\odot$, or if it is hydrogen-poor, degeneracy may set in which also gives a mass–radius relation similar to Equation (11). The result is that from here on mass transfer leads to expansion of the orbit while the transfer remains self-stabilizing, as inequality (5) is still not fulfilled. This explains the

existence of a period minimum of ~ 80 min for the cataclysmic-variable (CV) binaries with red-dwarf companions (Paczynski & Sienkiewicz 1981). [The only CV binary with a period below this minimum is GP Com (P = 47 min) which most probably has a hydrogen-poor degenerate companion; *cf.* Rappaport & Joss 1984.]

Fig. 7 outlines the above-described evolution of CV binaries and low-mass X-ray binaries which started out with a mass of the red-dwarf companion $\gtrsim 0.2\ M_\odot$. One would expect the companions to spiral out until finally $\tau_{GR} \simeq \tau_{Hubble}$. At that time the orbital period is ~ 2.2 h, the companion mass is ~ 0.017 $M_\odot$ and the mass-transfer rate is ~ $10^{-12.5}$ $M_\odot$ yr^{-1} (Henrichs & van den Heuvel 1983). [If additional angular-momentum loss mechanisms are invoked, *e.g.* 'magnetic braking' (Verbunt & Zwaan 1981) the values of the period-minimum and of the final mass may be different, *cf.* Eggleton 1983.]

(*b*) *The fate of low-mass X-ray binaries: coalescence?*

The assumption in the above-described evolutionary picture is that the mass transfer takes place in a steady and continuous fashion, such that in inequality (5) the 'conservative' Roche-lobe radius $R_{L,c}$ can be used. However, Ruderman & Shaham (1983a) suggest a different picture, namely, that when the companion's mass becomes

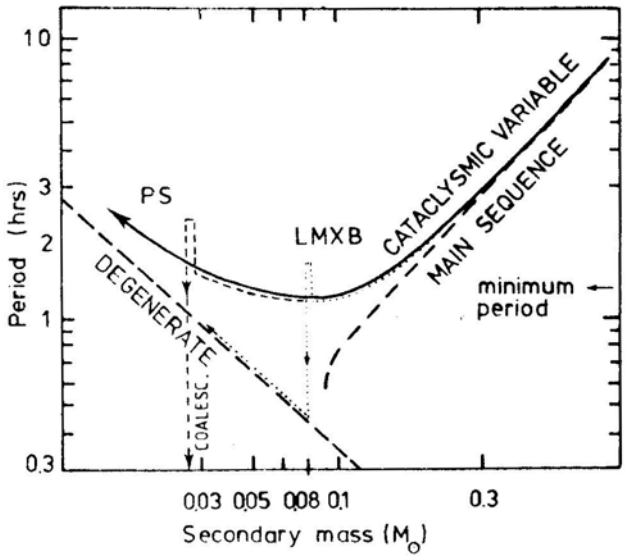


Figure 7. Schematic picture of the secular evolution of cataclysmic-variable (CV) binaries. Systems containing degenerate stars (black dwarfs or white dwarfs) would lie along the left-hand dashed line. Systems containing a (red) main-sequence companion in thermal equilibrium would lie along the dashed line on the right-hand side. The fully drawn curve is the model track of Paczynski & Sienkiewicz (1981, see explanation in the text). The dotted curve is the evolutionary track of a system in which the white dwarf collapses (due to the accretion) to a neutron star. This causes a sudden increase of the period, upon which the—now detached—companion of 0.08 $M_\odot$ shrinks to its degenerate radius. After angular momentum loss by GR, an ultra-short period low-mass X-ray binary results, like 4U 1626 – 67 (dotted track). The dashed track indicates the evolution of a similar system with M_{red} = 0.03 $M_\odot$. When the mass transfer resumes, this companion is disrupted, and a single millisecond pulsar may result (see text).

very low, mass and angular momentum will be lost from the disc edge in such a way that the Roche lobe will hardly expand, leading to runaway mass transfer and disruption of the low-mass star when its mass drops below about 0.01 $M_\odot$. An alternative possibility would be that the mass transfer gets interrupted. When, after such an interruption, the mass transfer starts again, the matter will first form a ring or disc around the neutron star, such that $R_{L,d}$ has to be used in inequality (5).

Fig. 6 shows that, for example, for $M_2^0 = 0.01\ M_\odot$, the $R_{L,d}$ curve is less steep than the R_s (M_2) curve for degenerate stars (or purely convective stars) such that the mass-transfer rate will assume a runaway character and the low-mass star will be disrupted on its dynamical timescale (a few minutes). Its matter will then form a disc which subsequently can be accreted.

In the preceding low-mass X-ray binary phase the rotation of the neutron star may already have been spun up to a millisecond period. Hence, a single millisecond pulsar would result (if the neutron star would still have a surface magnetic field; see also Section 4.4).

4.3.2 *Coalescence with a massive white dwarf*

Close binaries consisting of a neutron star and a massive (~ 1 $M_\odot$) white dwarf do exist in nature, *e.g.* PSR 0655 + 64. As pointed out in Section 3 such systems will result from spiral-in evolution of a neutron star with a companion in the mass range 5–8 $M_\odot$ (*cf.* Fig. 4). In analogy with the CV binaries the post-spiral-in orbital periods of these systems are expected to range from ~ 1 h to ~ 1 d. According to Equation (10) the orbits of such systems (for $M_2 \sim 1\ M_\odot$, $M_1 \sim 1.4\ M_\odot$) with $P \leqslant 16$ h will decay on a timescale shorter than τ_{Hubble}. Consider such a system at the moment that the white dwarf begins to overflow its Roche lobe. At that moment the orbital period is ~ 20 s and $\tau_{GR} \sim 30$ yr. From Fig. 6 we can deduce the further evolution of these systems. It appears that three mass ranges can be distinguished (for $M_1 = 1.4\ M_\odot$):

(1) for $M_2^0 > 0.7\ M_\odot$ both $R_{L,c}$ and $R_{L,d}$ fulfil inequality (5). Hence runaway mass transfer is unavoidable. This is expected to lead to total disruption of the white dwarf on its dynamical timescale (a few seconds; *cf* van den Heuvel & Bonsema 1984).

(2) for $0.5\ M_\odot < M_2^0 \leqslant 0.7\ M_\odot$, conservative mass transfer is self-stabilizing, whereas disc mass transfer is unstable and leads to disruption. As the mass transfer begins suddenly, the latter may correspond to the real case.

(3) for $0.035\ M_\odot < M_2^0 \leqslant 0.5\ M_\odot$, both conservative transfer and disc transfer are self-stabilizing. Hence, stable mass transfer will ensue.

Incase (1), at least ~ 0.1 $M_\odot$ of the mass of the white dwarf is expected to be accreted by the neutron star (this amount of accretion is the least required to provide the kinetic energy for expelling the remaining part of the white-dwarf mass; *cf.* van den Heuvel & Bonsema 1984). The progenitor systems were relatively wide B-emission X-ray binaries with a companion in the mass range 6–8 $M_\odot$ (see Fig. 4).The runaway velocities of such systems, imparted by the formation of their neutron stars, are expected to be $\leqslant$ 20–30 km s^{-1} (Rappaport & van den Heuvel 1982). Indeed, B-emission X-ray binaries are always found close to the galactic plane. At the onset of the spiral-in, the neutron star will have an age ~ 2–3 × 10^7 yr. With an orbital period after spiral-in of ~ 1 h, coalescence will occur ~ 2 × 10^7 yr later, *i.e.* $\geqslant$ 4 × 10^7 yr after the birth of the neutron star. At such an age its magnetic field will have decayed to a value ~ 10^{8-9} G (see Fig. 2), and spin-up to a very short period is possible.

4.3.3 *Coalescence with a second neutron-star*

The orbit of PSR 1913 + 16 is observed to decay on a timescale of $\sim 3 \times 10^8$ yr (Taylor & Weisberg 1982). Hence, coalescence of the two neutron stars in this system is inevitable on this timescale. When the neutron stars have the same mass they will both begin to overflow their Roche lobes when the separation of their mass centres has been reduced to ~ 30 km. At that time the orbital period and the timescale for orbital decay are both equal to about 1 ms (Clark & Eardley 1977; Henrichs & van den Heuvel 1983). In the equal-mass case the result of the coalescence will be either a black hole or a massive neutron star—in the latter case with a rotation period slightly below a millisecond. In the case that the neutron stars differ in mass by ~ 0.2–0.4 $M_\odot$, the lower-mass one will have the largest radius and will be the first to overflow its Roche lobe. This leads to runaway mass transfer and spiral-out to ~ 100 km followed by total disruption of this star (Clark & Eardley 1977). In this process a considerable fraction of its mass, ~ 0.5 $M_\odot$, may be ejected, and the remaining object is likely to be a rather massive neutron star (~ 1.9 $M_\odot$), again with a rotation period of slightly less than a millisecond.

However, at rotation periods < 1.5 ms, neutron stars are unstable to the radiation of gravitational waves by non-radial stellar modes (Papaloizou & Pringle 1978). This leads to a rapid increase in period until $P \sim 1.5$ ms. The neutron stars that result from the coalescence of a close neutron star binary are, therefore, always expected to be found with a rotation period $\geqslant 1.5$ ms.

4.4 *Discussion and Conclusions*

In Table 3 the observed properties of PSR 1937 + 214 are compared with the predictions of each of the above-described coalescence models. The following remarks can be made:

(1) All of the 3 models can provide the ~ 0.12 $M_\odot$ accretion required for spin-up to a 1.5 ms rotation period.

(2) Two models predict with certainty that coalescence will occur. For the red-dwarf model, coalescence is predicted only if angular-momentum loss from the disc edge occurs, or if the mass-transfer can be interrupted. The latter case should be seen in combination with the condition that the magnetic field strength of 5×10^8G requires a neutron star age $\leqslant 10^8$ yr (see Fig. 2). This is much shorter than the age of the low-mass X-ray binary progenitor system, as such systems belong to an old disc population with ages $> 5 \times 10^9$ yr (*cf.* Lewin & Joss 1983; van Paradijs 1983). Both the above conditions can be fulfilled simultaneously if one assumes that the neutron star was formed $\sim 0.5 \times 10^8$ yr ago by the accretion-induced collapse of a white dwarf. This is because in that case (i) the progenitor system suddenly became detached due to the sudden increase in orbital separation produced by the SN mass loss of $\geqslant 0.1$ $M_\odot$ (see Section 3, and van den Heuvel 1977). Hence the mass transfer became interrupted, and (ii) if the mass of the red dwarf at that moment was < 0.035 $M_\odot$—as required for coalescence—the time interval required for it to fill its Roche lobe again (determined by GR losses) is $< 10^8$ yr. Thus, at the time of coalescence the neutron star is indeed expected to still have a magnetic field strength of ~ 10^8–10^9 G. It thus appears that the red-dwarf-coalescence model is viable if the progenitor was a CV binary, in which the red

Table 3. Comparison between observed properties of the 1.5-ms pulsar and those expected on the basis of the three different coalescence models described in the text.

Observed property	Is the observed property expected in the case of coalescence with a: Red dwarf	Massive white dwarf $M > 0.7\,M_\odot$	Neutron star
1. 1.5-ms rotation period	Yes	Yes	Yes
2. Single (*i.e.* will coalescence work?)	Yes, if system went through a detached phase (may require formation of neutron star by accretion-induced collapse)	Yes	Yes
3. Surface dipole magnetic field strength $\sim 5 \times 10^8$ G (*i.e.* neutron star age $\sim 10^8$ yr)	Yes, if neutron star was formed by accretion-induced collapse of a white dwarf	Yes	Not necessarily (if resulting neutron star is hot, B_s may be $\sim 10^{12\text{–}13}$ G)
4. Position close galactic plane	Not necessarily (progenitor: old disc population)	Yes	Not necessarily (runaway velocity $\gtrsim 100\,\mathrm{km\,s^{-1}}$)
Prospective progenitor system	CV binary with $M_{\mathrm{red}} < 0.035\,M_\odot$	Be-X-ray binary	Massive X-ray binary

component at the moment of the white-dwarf collapse had a mass < 0.035 $M_\odot$. In a similar white-dwarf-collapse model with a red companion more massive than 0.035 $M_\odot$, the resulting system after the detached phase presumably will be a low-mass X-ray binary resembling 4U 1626 – 67 (P_{orb} = 41 min) and 4U 1916 – 05 (P_{orb} $\simeq$ 50 min), in which the red stars have a mass between 0.01 $M_\odot$ and 0.1 $M_\odot$ (Rappaport & Joss 1984). These two alternatives for the evolution of CV binaries with accretion-induced collapse are represented by the dashed and dotted tracks, respectively, in Fig. 7.

(3) Neutron stars resulting from the coalescence of a close neutron-star binary might be hot and completely melted (Ruderman & Shaham 1983b), in which case a strong magnetic field might be re-generated (Flowers & Ruderman 1977). If this is the case (which seems not yet certain) this coalescence model would not be applicable for the origin of PSR 1937 + 214. [Still, however, a most interesting object would result: If a 1.5-ms pulsar has the same B_s value as the Crab pulsar, its total energy emission in the form of magnetic dipole radiation and energetic particles will be $\sim 2 \times 10^{43}$ erg s^{-1}, with a decay timescale of ~ 3 yr. The only compact objects known to have such a large energy emission on a short timescale are active galactic nuclei and quasars.]

(4) Coalescence with a red dwarf or a neutron star does not necessarily predict a position close to the galactic plane, as (i) low-mass X-ray binaries and CV binaries tend to belong to a relatively old stellar population with a spread in z-values of at least 0.5–1 kpc; (ii) neutron-star binaries like PSR 1913 + 16 are expected to have runaway velocities >100 km s^{-1} as a result of the two SN explosions that took place in them (see Section 4.1).

We conclude from the above that all three coalescence models are, in principle, viable for the formation of the 1.5-ms pulsar with the above-discussed provisions for the red-dwarf and neutron-star cases (see also Table 3). Only in the case of coalescence with a white dwarf of mass > 0.7 $M_\odot$ is the position close to the galactic plane a natural consequence.

Acknowledgements

I am grateful to H. Henrichs, V. Radhakrishnan, R. Taam and G. J. Savonije for discussions. Part of this work was started at the CECAM workshop on 'Late Stages of Close Binary Evolution', Paris-Meudon, summer 1982.

References

Alpar, M. A., Cheng, A. F., Ruderman, M. A., Shaham, J. 1982, *Nature*, **300,** 728.
Arons, J. 1983, *Nature*, **302,** 301.
Backer, D. C. 1984, *J. Astrophys. Astr.*, **5,** 187.
Blandford, R. D., DeCampli, W. M. 1981, in *IAU Symp.* 95: *Pulsars*, Eds W. Sieber & R. Wielebinski, D. Reidel, Dordrecht, p. 371.
Boriakoff, V., Buccheri, R., Fauci, F. 1983, *Nature*, **304,** 417.
Bradt, H. V. D., McClintock, J. E. A. 1983, *A. Rev. Astr. Astrophys.*, **21,** 13.
Bravo, E., Isern, J., Labay, J., Canal, R. 1983, *Astr. Astrophys.*, **124,** 39.
Brecher, K., Chanmugam, G. 1983, *Nature*, **302,** 124.
Clark, J. P., Eardley, D. 1977, *Astrophys. J.*, **215,** 311.
Damashek, M., Backus, P. R., Taylor, J. H., Burkhardt, R. 1982, *Astrophys. J.*, **253,** L57.
Delgado, A. J., Thomas, H. C. 1981, *Astr. Astrophys.* **96,** 142.

Eggleton, P. P. 1983, in *IAU Symp.* 72*: Cataclysmic Variables and Related Objects*, Eds M. Livio *&* G. Shaviv, D. Reidel, Dordrecht, p. 239.
Fabian, A. C., Pringle, J., Verbunt, F., Wade, A. 1983, *Nature*, **301,** 222.
Faulkner, J. 1971, *Astrophys. J.*, **170,** L99.
Flannery, B. P., van den Heuvel, E. P. J. 1975, *Astr. Astrophys.*, **39**, 61.
Flowers, E., Ruderman, M. A. 1977, *Astrophys. J.*, **215,** 302.
Goldreich, P., Julian, W. H. 1969, *Astrophys. J.*, **157**, 869.
Gunn, J. E., Ostriker, J. P. 1970, *Astrophys. J.*, **160**, 979.
Habets, G. 1984, in *IAU Symp.* 105: *Observational Tests of Stellar Evolution*, Eds A. Maeder & A. Renzini, D. Reidel, Dordrecht (in press).
Helfand, D. J., Ruderman, M. A., Shaham, J. 1983, *Nature*, **304**, 423.
Henrichs, H. F. 1983, in *Accretion-Driven Stellar X-ray Sources*, Eds W .H. G . Lewin & E. P. J. van den Heuvel, Cambridge Univ. Press, p. 392.
Henrichs, H. F., van den Heuvel, E. P. J. 1983, *Nature*, **303**, 213.
Iben, I, Tutukov, V. 1984, *Astrophys. J. Suppl. Ser.*, **54**, 335.
Joss, P. C., Rappaport, S. A. 1983, *Nature*, **304**, 419.
Kieboom, K. H., Verbunt, F. 1981, *Astr. Astrophys.*, **95**, L11.
Lewin, W. H, G., Joss, P. C. 1983, in *Accretion-Driven Stellar X-ray Sources*, Eds W. H. G. Lewin *&*. E. P. J. van den Heuvel, Cambridge Univ. Press, p. 41.
Lyne, A. 1981, in *IAU Symp. 95: Pulsars*, Eds W. Sieber & R. Wielebinski, D, Reidel, Dordrecht, p. 423.
Manchester, R. N., Newton, L. M., Cooke, D. J., Backus, P. R., Damashek, M., Taylor, J. H., Condon, J. J. 1983, *Astrophys. J.*, **268**, 832.
Manchester, R. N.. Taylor, J. H. 1977, *Pulsars*, Freeman, San Francisco.
Manchester, R. N., Taylor, J. H. 1981, *Astr. J.*, **86**, 1953.
Meyer. F, Meyer-Hofmeister, E. 1979, *Astr. Astrophys.*, **78**, 167.
Nomoto, K. 1980, *Space Set Rev.*, **27**, 563.
Nomoto, K. 1982, *Astrophys. J.*, **253**, 798.
Pacini, F. 1968, *Nature*, **219**, 145.
Pacini F. 1983, *Astr. Astrophys*, **126**, L11.
Paczyński, B. 1970, *Acta astr.*, **20**, 47.
Paczyński, B. 1971, *A. Rev. Astr. Astrophys.*, **9**, 183.
Paczyński, B. 1976, in *IAU Symp.* 73: *Structure and Evolution of Close Binaries*, Eds P. P. Eggleton, S. Mitton & J. Whelan, D. Reidel, Dordrecht, p. 75.
Paczyński, B. 1983, *Nature*, **304**, 421.
Paczyński, B., Sienkiewicz, R. 1971, *Acta astr.*, **21**, 73.
Paczyński, B., Sienkiewicz, R. 1981, *Astrophys. J.*, **248**, L27.
Papaloizou, J., Pringle, J. E. 1978, *Mon. Not. R. astr. Soc.*, **184**, 501.
Plavec, M. 1968, *Adv. Astr. Astrophys.*, **6**, 201.
Pringle, J. E. 1981, *A. Rev. Astr. Astrophys.*, **19**, 137.
Rappaport, S. A., Joss, P. C. 1983, in *Accretion-Driven Stellar X-ray Sources*, Eds W. H. G, Lewin & E. P. J. van den Heuvel, Cambridge Univ. Press., p. 1.
Rappaport, S. A., Joss, P. C. 1984, *Astrophys. J.* (in press).
Rappaport, S. A., Joss, P. C, Webbink, R. F. 1982, *Astrophys. J.*, **254**, 616.
Rappaport, S. A., van den Heuvel, E. P. J. 1982, in *IAU Symp. 98: B-emission Stars*, Eds M. Jaschek & H. G. Groth, D. Reidel, Dordrecht, 327.
Radhakrishnan, V. 1982, *Contemp. Phys.*, **23**, 207.
Radhakrishnan, V., Srinivasan, G. 1981, paper presented at *2nd Asian-Pacific Regional IAU Meeting*, Bandung.
Radhakrishnan, V., Srinivasan, G. 1982, *Current Sci.*, **51**, 1096.
Ruderman, M. A., Shaham, J. 1983a, *Nature*, **304**, 425.
Ruderman, M. A., Shaham, J. 1983b, *Comments Astrophys.*, **10**, 15.
Ruderman, M. A., Sutherland, P. G. 1975, *Astrophys. J.*, **196**, 51.
Savonije, G. J. 1983a, *Nature*, **304**, 422.
Savonije, G. J. 1983b, in *Accretion-Driven Stellar X-ray Sources*, Eds W. H. G. Lewin & E. P. J. van den Heuvel, Cambridge Univ. Press, p. 343.
Seward, F. D., Harnden, F. R., Helfand, D. 1984, *IAU Circ. No.* 3928.
Sion, E. M., Acierno, M. J., Tomsczyk, S. 1979, *Astrophys. J.*, **230,** 832.

Smarr, L. L., Blandford, R. D. 1976, *Astrophys. J.*, **207**, 574.
Srinivasan, G., van den Heuvel, E. P. J. 1978, paper presented at 9*th Texas Symp. on Relativistic Astrophysics*, München (unpublished).
Srinivasan, G., van den Heuvel, E. P. J. 1982, *Astr. Astrophys.*, **108,** 143.
Stevenson, D. J. 1980, *Proc. Conf. Physics of Dense Matter: J. Phys. Coll.* **41,** C2/53.
Sutantyo, W. 1981, paper presented at *2nd Asian-Pacific Regional Meeting IAU*, Bandung.
Taam, R. E. 1983, *Astrophys. J.*, **270**, 694.
Taam, R. E., van den Heuvel, E. P. J. 1984, *Astr. Astrophys.*, (in press).
Taam, R. E, Bodenheimer, P., Ostriker, J. 1978, *Astrophys. J.*, **222**, 269.
Taylor, J. H. 1981, in *IAU Symp. No. 95: Pulsars*, Eds W. Sieber & R. Wielebinski, D Reidel, Dordrecht, p. 361.
Taylor, J. H., Weisberg, J. M. 1982, *Astrophys. J.*, **253,** 908.
van den Heuvel, E. P. J. 1977, *Ann. N. Y. Acad. Sci.*, **302,** 14.
van den Heuvel, E. P. J. 1981, in *IAU Symp. No.* 95: Pulsars, Eds W. Sieber & R Wielebinski, D. Reidel, Dordrecht, p, 379.
van den Heuvel, E. P. J., Bonsema, P. 1984, *Astr. Astrophys.* (in press).
van den Heuvel, E. P. J., Taam, R. E. 1984, *Nature*, **309,** 235.
van Paradijs, J. A. 1983, in *Accretion-Driven Stellar X-ray Sources*, Eds W. H. G. Lewin &E. P. J. van den Heuvel, Cambridge Univ. Press, p. 189.
Verbunt, F., Zwaan, C. 1981, *Astr. Astrophys.*, **100,** L7.
Webbink, R. F., 1979, in *IAU Coll.* 53: *White dwarfs and Variable Degenerate Stars*, Eds H. M. Van Horn & V. Weidemann, Univ. Rochester, p. 426.
Webbink, R. F., Rappaport, S. A., Savonije, G. J. 1983, *Astrophys. J.*, **270,** 678.
Whelan, J., Iben, I. 1973, *Astrophys. J.*, **186,** 1007.
Zahn, J. P. 1977, *Astr. Astrophys.*, **57,** 383; erratum: 1978, *Astr. Astrophys.*, **67,** 162.

J. Astrophys. Astr. (1984) **5,** 235–250

Gravitational Lensing by Stars in a Galaxy Halo: Theory of Combined Weak and Strong Scattering

Rajaram Nityananda *Princeton University Observatory*, *Peyton Hall*, *Princeton*, *NJ 08544*, *USA* and *Raman Research Institute*, *Bangalore 560080**

J. P. Ostriker *Princeton University Observatory*, *Peyton Hall*, *Princeton*, *NJ 08544*, *USA*

(*Invited article)*

Abstract. The theory of gravitational lensing of background quasars by stars in the halo of a galaxy is considered. In the limiting case of small 'optical depth', only one star is close enough to the beam to cause strong scattering, and the effect of all the other stars is treated as a perturbation with both systematic and random components. The perturbation coming from weak scattering can increase the number of images and the amplification in those cases where the amplification is already high; such events are preferentially selected in flux limited observations. The theory is applicable to the apparent association of background quasars with foreground galaxies. A comparison with earlier work on the same problem is given. The relevance of these results to gravitational lensing by galaxies as perturbed by random inhomogeneities surrounding the ray path is also briefly discussed.

Key words: gravitation lens—galaxy halos—quasars

l. Introduction

The possibility that the luminosity function of quasars derived from observations could be significantly influenced by gravitational lensing was originally proposed by Barnothy & Barnothy (1968) and more recently considered by Turner (1980). Avni (1981) and Peacock (1982) have subsequently examined this problem, including small effects due to the deamplification which is implied by flux conservation. The fluctuations in the intensity of the gravitationally lensed image of the quasar 0957 +561 due to 'mini-lensing' by stars close to the beam were considered by Chang & Refsdal (1979) and Gott (1981). The latter emphasised the unique opportunity this affords to detect low mass ($\sim 0.001\ M_{\odot}$) objects in galactic halos. Yet another aspect of lensing by stars in galactic halos was pointed out by Canizares (1981), *viz.*, that it could produce an apparent association between quasars and foreground galaxies. Vietri & Ostriker (1983) have reconsidered this problem, including not only the effects of individual stars but also that of the galaxy as a whole as well as flux conservation. They also introduced the very useful concept of an optical depth τ for significant amplification (defined below). All the problems described above centre around the distribution of amplifications produced by an encounter with a single galaxy. The basic concepts involved are: (1) lensing by individual stars; (2) superposition of the weak amplification caused by many distant encounters with strong amplification caused by a close encounter;

* Permanent address.

(3) effect of the smoothed-out potential of the galaxy; (4) the deamplification required by flux conservation. The present analysis of the problem closely follows Vietri & Ostriker (1983, hereafter VO). They use a formalism similar to radiative transfer and treat the case $\tau \ll 1$. As discussed in the next section, we take an alternative view of the same problem which clarifies and extends their work. Most of the new results concern situations in which weak lensing effects are combined with a strong event. We find that such weak events do not simply superpose with the strong ones but can have a. significant, even surprisingly large, effect on the total amplification and on the number and geometry of the images. As Turner, Ostriker & Gott (1984; hereafter TOG) have shown, lensing events in a flux-limited sample are predominantly those with high amplification, and hence they are the ones where the effects discussed in this paper could be important.

The plan of this paper is as follows. In Section 2 we discuss two alternative points of view on gravitational lensing to clarify the issues of deamplification and flux conservation. Section 3 reviews the basic definitions, notation, and equations. Section 4 treats the superposition of weak and strong amplifications and Section 5 includes the smooth potential of the galaxy. Section 6 is a discussion and summary.

2. Filled and empty beams, flux conservation and negative amplification

There are two equivalent viewpoints in gravitational lensing which differ in the zero order description of the propagation of light. In the first, which we call the 'filled-beam' approach, the matter is first smoothed out to make the cosmology truly homogenous. One then considers under and over-densities which cause a beam of light to diverge or converge giving deamplification or amplification with respect to the flux calculated for a homogenous universe. In the alternative 'empty-beam' approach, the starting point is an evacuated tube in an otherwise homogenous universe. If the matter in the real universe is sufficiently lumpy so that we view at distant sources via empty regions, this may be a better first approximation. Zeldovich (1964) and Feynman (quoted by Gunn 1967) pointed out that this reduces the angular size and flux with respect to the filled-beam case. Gunn (1967) discussed the fluctuations produced by the discreteness of the matter outside the beam. In their classic paper on the lensing effects of a cosmological distribution of point masses, Press & Gunn (1973) used the empty beam approach. In this approach all density fluctuations are positive and act to increase the observed flux.

It is of course possible that a substantial fraction of the mass density in the universe is more smoothly distributed than the galaxies. One can then use a tube in which the density of galaxies alone has been removed. For reference, we give the deamplification as a function of redshift for an $\Omega = 1$ universe in which a fraction Ω_g of the density has been removed from a tube enclosing the beam. With the definitions

$$n_a \equiv [1-(1+24\Omega_g)^{1/2}]/2; \quad n_b \equiv [1+(1+24\Omega_g)^{1/2}]/2, \tag{1}$$

the effect of removing a fractional density Ω_g is to multiply the flux from point sources by the factor $A_g (A_g < 1)$ given below (Dashevskii & Slysh 1965)

$$A_g = \frac{(n_b - n_a)^2[(1+z_q)^{1/2}-1]^2}{[(1+z_q)^{n_b/2}-(1+z_q)^{n_a/2}]^2}. \tag{2}$$

For an empty tube ($\Omega_g = 1$), this takes the simpler form (Zeldovich 1964)

$$A_g = \frac{25[(1+z_q)^{1/2}-1]^2}{[(1+z_q)^{3/2}-(1+z_q)^{-1}]^2}. \tag{3}$$

To take a somewhat extreme example, an empty tube in an $\Omega = 1$ universe produces deamplification by a factor 0.42 at a redshift of 3. So long as we are dealing with smaller redshifts, it is clear that deamplification effects amount to a fraction of a magnitude and cannot be more important than the comparable or greater uncertainties which are already present in the quasar luminosity function. The effect of this deamplification on the observed surface density of quasars also depends on the slope of the relation between number counts and apparent magnitude (VO). Typically, the fractional change in the number counts at a given magnitude due to deamplification is less than the fractional flux change. In the association problem considered by Canizares (1981), one compares the density of quasars in two fields one of which is close to a foreground galaxy. The deamplification is common to both the fields and hence does not affect the result.

The same empty and filled beam approaches also apply to lensing by stars in galaxies. VO use the filled beam approach and represent the galaxy by a smooth mass distribution. Individual stars are then represented by point masses with surrounding compensating negative mass spheres to avoid double counting. In the present paper, we use the empty beam approach and thus deal only with overdensities and amplification.

3. Review of lensing definitions and equations

We briefly review the basic equations governing lensing, following the treatment of Young (1981). The geometry of a gravitational lensing event is shown in Fig. 1. The

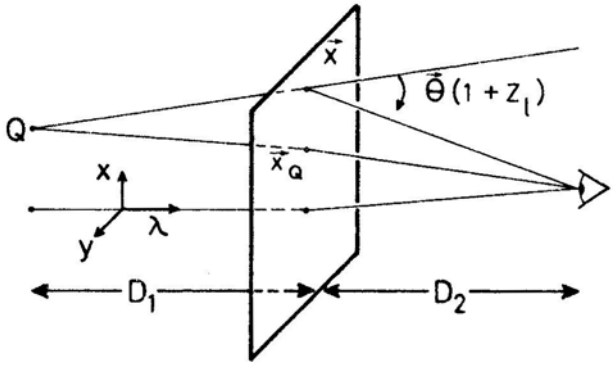


Figure 1. Geometry of a gravitational lensing event. Rays from the background source (quasar) at Q at redshift z_q undergo a proper vector deflection θ at a redshift z_1 and propagate to the observer. x and y are transverse proper distances. x_Q is the intercept on the deflector plane of the unperturbed ray from Q to the observer. D_1 and D_2 are affine distances.

source (quasar) is located at a red shift z_q and the lensing object (also called the deflector) at a red shift z_l We use proper distances in the deflector plane as Cartesian coordinates x, y. The unperturbed ray connecting the quasar and the observer in the absence of gravitational lensing cuts the deflector plane at x_q, y_q. The deflection at a point x, y in the deflector plane is a vector θ whose components θ_x and θ_y are expressed in terms of the Newtonian gravitational potential ϕ (x, y, z) in the weak field limit (Bourassa & Kantowski 1975).

$$\theta_x = -\frac{2}{c^2}\frac{\partial}{\partial x}\int \phi(x, y, z)\,dz \equiv -\frac{2}{c^2}\frac{\partial}{\partial x}\Phi(x, y),$$
$$\theta_y = -\frac{2}{c^2}\frac{\partial}{\partial y}\int \phi(x, y, z)\,dz \equiv -\frac{2}{c^2}\frac{\partial}{\partial y}\Phi(x, y). \tag{4}$$

We have denoted the integral of the gravitational potential along the line of sight (z direction) by $\Phi(x, y)$ in (2). This is related to the surface density Σ (x, y) by Poisson's equation

$$\frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} = 4\pi G\Sigma(x, y). \tag{5}$$

Strictly speaking, the potential due to a two-dimensional mass distribution suffers from a logarithmic divergence if it is required to vanish as $x, y \to \infty$. We can cure this by choosing any (finite) point as the zero of potential, without affecting the gradient of the potential which determines the deflection of a ray. This 'two-dimensional' view of lensing is valid so long as the extent of the deflector along the line of sight is much smaller than its distance from either the source or observer. Turner, Ostriker & Gott (1984) have shown how a uniform sheet of matter at a different red shift from a given galaxy can be replaced by an equivalent sheet at the same red shift so that the two-dimensional picture can then be used.

In the 'empty-beam' approach, the rays are regarded as travelling in an evacuated tube in the universe, with deflections by galaxies and stars put in through Equation (4). Press & Gunn (1973) have introduced a particularly convenient coordinate system (shown in Fig. 1) in which distances along the line of sight are measured in terms of the affine parameter* λ. In the transverse direction, the proper lengths x and y are used. Of course, it is assumed that x and y are much smaller than characteristic distances along the ray. In this coordinate system, light rays close to the axis are straight lines so long as there is circular symmetry around the beam and no matter. In a Friedman cosmology with a scale factor $a(t)$, we have $d\lambda = ca(t)dt$. The affine length normalized in this way agrees with the proper distance at the present epoch, and is less than it by the factor $1 + z$ at earlier epochs. The angles in Fig. 1 are thus $1 + z$ times proper angles. From the geometry of Fig. 1, we have the following relation between x, the image position, and x_q, the unperturbed quasar position.

$$(\mathbf{x} - \mathbf{x}_q)(1/D_1 + 1/D_2) = -\boldsymbol{\theta}(1 + z_l). \tag{6}$$

It is convenient to define the effective distance D by

$$(1 + z_l)/D \equiv 1/D_1 + 1/D_2. \tag{7}$$

* The properties of the affine parameter are treated in the texts by Misner, Thorne & Wheeler (1972) and Hawking & Ellis (1973).

For a given quasar position x_q one can solve for the values of x, each corresponding to an image of the source. Further, the Jacobian of x with respect to x_q gives the amplification A. Young (1981) has used the concepts of shear κ and convergence γ of an infinitesimal bundle of rays (defined below) to write the amplification A in a simple form.

$$\gamma_1 \equiv -\frac{D}{2}\left(\frac{\partial\theta_x}{\partial x}-\frac{\partial\theta_y}{\partial y}\right); \quad \gamma_2 \equiv -\frac{D}{2}\left(\frac{\partial\theta_x}{\partial y}+\frac{\partial\theta_y}{\partial x}\right); \tag{8}$$

$$\gamma^2 = \gamma_1^2+\gamma_2^2; \quad \kappa \equiv -\frac{D}{2}\left(\frac{\partial\theta_x}{\partial x}+\frac{\partial\theta_y}{\partial y}\right); \quad A = [(1-\kappa)^2-\gamma^2]^{-1}. \tag{9}$$

We can substitute the expression (4) for the deflection into the definition (8) of the convergence, which turns out to be proportional to the two-dimensional Laplacian (5) of the potential, *i.e.*, to the surface density $\Sigma(x, y)$. The convergence thus vanishes when there is no matter in the beam. The shear γ represents tidal deformation produced by matter not in the beam. Note that κ and γ^2 are scalars under rotation by an angle α in the x-y plane while γ_1 and γ_2 transform as components of a second-rank symmetric tensor, *viz.*

$$\gamma_1' = \gamma_1 \cos 2\alpha + \gamma_2 \sin 2\alpha, \quad \gamma_2' = -\gamma_1 \sin 2\alpha + \gamma_2 \cos 2\alpha. \tag{10}$$

It is also clear from (9) that for weak scattering ($\kappa \ll 1$, $\gamma \ll 1$), the convergence makes a first-order contribution to the amplification while the shear appears only in the second order, since it involves expansion in one direction and contraction in the perpendicular one.

We now review briefly the lensing properties of a point mass (VO give more details). There is a critical radius r_0 in the deflector plane given in terms of the Schwarzschild radius r_s and the effective distance D by

$$r_0^2 = 2r_s D. \tag{11}$$

When the unperturbed ray approaches within r_0 of the point mass, the amplification is significant (greater than 1.34) and the two images become of comparable intensity (the intensity ratio is less than 6.9). Thus πr_0^2 is a natural cross section for a strong lensing event. VO introduce the optical depth τ in terms of the density n of stars by

$$d\tau = \pi r_0^2 n\, dl.$$

When all the lensing matter lies at the same red shift, r_0 is effectively constant and the optical depth can then be expressed in terms of the surface density n_s of stars

$$\tau \equiv n_s \sigma_0 = n_s \pi r_0^2. \tag{12}$$

For small optical depth ($\tau \ll 1$), circles of radius r_0 around the projected position of each star in the x-y plane do not overlap and are in fact separated typically by $\sim \tau^{-1/2} r_0$ which is greater than r_0. For a $\tau \ll 1$, strong lensing event is due to the beam passing within $\sim r_0$ from a single star. The effect of the other stars can be treated as a perturbation (see Section 4 below).

The finite angular size of the source and the breakdown of geometrical optics can both limit the amplification produced by lensing. This was discussed briefly by Gott (1981) and VO; we go into some more details here. For high amplification A, the closest approach r of the unperturbed ray is less than the critical radius r_0 and we have $A = r_0/r$.

To ensure that the amplification does not vary appreciably over an extended source for which r varies by Δr we must have $\Delta r < fr$ where f is a safety factor less than 1. This defines a critical linear size in the deflector plane, below which we obtain the same amplification as for a point source.

$$\Delta r < fr_0 / A. \tag{13}$$

Gott (1981) and VO used the weaker condition

$$\Delta r < fr_0/A^{1/2},$$

which implicitly assumes that the amplification is isotropic. For a point mass the amplification is entirely in the tangential direction and our condition (11) applies. This correction raises by a factor A, the minimum mass of a star which can act as a 'mini-lens' for a quasar of a given angular size. The interpolation formula given by Peacock (1982) for the amplification of extended sources is consistent with the criterion (13) since the maximum amplification scales inversely as the source size.

We now consider the possible breakdown of geometrical optics which involves two distinct effects. A ray connecting the source and the observer represents a path of stationary phase. We can draw the so-called Fresnel zone (*cf.* Born & Wolf 1975) in the deflector plane with a radius r_F around each image. Paths passing within this zone differ by less than half a wavelength and hence contribute significantly to the observed intensity. In terms of the effective distance D and the observed wavelength W_0, we have

$$r_F = [Dw_0/(1+z_l)]^{1/2}. \tag{14}$$

The geometry of point-mass lensing shows that an event with amplification A involves rays which pass within $r_0/2A$ of the critical radius r_0. To ensure that the amplification does not vary significantly over r_F, we need

$$fr_0/2A > r_F$$

or equivalently

$$w_0/(1+z_l) < f^2 r_0^2/4A^2 D = f^2 r_S/2A^2. \tag{15}$$

VO give a similar criterion for a typical lensing event ($A - 1$ of the order of unity). It is interesting that apart from red shift and amplification-dependent factors, the scale of wavelengths is set by the Schwarzschild radius of the lensing mass. We note, as a consequence of (15), that geometrical optics breaks down more easily, *i.e.*, at shorter wavelengths, for high amplification events. We note that even for high amplification ($A \sim 10$) and small masses ($10^{-2}\ M_\odot$) the criterion (15) is comfortably fulfilled upto decametre wavelengths and geometrical optics is valid.

Even after fulfilling the condition (15) which guarantees the validity of geometrical optics for the intensity of each ray, we can get interference effects between the two images for a source of small enough angular size. From Fig. 1, the two rays part at the angle $2r_0/D_1\,(1+z_q)$ at the source. They will be incoherent only for wavelengths Satisfying

$$\frac{w_0}{1+z_q} < \frac{2r_q \cdot 2r_0}{D_1(1+z_q)}, \tag{16}$$

where $2r_q$ is the linear size of the source. This is basically the same as Gott's (1981) condition that the gravitational lens of size $2r_0$ should be able to resolve the source of

angular size $2r_0 /D_1$. For smaller source or lens sizes, one can obtain fringes—but subject to one more condition. Given a path difference between the two rays of n wavelengths, one needs a fractional bandwidth of less than $1/n$. The path difference is directly related to the time delay between the two images which has been calculated by Refsdal (1964). Using this result we find

$$n = (1+z_1)\, r_s / A w_0. \tag{17}$$

Comparing (17) and (15), we see that the fractional bandwidth has to be less than a critical value to get two beam interference.

$$\Delta w_0/w_0 < f_2/2A.$$

4. Superposition of weak and strong amplifications

Gott (1981) has shown that rays from background objects which are not doubly imaged by an isothermal galaxy encounter an optical depth due to individual stars that is less than 1/4. The condition $\tau \ll 1$ applies to all the cases of interest in this paper. We thus have two possibilities: (1) rays which pass at a distance significantly greater than r_0 from the projected position of all stars, (2) rays which pass one star at a distance $\sim r_0$ and hence typically at $> r_0$ from all the others. In case (1) we are dealing with the superposition of weak amplifications. VO (in their Appendix A) show that it is correct, on the average, to multiply the amplifications which individual stars would have produced acting on their own. In case (2), we have to combine a weak and a strong amplification. Although VO have used the superposition principle in this case as well, the result proved in their Appendix A actually applies to the superposition, on the average, of the shears F_1 and F_2 produced by two stars. Since the relation (9) between shear and amplification is nonlinear, we do not expect the superposition principle to hold for the amplification in case (2) as shown below.

Let the image considered lie at the origin and let $(r_i\ \theta_i)$ be the polar coordinates of stars of mass m. The shear components γ_1, γ_2 defined in (8) can easily be evaluated

$$\gamma_1 = -\sum_i \left(\frac{r_0}{r_i}\right)^2 \cos 2\theta_i; \quad \gamma_2 = -\sum_i \left(\frac{r_0}{r_i}\right)^2 \sin 2\theta_i. \tag{18}$$

For example, we consider the case when a close encounter with one star produces a shear F_1 of magnitude close to one (strong amplification) and is at a polar angle zero, while a second star has a shear $F_2 \ll 1$ at polar angle θ. The amplification A is given by

$$\begin{aligned} A^{-1} &= 1 - (F_1 + F_2 \cos 2\theta)^2 - (F_2 \sin 2\theta)^2 \\ &= 1 - F_1^2 - F_2^2 - 2F_1 F_2 \cos 2\theta. \end{aligned} \tag{19}$$

The average amplification $\bar{A}$ is given by averaging over θ:

$$\bar{A} = [(1 - F_1^2 - F_2^2) - 4F_1^2 F_2^2]^{1/2}. \tag{20}$$

We can compare this to the result given by the superposition principle.

$$\bar{A}_{Su} = (1 - F_1^2)^{-1}(1 - F_2^2)^{-1}. \tag{21}$$

We see from (20) and (21) that even for $F_2 \ll 1$, if we have $F_2 \sim 1 - F_1^2$, the true amplification can be significantly greater than that given by superposition. In the

limiting case $A_1^{-1} = 1 - F_1^2 \ll 1, F_2 \ll 1$, we find

$$\bar{A}/\bar{A}_{Su} = |1 - 4F_2^2 A_1^2|^{-1/2}. \tag{22}$$

This shows that a weak shear F_2 is boosted by a factor of the order of A_1 the amplification due to the strong scattering alone.

In the rest of this section, we treat the problem of gravitational lensing by a point mass, perturbed by a weak shear $\gamma \ll 1$ produced by neighbouring masses. As suggested by (22), the imaging is strongly perturbed for $F_2 \sim 1 - F_1^2$, *i.e.*, for amplifications of the order of γ^{-1}. As shown below, one can even obtain four images instead of two. For a surface density n_s, typical nearest neighbour distances are given by $r \sim n_s^{-1/2}$. From (18), the shear γ is typically given by

$$\gamma \sim r_0^2 n_s \sim \tau. \tag{23}$$

We evaluate the probability distribution of the shear γ and average the probability distribution of amplifications over it. This approach to the problem does full justice to the tensorial nature of the shear, the deviations from the superposition principle, and the random distribution of the projected star positions.

As just noted, a weak shear γ significantly affects high amplification events with $A \sim \gamma^{-1}$. The equations governing lensing by a point mass in the presence of shear were given by Chang & Refsdal (1979) in the context of the *B* image of the quasar 0957 + 561 (Walsh, Carswell & Weymann 1979). Since $\gamma \sim 1$ in this case, the problem had to be solved numerically. We need the case $\gamma \ll 1$. Defining dimensionless distances in the deflector plane by

$$X \equiv (X_1, X_2) \equiv x/r_0; \qquad X_q \equiv (X_{q1}, X_{q2}) \equiv x_q/r_0, \tag{24}$$

the basic lensing Equation (3) reads

$$X_1 - X_{q1} = \frac{X_1}{X_1^2 + X_2^2} + \gamma X_1; \quad X_2 - X_{q2} = \frac{X_2}{X_1^2 + X_2^2} - \gamma X_2. \tag{25}$$

The first term on the right sides of (24) and (25) represents the deflection produced by a point mass and the second that due to the shear γ.The amplification (9) now reads

$$A^{-1} = 1 - \left[\frac{X_1^2 - X_2^2}{(X_1^2 + X_2^2)^2} - \gamma\right]^2 - \left[\frac{2X_1 X_2}{(X_1^2 + X_2^2)^2}\right]^2. \tag{26}$$

In terms of polar coordinates defined by

$$X_1 = R\cos\theta, \quad X_2 = R\sin\theta; \quad X_{q1} = R_q\cos\theta_q, \quad X_{q2} = R_q\sin\theta_q,$$

we can rewrite (24) and (25) as radial and tangential equations

$$R - R_q\cos(\theta - \theta_q) = 1/R + \gamma R\cos 2\theta, \tag{27}$$

$$-R_q\sin(\theta - \theta_q) = \gamma R\sin 2\theta, \tag{28}$$

with the amplification give by

$$A^{-1} = 1 - \left(\frac{\cos 2\theta}{R^2} - \gamma\right)^2 - \left(\frac{\sin 2\theta}{R^2}\right)^2. \tag{29}$$

Further, in the high amplification limit we can set

$$R = 1 + \delta R; \quad \delta R \ll 1.$$

As is clear from (29), we have $\delta R \sim \gamma$ for $A \sim \gamma^{-1}$. Retaining terms of order γ, Equations (27) and (29) simplify to

$$2\delta R = R_q \cos(\theta - \theta_q) + \gamma \cos 2\theta, \tag{30}$$

$$— R_q \sin(\theta - \theta_q) = \gamma \sin 2\theta, \tag{31}$$

$$A^{-1} = 4\delta R + 2\gamma \cos 2\theta. \tag{32}$$

Equations (30)–(32) describe the high-amplification limit of lensing by a point mass with weak external shear. Equation (31) gives the angular position of the image, giving four solutions for $R \ll \gamma$ and two for $R_q \gg \gamma$. For each value of θ satisfying (31), one can compute the radial position of the image from (30) and the amplification from (32). It is clear from (30)–(32) that R_q R and A all scale with the shear γ and we only need to solve for $A\gamma$ in terms of R_q/γ and θ_q.

Figure 2 shows the probability distribution of amplifications obtained by numerical solution of (30)–(32). At low amplifications, we have $P(A) \propto dA/A^3$ as for isolated point masses (VO). As the amplification approaches γ^{-1} , the probability falls below its unperturbed value. This is consistent with Equation (22) which predicts that shear shifts events to higher amplification. Then, at $A = \gamma^{-1}$ there is a sharp rise in $P(A)$, due to the onset of events with four images instead of two. At still higher values of A, the curve returns to its A^{-3} form. The net effect of the shear is thus to shift a certain fraction of the events with $A \sim \gamma^{-1}$ to amplifications higher by a factor of two or more.

Just as we define a cross-section $\sigma_0 = \pi r_0^2$ for strong lensing events, one can define a cross-section σ_4 for events in which the shear is typically strong enough to produce four images. This is comparable to the cross-section for producing amplifications higher than τ^{-1}. We have $\sigma_4 \simeq \sigma_0 \tau^2$, for small optical depth. As τ approaches 1, it is more and more probable that more than two images form.

5. Effect of the smooth potential of the galaxy

So far, we have discussed lensing by individual stars, perturbed by the random shear produced by their neighbours close to the ray, as if they made up a homogeneous sheet. However, there is also a systematic potential acting on a light ray produced by the overall mass distribution of the galaxy. For clarity, we first look at the case when this mass is dominated by stars which act as point masses, rather than something like neutrinos which act as a continuous distribution. There are two basic effects: (1) a systematic shear γ_g and (2) an increase, by a factor of $1 + 2\tau$, of observed solid angles over those in the absence of lensing.

Imaging by a galaxy has been discussed by Gott & Gunn (1974) and reviewed more recently by VO and TOG. The smoothed-out potential of the galaxy idealised as a singular isothermal sphere, produces a constant deflection angle θ_g given by

$$\theta_g = 4\pi\sigma^2 / c^2, \tag{33}$$

where σ is the one-dimensional velocity dispersion of the matter making up the isothermal. The critical radius r_g defined by

$$r_g = D\theta_g \tag{34}$$

plays a role similar to that of r_0 in the case of a point mass with the cross-section for

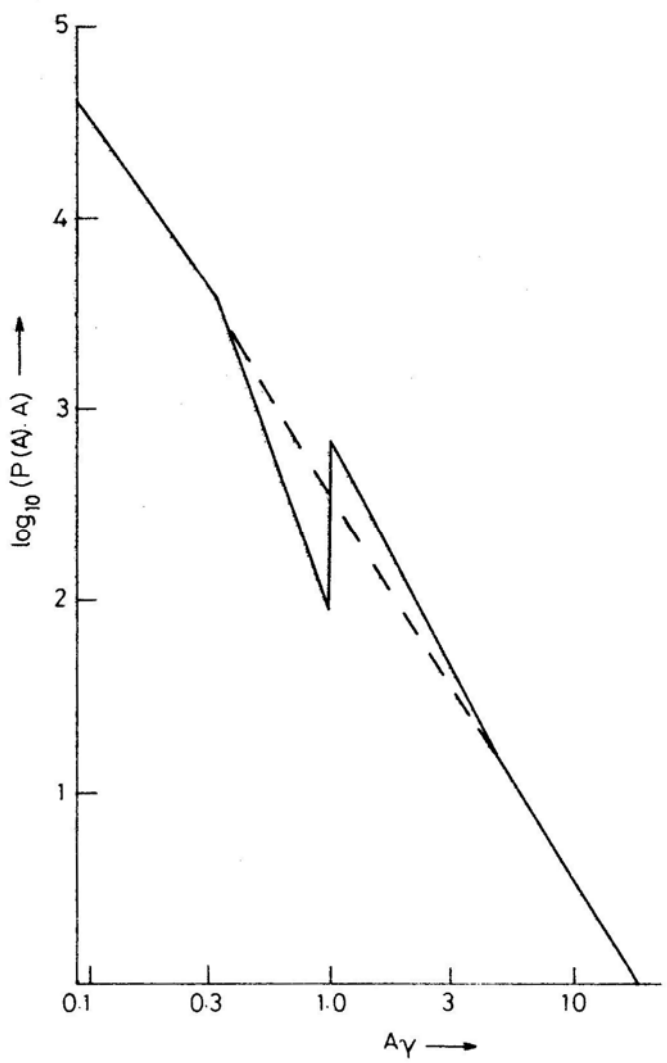


Figure 2, Probability distribution $AP(A)$ (in arbitrary units) of finding amplification A in a given logarithmic interval. The amplification along the x-axis is measured in units of γ^{-1} where γ is a shear (see Equations 30–32).

galaxy lensing given by πr_g^2. There is significant amplification accompanied by formation of two images with comparable intensity when the impact parameter is less than r_g. For impact parameters greater than r_g, only one image is formed. Gott (1981) has shown that the optical depth τ for scattering by the stars making up the galaxy seen by rays passing at a distance r from the centre is given by

$$\tau(r) = r_g/2r \tag{35}$$

This result is independent of the stellar mass so long as geometrical optics holds, *i.e.*, condition (15) is fulfilled.

The convergence κ and shear γ produced by the smooth potential of an isothermal

galaxy can be calculated from Equations (8), (33) and (35), the result being

$$\kappa_g(r) = -\gamma_g(r) = r_g/2r = \tau(r). \tag{36}$$

We now show that the rays which propagate in between the stars experience the shear γ but not the convergence κ. The argument is similar to that used in discussions of the Lorentz local field in dielectrics (*cf.* Kittel 1966). We use the fact that the spacing between the stars (projected onto the deflector plane) is five or six orders of magnitude smaller than length-scales like r_g and r associated with the lensing galaxy. It is therefore possible to choose an area element at r with radius r_d with the following properties: (1) The surface density n_s does not vary appreciably within r_d. (2) The discreteness of the mass distribution outside the circle r_d can be neglected in calculating the shear and convergence of rays near the centre. The distribution of the matter outside r_d is thus that of a smooth galaxy with a uniform disc of radius r_d removed. The convergence κ is entirely produced by the local surface density (as noted following Equations 8 and 9) and hence vanishes when the disc is removed. The removal of a circularly symmetric disc leaves the shear γ_g unaffected. We are still left with the contribution of the stars within r_d. A single star at a distance l produces a shear proportional to $1/l^2$ (Equation 18). The area element $l\mathrm{d}l$ translates into a $\gamma^{-2}\,\mathrm{dg}$ probability for shear γ. We show in the appendix that the distribution of the random as well as the total shear is in fact dominated by nearest neighbour contributions and has a γ^{-2} tail.

We now discuss the effect of lensing on solid angles in the sky. As just shown, the magnification produced by a smooth galaxy does not apply to a beam (coming from a sufficiently small source) which slips in between the stars. Nevertheless, this magnification does apply to extended sources and hence to the angles between quasars on the sky. There is no paradox involved here, just a difference of scales. Fig. 3 shows rays from two point sources which undergo a relative deflection $\sim \theta_g$ as computed from the smooth galaxy potential. If either or both underwent a close encounter with a star, this would contribute extra deflections of the order of r_0 / D. The correction coming from the discreteness of the mass distribution is thus of the order

$$r_0/D\theta_g \sim (m_{\mathrm{star}} / m_{\mathrm{galaxy}})^{-1/2} \sim N_{star}^{-1/2} \tag{37}$$

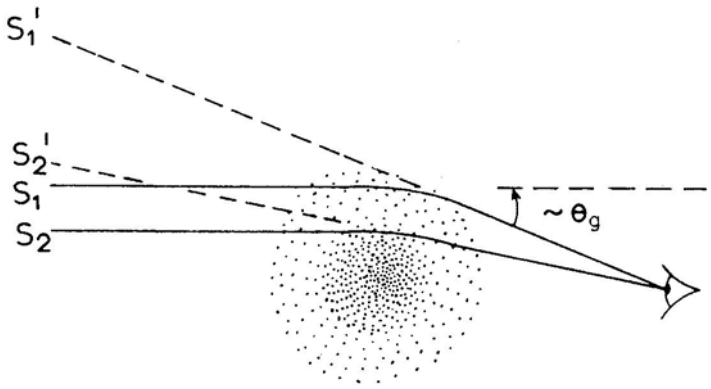


Figure 3. Magnification of the angle between two point sources S_1 and S_2 by a galaxy. S_1' and S_2' are their images. As argued in Section 5, the relative deflection of the two rays is dominated by the smooth potential of the galaxy.

and hence small. Basically, the deflection produced by a point mass falls off slowly (as $1/r$). The relative deflection of two rays separated by many stars is thus not dominated by nearest neighbours and the continuum picture applies with small corrections. Solid angles at a distance r from the centre of an isothermal galaxy are scaled up by a factor

$$A \simeq [1 - \kappa_g(r)]^{-2} \simeq 1 + 2\tau(r), \tag{38}$$

where terms of order τ^2 have been neglected. The number of quasars per unit solid angle goes down by the same factor.

It is now straightforward to deal with the case when there is a significant amount of smoothly distributed matter in the galaxy, in addition to the stars. The convergence κ_s due to the surface density of this matter should be included in (9) while calculating the amplification. As before, the stars do not contribute to κ.

6. Discussion and summary

The main purpose of this paper has been to discuss the problem of imaging background sources such as quasars by point masses (such as low-mass stars) distributed in an isothermal halo around a galaxy, stressing the compound effect of weak and strong lensing. We have tried to include in a consistent way, the lensing by a single mass as modified by its neighbours and the potential of the whole galaxy. The importance of this modification is measured by the optical depth τ and the calculations presented in this paper are valid for $\tau < 1$. In practice, this includes all lines of sight which pass more than 10 kpc from the centre of a typical massive galaxy. Two new points are (1) weak shear can strongly perturb high amplification encounters multiplying the number of images, and (2) the probability distribution of the random shear coming from neighbours of a given point mass has a tail extending to many times the typical shear. Combining these, the net effect is that an appreciable fraction of events with amplifications of the order of $\frac{1}{2}\tau^{-1}$ are further brightened by approximately one magnitude. This should be important for the QSO–galaxy association problem (Canizares 1981, VO).

The general idea that weak shear can appreciably perturb high-amplification events is applicable to the cosmological situation where quasars are lensed by galaxies along the line of sight. TOG have pointed out that there is a strong bias towards high amplification events in flux limited observations of gravitationally lensed quasars. They estimate that the average amplification of a lensed quasar could be in the range 10–40 or even higher. The basic reason for this bias is the steep fall in the luminosity function of quasars at the bright end. At a given observed flux, this favours higher amplification of intrinsically fainter sources. Such events will be sensitive to the random cosmological shear coming from the lumpiness of matter in the universe which is always present (Gunn 1967). This random shear is of the same order as the optical depth for lensing by galaxies which TOG estimate to be 0.05–0.1. We therefore have a situation where the distributed cosmologically-induced shear can play a significant role in determining image geometry and amplification. In addition, the lensing galaxy can be in a cluster as in the case of the QSO 0957 + 561 (Young *et al.* 1981). While TOG emphasize the role of the convergence produced by the cluster potential in increasing image splittings, there is also a significant shear produced by the cluster, especially when it is not centred on the beam. We hope to return to these applications in detail in a future publication.

Acknowledgements

We would like to thank J. E. Gunn for his illuminating remarks on the relation between lensing by point masses and by a continuous sheet. One of us (RN) would like to express his thanks to the Astrophysical Sciences Department, Princeton University, for support and hospitality while this work was being done, and to IAU Commission 38 for travel funds under their exchange of astronomers programme.

Appendix

Probability Distribution of the Shear by Randomly Distributed Stars

The tidal force (that is, shear) produced at the origin by a star at (r_i, θ_i) falls as r_i^{-2}. The law of superposition of shears, expressed by (18), shows that we can regard the total shear as the result of a random walk with steps of length (r_0^2/r_i^2) and direction $2\theta_i$. Chandrasekhar (1943) has reviewed the Holtsmark-Markoff method of calculating the distribution of the net displacement in such random walks and we follow it here.

It is convenient to use the following scaled variables to present the coordinates of the stars and the shear components occuring in (18),

$$\begin{aligned} &S_1 \equiv \gamma_1/n_s r_0^2 \equiv \gamma_1 \pi/\tau; \quad S_2 \equiv \gamma_2 \frac{\pi}{\tau}; \qquad X_i \equiv r_i \cos 2\theta_i / n_s^{-1/2} \\ &Y_i \equiv r_i \sin 2\theta_i / n_s^{-1/2}; \quad S^2 \equiv S_1^2 + S_2^2; \quad R_i^2 \equiv X_i^2 + Y_i^2. \end{aligned} \tag{A1}$$

Equation (18) for the shear produced by a random collection of stars then reads

$$S_1 = -\sum_i \frac{X_i}{R_i^3}; \quad S_2 = \sum_i \frac{Y_i}{R_i^3}. \tag{A2}$$

The probability distribution $P(S_1, S_2)$ of the scaled shear components is easier to compute in terms of its Fourier transform, the so-called characteristic function $Q(t_1, t_2)$ defined by

$$Q(t_1, t_2) \equiv \langle e^{i(t_1 S_1 + t_2 S_2)} \rangle \equiv \int\int \mathrm{d}S_1 \, \mathrm{d}S_2 P(S_1, S_2) e^{i(t_1 S_1 + t_2 S_2)}. \tag{A3}$$

Let all the stars which lie at distances less than r_d be included in (A2). In terms of scaled variables, we have

$$R_i < R_d = r_d / n_s^{-1/2}. \tag{A4}$$

The probability distribution of R_i is uniform over the disc $R_i < R_d$.

$$P(R_i)\,\mathrm{d}R_i = 2R_i \,\mathrm{d}R_i / R_d^2. \tag{A5}$$

The total number of points is given by

$$N = \pi r_d^2 n_s = \pi R_d^2 \tag{A6}$$

with small fluctuations since r_d is chosen to enclose many points.

The characteristic function Q in (A3)is the expectation of a product in N independent

random variables and hence factors

$$Q(t_1, t_2) = \left[\left\langle \exp\left(i\frac{t_1 X}{R^3} + i\frac{t_2 Y}{R^3} \right) \right\rangle \right]^N \equiv [q(t_1, t_2)]^N. \tag{A7}$$

It is convenient to introduce polar coordinates related to t_1, t_2 *by*

$$t_1 = t\cos\phi, \qquad t_2 = t\sin\phi \tag{A8}$$

Using (A5), one of the factors in (A7) reads

$$q(t_1, t_2) = \int \frac{d\theta}{2\pi} \exp\left\{ \frac{it\cos(2\theta - \phi)}{R^2} \right\} \frac{2R\,dR}{R_d^2}$$

$$= \frac{2}{R_d^2} \int_0^{R_d} J_0(t/R^2) R dR. \tag{A9}$$

J_0 is the usual zero-order Bessel function. Clearly, q is very close to 1, but raised to a very high power in (A7). Substituting (A9) into (A7), we have

$$Q(t, \phi) = \left[1 - \frac{2}{R_d^2} \int_0^{R_d} [1 - J_0(t/R^2)]\, R\,dR \right]^{\pi R_d^2}. \tag{A10}$$

Making the substitution

$$u = t/R^2$$

and using the integral

$$\int_0^\infty \left(\frac{1 - J_0(u)}{u^2} \right) du = 1$$

we can take the limit of large R_d in (A10) which simplifies to

$$Q(t, \phi) = e^{-\pi t}. \tag{A11}$$

As expected for circular symmetry, there is no ϕ dependence. The probability distribution of S_1 and S_2 is given by inverting the Fourier transform in (A3)

$$P(S_1, S_2) = \int\int \frac{dt_1\, dt_2}{(2\pi)^2} e^{-i(t_1 S_1 + t_2 S_2)} e^{-\pi(t_1^2 + t_2^2)^{1/2}}. \tag{A12}$$

The probability is a function of the magnitude S of the shear and can be written

$$P(S)\,dS = 2\pi S\,dS \int \frac{t\,dt}{2\pi} J_0(St) e^{-\pi t}. \tag{A13}$$

Fig. 4 shows the function $P(S)$.

We now superpose a systematic shear S_0, taken for convenience to be along the x-axis. The new probability distribution $P'(S_1, S_2)$ is obtained from the old one by a shift along the S_1 direction.

$$P'(S_1, S_2) = P(S_1 - S_0, S_2). \tag{A14}$$

The corresponding characteristic function $Q'(t_1, t_2)$ is given

$$Q'(t_1, t_2) = Q(t_1, t_2) e^{iS_0 t_1} = e^{-\pi t} e^{iS_0 t\cos\phi}. \tag{A15}$$

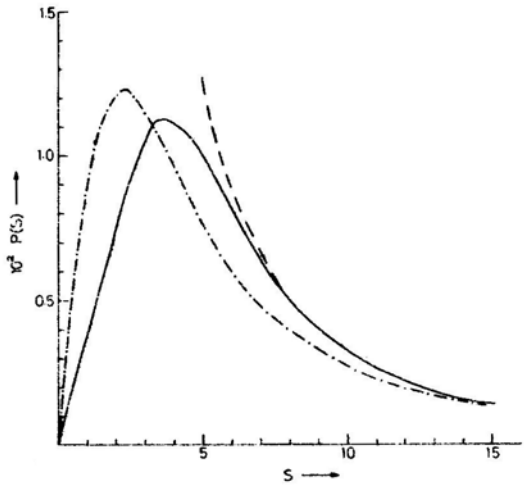


Figure 4. Probability distribution of the normalised shear S defined in (Al). The dot dashed line is the probability distribution $P(S)$ defined in (A 13) which includes only the random part of the shear. The solid line gives the probability distribution $P''(S)$ defined in (A l7) which includes the systematic shear produced by the smoothed-out galaxy potential. The dashed line gives the asymptotic form S^{-2} dominated by the contribution of one star very close to the beam.

Equations (A14) and (A15) show that the problem is no longer circularly symmetric about S_1, $S_2 = 0$. However, we are only interested in the magnitude S of the shear and can therefore average (A14) over θ, which is equivalent to averaging (A l5) over ϕ. The resulting characteristic function $Q''(t)$ and probability distribution $P''(S)$ are given by

$$Q''(t) = \int \frac{d\phi}{2\pi} e^{-\pi t} e^{itS_0 \cos\phi} = e^{-\pi t} J_0(S_0 t) \tag{A16}$$

$$P''(S) = \int \frac{t\,dt}{2\pi} e^{-\pi t} J_0(S_0 t) J_0(St) \cdot 2\pi S. \tag{A17}$$

In the case when the galaxy is entirely made out of stars with no smooth component, we can use Equation (36) to evaluate the scaled systematic shear S_0.

$$S_0 = \gamma_g(r) \cdot \pi / \tau_g(r) = \pi.$$

The probability (A17) is also plotted in Fig. 4 for this case. Also shown in the figure is the function πS^{-2}. Simple arguments show that P and P'' approach this form asymptotically. At large S, we have one star at a distance $R \ll 1$, with the shear given by

$$S = R^{-2}.$$

Since the surface density of stars has been scaled to one in (Al), the probability distribution of S is given by

$$2\pi R\,dR = \left|2\pi S^{-1/2}(-\tfrac{1}{2}) S^{-3/2}\,dS\right| = \pi S^{-2}\,dS.$$

This asymptotic form, dominated by the nearest neighbour contribution, also follows from the πt cusp at the origin in the two-dimensional Fourier transform (A11) of the probability distribution.

References

Avni, Y. 1981, *Astrophys. J.*, **248,** L95.
Barnothy, J. M., Barnothy, M. F. 1968, *Science*, **162**, 348.
Born, M., Wolf, E.1975, *Principles of Optics*, Pergamon, New York.
Bourassa, R. R., Kantowski, R.1975, *Astrophys. J.*, **195,** 13.
Canizares, C. R. 1981, *Nature*, **291,** 620.
Chandrasekhar, S. 1943, *Rev. mod. Phys.*, **15,** 1.
Chang, K, Refsdal, S. 1979, *Nature*, **282,** 561.
Dashevskii, V. M., Slysh, V. I. 1965, *Astr. Zu.*, **42,** 863: 1966, *Soviet Astr.*, **9,** 671.
Gott, J. R. 1981, *Astrophys. J.*, **243,** 140.
Gott, J. R., Gunn, J. E. 1974, *Astrophys.* J., **190,** L105.
Gunn, J. E. 1967, *Astrophys. J.*, **150,** 737.
Hawking, S. W., Ellis, G. F. R. 1973, *The Large Scale Structure of Space-Time*, Cambridge University Press.
Kittel, C. F.1966, *Introduction to Solid State Physics*, 3 edn, Wiley, New York.
Misner, C. W., Thorne, K. S., Wheeler, J. A. 1972, *Gravitation*, Freeman, San Francisco.
Peacock, J. A.1982, *Mon. Not. R. astr. Soc*, **185,** 987.
Press W. H., Gunn, J. E. 1973, *Astrophys. J.*, **185,** 397.
Refsdal, S.1964, *Mon. Not. R. astr. Soc*, **128,** 307.
Turner, E. L. 1980, *Astrophys.* J., **242,** L135.
Turner, E. L., Ostriker, J. P., Gott, J. R. 1984, *Astrophys.* J., in press (TOG).
Vietri, M., Ostriker, I P . 1983, *Astrophys. J.*, **267,** 488(VO).
Walsh, D., Carswell, R. F., Weymann, R. J. 1979, *Nature*, **279,** 381.
Young, P. J., 1981, *Astrophys. J.*, **244,** 756.
Young, P. J., Gunn, J. E., Kristian, J., Oke, J. B., Westphal, J. A. 1981, *Astrophys. J.*, **244,** 736.
Zel'dovich, Ya. B. 1964, *Astr. Zu.*, **41,** 19: 1964, *Soviet Astr.*, **8,** 13.

Note added in proof

K. Chang & S. Refsdal (1984, *Astr. Astrophys.*, **132,** 168) have recently published a detailed numerical study of Lensing by a star in a galaxy which provides additional shear and convergence. M. Vietri (private communication) has pointed out that many of the claimed cases of quasar-galaxy associations correspond to large angular separations where the probability of Lensing would be very small. We thank M. Vietri for drawing our attention to the work of Chang & Refsdal and also for his close and critical reading of the manuscript.

J. Astrophys. Astr. (1984) **5,** 251–271

Searches for Proton Decay and Superheavy Magnetic Monopoles

B. V. Sreekantan *Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Bombay 400005*

(*Invited article*)

1. Motivation

1.1 *Proton Decay*

By the mid-thirties of this century it had become clear that all matter, consisting of molecules and atoms, was reducible to just three fundamental particles—the proton, the neutron and the electron. The proton and the electron were regarded as absolutely stable and the neutron though unstable in its free state was stable when bound inside the nucleus. The discovery of the positron, the anti-particle of the electron, established the Dirac theory of the electron. This theory was based on quantum mechanics and the special theory of relativity, and could explain the behaviour of charged particles in passing through matter and almost all aspects of emission and absorption of radiation by molecules and atoms. The finer radiative effects, however, needed the development of quantum electrodynamics (QED) for their explanation. In QED, the simplest example of a guage theory, the photon or quantum of radiation mediated the electromagnetic force. At this stage the only things which seemed to need further clarification were the nature of the nuclear force—that bound the protons and the neutrons together inside the nuclei—and the presence of a very penetrating component in the cosmic radiation, and the occurrence of atmospheric cascade showers.

The discovery of the muon in 1937 solved the problem of the penetrating radiation and the discovery of the pion in 1947 was a great boost to the theory of nuclear forces based on Yukawa's ideas. The discovery of π° mesons (which are produced along with charged pions), and their decay into γ rays and muons solved the problem of cascade showers. The β-decay of the neutron and of other radioactive substances led to the postulate of a massless spin -½ particle—the neutrino. Just when one was getting the feeling that the nuclear forces were becoming tractable, and all physical phenomena could be elegantly explained on the basis of a few elementary particles, things began to happen rapidly in cosmic-ray research, and later at higher and higher energy accelerators, that changed the whole course of this area of physics. A large number of extremely unstable fundamental particles were discovered with lifetimes ranging from 10^{-8} s to 10^{-22} s and with masses extending upto 10 GeV; some of them were fermions and some were bosons, and some of them required additional quantum numbers like strangeness, charm *etc.* Their number kept on increasing with the increase in the energy of the accelerators.

These particles were classified on the basis of their masses as leptons, the light ones; mesons of medium mass, and baryons, the heavy ones. Among these the strongly interacting particles are known as hadrons. Examples of hadrons are baryons, *e.g.* the proton, the neutron, Λ°, Σ^+ *etc.* and mesons, *e.g.* the pion, the kaon, ρ, ω *etc.* The

electron, the positron, $\mu^{\pm}$, $\tau^{\pm}$ and neutrinos ν_e, ν_μ, ν_τ make up the lepton family. Each particle has a corresponding anti-particle.

While classification of the hadrons (whose number rose to hundreds) according to their grouping into charge multiplets, super-multiplets and the eightfold way *etc.* brought some order into this jungle of fundamental particles, the real simplification and progress however came with the introduction of the 'quark' model and its development into the theory known as 'quantum chromodynamics' (QCD). In this picture all the baryons and mesons are composites of more elementary entities called 'quarks' and their antiparticles 'antiquarks'. There are six 'flavours' of quarks, and in each flavour there are three varieties distinguished by their 'colour'. The properties are given in Table 1 which also illustrates the quark compositions of some of the hadrons.

According to the quark model, the proton is a composite of two up-quarks and a down-quark (*uud*), the three quarks having different colours such that they combine to make the proton 'colourless'. The pion, say the π^-, is a combination of a quark and an antiquark ($d\bar{u}$), and so on. Thus the hundreds of hadrons reduce to combinations of just 18 quarks and 18 antiquarks. Clearly the spin, the charge and the baryon number of quarks have been adjusted to give the observed properties of the combinations—the mesons and the baryons. The strange (*s*), the charmed (*c*), the top (*t*), and the bottom (*b*) quarks are necessary to account for the new particles *K*, Λ, Ψ/J, *Y*, *etc.* which have nonzero values of special quantum numbers like the strangeness.

In QCD, the quark-quark forces are mediated by the exchange of eight massless vector bosons called 'Gluons'. Each gluon carries a colour charge, *i.e.* the strong charge. The absorption or emission of a gluon by a quark changes its colour. This is a major difference between QED and QCD. In QED, the charge of a particle is unchanged by emission or absorption of a photon since the photon carries no electric charge. An important property of the QCD force is that it depends on the momentum carried by the gluon, and the higher the momentum the weaker the force. This feature of the force has two important consequences. It makes the quarks behave as free particles at extremely short distances—the so-called property of asymptotic freedom. It is also believed that the force increases with distance leading to permanent confinement of quarks within the hadrons. Enormously large energies would be required to free the quarks from the particles in which they are bound. In such attempts, quark-antiquark

Table 1. Properties assigned to the different flavours of quarks.

Flavour	Spin	Charge	Baryon number	Charm number	Strangeness number	Mass
d	1/2	$-\frac{1}{3}e$	1/3	0	0	300 Mev
u	1/2	$\frac{2}{3}e$	1/3	0	0	300 MeV
s	1/2	$-\frac{1}{3}e$	1/3	0	1	500 MeV
c	1/2	$\frac{2}{3}e$	1/3	1	0	1500 MeV
t	1/2	$\frac{2}{3}e$	1/3	0	0	Not seen yet
b	1/2	$-\frac{1}{3}e$	1/3	0	0	5000 MeV

Typical combinations: $p = uud$, $\Omega^- = sss$, $\Delta^+ = uuu$, $\pi^- = d\bar{u}$, $\Delta^- = ddd = d_r d_g d_b$, $K^+ = u\bar{s}$, $Y = bb$, $D = c\bar{u}$, $\bar{c}d$, $F = c\bar{s}$.

combinations would be generated resulting in the emergence of bound particles rather than free quarks. This would explain the anomaly that free quarks have not been observed so far in searches with cosmic rays and accelerators. QCD which is a gauge theory has been successful in explaining many of the features of high-energy interactions observed at the accelerators. The nuclear force that binds the protons and neutrons in the nuclei in this scheme, emerges as a residue of the quark-quark forces.

Another exciting development that has taken place in the last two decades is the unification of the weak and electromagnetic forces, again as a gauge theory. While in QED the electromagnetic interactions are considered to be mediated by photons, the weak interactions in the electro-weak theory are assumed to be mediated by massive intermediate vector bosons $W^{\pm}$ and Z^0. As was already noted in the context of QCD, in non-Abelian gauge theories (of which the electro-weak is also one) the identity of the particle is changed when the mediating bosons are absorbed or emitted.

An electron may emit a charged boson (W^-) and get transformed into a neutrino. In the unified electro-weak theory, the photon, the $W^{\pm}$, and the Z^0 belong to the same family. At extremely high energies ($> 10^{11}$ GeV) the strength of their interactions becomes identical. The force is then long range and the weak coupling constant is the same as that of the electromagnetic force. Below this energy, due to a mechanism known as spontaneous symmetry breaking, the $W^{\pm}$ and Z^0 acquire mass, and the long-range character of the weak force becomes an extremely short-range one. The Weinberg–Salam electro-weak theory made specific predictions—(i) the existence of neutral currents involving the exchange of the neutral Z^0 particles, and (ii) the mass of $W^{\pm}$ as 81 GeV and that of Z^0 as 93 GeV. Experimentally, the neutral currents were established in 1974. The $W^{\pm}$ and Z^0 with precisely the masses predicted, were discovered at the CERN $\bar{p}p$ Collider in 1983, thus giving the final stamp of success to the electro-weak theory.

The initial success of the gauge theories in unifying electromagnetic and weak forces propelled many theorists (Pati & Salam 1973; Georgi & Glashow 1974; Georgi, Quinn & Weinberg 1974) to explore the feasibility of constructing a single gauge theory of all the three forces—the strong, the weak and the electromagnetic—particularly since strong interactions had been incorporated into a successful gauge theory, *i.e.* QCD. Pati (1983) gives an excellent account of the historical development, the current status and future prospects of these theories with an exhaustive list of references. Experimental support for such unification came from the inelastic scattering experiments of electrons and neutrinos on nucleons, and from the study of the behaviour of coupling constants of strong and electro-weak interactions with increasing energy transfer. In these theories, all the coupling constants converge to a common value at the unification energy of 10^{15} GeV as shown in Fig. 1.

While there area number of models $SU(4)^4$, SU(5), SO(10), E_c, E_7, SU(7) *etc.*, named after the mathematical group of symmetries that connect the forces, in the following we shall consider the simple SU(5) model to illustrate the general direction of progress.

In the SU(5) model, there are 24 vector bosons that couple 24 different currents. The photon, $W^{\pm}$, Z^0 and the 8 gluons of QCD, constitute a subset of 12 vector bosons. The other 12 are new massive leptoquarks (X, Y) which carry both flavour and colour, three of charge $-(1/3)e$ and three of charge $-(4/3)e$ and the six corresponding antiparticles. It is the leptoquarks that bring about the interaction between the quarks and leptons and their mass is around 10^{15} GeV/c^2, the unification energy. The two important predictions of the grand unification theories (GUTs) are: (i) the proton (and the

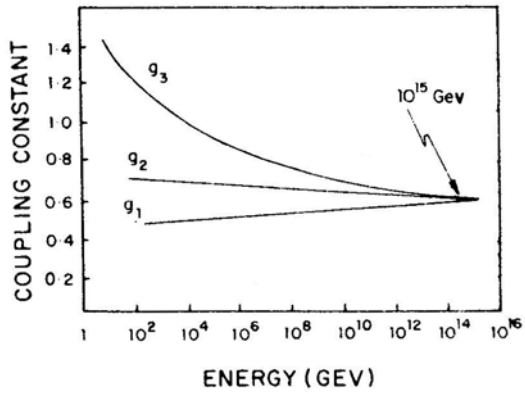


Figure 1. The energy dependence of the strong and electroweak coupling constants showing convergence at ~ 10^{15} GeV. While g_3 is the strong coupling constant, g_1 and g_2 are linear orthogonal combinations of the electromagnetic and weak coupling constants.

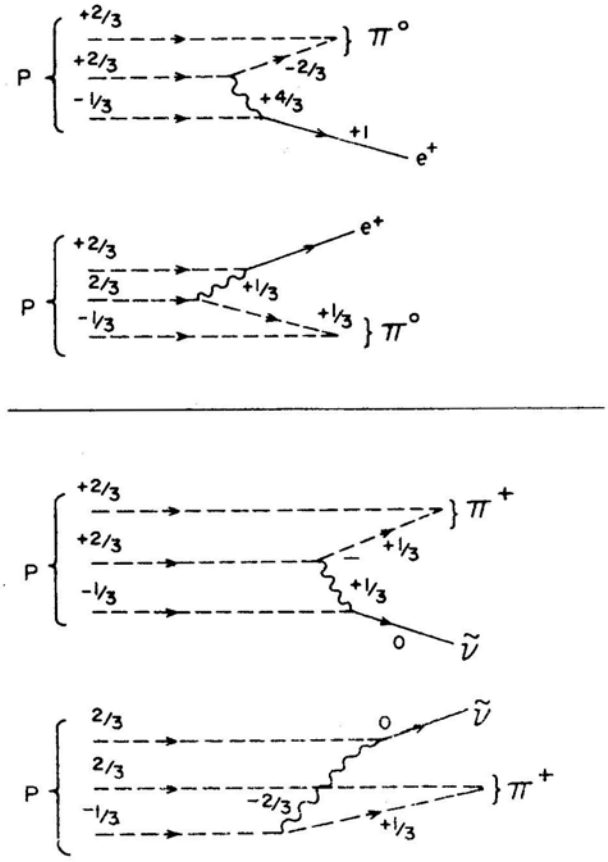


Figure 2. Emission and absorption of leptoquarks through quark-quark interactions leading to proton decay.

neutron even if it were bound) must decay violating baryon number conservation, and (ii) super-heavy magnetic monopoles must exist.

The typical mechanism of proton decay brought about by the emission and absorption of leptoquarks by quarks is illustrated in Fig. 2. For example, by the emission of a leptoquark of charge (4/3)*e*, the quark with charge (2/3)*e* transforms into an antiquark of charge — (2/3)*e*, which together with a quark of charge (2/3)*e* becomes a π^0. The emitted (4/3)*e* leptoquark is absorbed by the –(l/3)*e* quark of the proton and transformed into a positron. A leptoquark which plays this dual role is called a diquark. Note that the electric charge and colour are conserved in these reactions throughout.

The probability of this type of interaction depends on the mass of the *X* particle. In fact, the lifetime for decay is proportional to the fourth power of M_X, the mass of the *X*–particle *i.e.*,

$$T_p = \frac{k}{\alpha^2} \cdot \frac{M_X^4}{m_p^5}$$

where $k \sim 1$ and $\alpha = 0.02$. If $M_X = 2 \times 10^{14}$ GeV/c^2 then $T_P = 3 \times 10^{29}$ yr.

Table 2. Predictions for branching ratios for proton and neutron decay into the major two-body decays in the SU(5) model.

Proton decay:

Mode	(a) REC	(b) NR	(c) R	(d) R	(e) R	(f) NR	(f) REC	(f) R
$e^+\pi^0$	31	37	9	13	31	36	40	38
$e^+\rho^0$	15	2	21	20	21	2	7	11
$e^+\eta$	11	7	3	.1	5	7	1.5	0
$e^+\omega$	18	18	56	46	19	21	25	26
$\nu_e^c\pi^+$	12	15	3	5	11	14	16	15
$\nu_e^c\rho^+$	6	1	8	7	8	1.0	2.6	4
μ^+K^0	1	19	—	7	.5	18	8	5
$\nu_\mu^c K^+$	2	0	—	.5	—	0	.2	.6

(a) Machacek 1979; (b) Gavela *et al.* 1981a, b; (c) Donoghue 1980; (d) Golowich 1980: (e) Din *et al.* 1980; (f) Kane & Karl 1980.

Neutron decay:

Mode	(a) REC	(b) NR	(c) R	(d) R	(e) NR	(e) REC	(e) R
$\nu_e^c\pi^0$	5	8	2	3	8	7	7
$\nu_e^c\rho^0$	3	.5	5	4	.6	1.2	1.8
$\nu_e^c\eta$	2	1.5	1	—	1.5	—	—
$\nu_e^c\omega$	3	3.5	14	10	5	5	5
$e^+\pi^-$	54	74	23	32	79	72	68
$e^+\rho^-$	27	4	55	48	6	12	19
$\nu_\mu^c K^0$	0	10	—	2	1.1	3	0.6

(a) Machacek 1979; (b) Gavela *et al* 1981a, b; (c) Donoghue 1980; (d) Golowich 1980;
(e) Kane & Karl 1980.
See Langacker (1982) for details of references.

The various decay modes of protons and neutrons according to SU(5) are given in Table 2. The range of branching ratios for the different decay modes is due to the differences in the approaches of different authors, and the parameters used by them (Langacker 1982). It is important to note that in the SU(5) model, the dominant lepton secondary in the decays of the proton and the neutron, is the electron and not the muon.

1.2 *Grand Unification Monopoles*(*GUMS*)—*Catalysed Proton Decays*

In his attempt to understand the quantization of electric charge, half a century ago Dirac (1931) had proposed the existence of magnetic monopoles with magnetic charge a multiple of $g = \hbar c/2e = (137/2)e$. Experimental searches over the 50 years since then have not provided any evidence for them. However, the advent of grand unification theories have revived the interest in the field. Breaking of the GU Symmetries down to SU(3) × SU(2) × U(l) predicts the existence of the t'Hooft-Polyakov type magnetic monopoles (t'Hooft 1974; Polyakov 1974) with masses of the order of the grand unification mass 10^{16} GeV/c^2, which corresponds to 10^{-8} g—the mass of a bacterium. The GUT monopole, or GUM as it is called, has a long-range magnetic field. However, being a massive quark condensate with a core of 10^{-30} cm or so, it will need to have a state of perfect symmetry that could have existed perhaps only immediately after the big bang, when quarks and leptons would have been identical as also photons, heavy vector bosons and gluons. In such a state the gauge hierarchy gets broken in stages as one proceeds outwards from the core, and the various particles begin to assume their identities. Thus if GUMs are available, they will constitute a wonderful laboratory for exploration of GUT effects.

Rubakov (1981) and independently Callan (1982a, b) have shown that the grand unification monopole, when it passes through matter, can induce proton decay through reactions of the type $M + p \rightarrow Me^+\pi^0$, $M\mu^+K^0$, $Me^+\mu^+\mu^-$ *etc.* The monopole would come out unscathed in the reaction, but would cause the break up of the proton in a manner identical to what happens in proton decay. A surprising feature of this phenomenon that has made it extremely interesting from the point of view of the experimentalist, is that the cross-section according to Rubakov and Callan for this catalysed proton decay is of the same order as a strong interaction cross-section ($\lambda \sim 30$ cm in iron). If this is the case, then multiple decays should be recorded in a few metres of matter traversed by a GUM. Wilczek (1982), however, has suggested that the cross-section may be more like that of a weak interaction.

1.3 *Leptoquarks*, *the Grand Unification Monopoles* (*GUMS*) *and the Early Universe*

The masses of leptoquarks and GUMs being of the order of or more than 10^{15} GeV/c^2, it is extremely unlikely that these will ever be produced in terrestrial accelerator laboratories. It has been estimated by an enterprising scientist that a linear accelerator for the purpose would have to extend from the earth to the moon. This being the case, the question arises as to whether these super-massive particles remain only as predictions of GUTs, or whether there exists any possibility that they could have been produced at some point of time in the history of the universe and played a role in its

evolution, perhaps, even surviving up to the present time as relics. It turns out that if we believe m the big-bang origin of the universe, then at times less than 10^{-35} s, conditions were suitable for production of these massive particles.

This extrapolation to such small values (over 53 decades of time) is based upon current ideas of the universe that have emerged from a variety of astronomical observations in the different bands of the electromagnetic spectrum. Constrained by observations of the rate of expansion of the universe, the temperature and density of the universal microwave background and the cosmic abundance of light nuclei, and using all of physics relevant to the different stages of the evolution of the universe it has become possible to construct scenarios of what happened immediately after the big bang. This is illustrated in Fig. 3 which is adapted from Schramm (1983). As is evident from it, the astrophysics of what followed the big bang is best described in the language of modern particle physics.

In the time interval between 10^{-43} to 10^{-35} s after the big bang, the temperature of the universe could have been higher than 10^{12} GeV (10^{25} K) and the density greater than 10^{74} g cm^{-3} sufficiently high to produce particles of mass equal to or greater than 10^{15} GeV/c^2. It is precisely in this interval that the super-heavy particles and the monopoles could have been produced. If they were, two astrophysical puzzles automatically get solved.

Based on the measurement of the microwave background and the estimated amount of matter in the universe, the ratio of the number of photons to baryons is 10^9–10^{10}; the universe is dominated by radiation rather than matter. Also, all attempts to detect anti-matter in the primary cosmic radiation have resulted only in setting upper limits, clearly showing that in the universe matter dominates over antimatter. In the framework of GUTs, both these large-scale asymmetries can be explained on the basis of a perfectly symmetric origin of the universe. In one of the typical models, it is assumed that initially the massive X, Y particles and their antiparticles were produced in equal numbers; their decays led to creation of quarks and leptons and their antiparticles. As the temperature

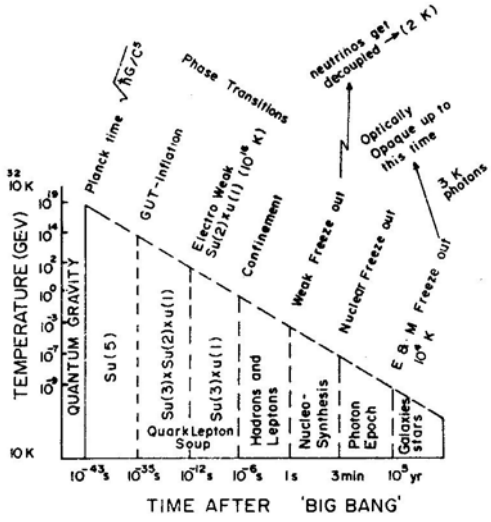


Figure 3. Physics immediately after the big bang at different time intervals—symmetry breaking and phase transitions.

fell, nucleons and antinucleons were produced but not in exactly equal numbers. Their annihilation produced the observed background radiation, and the excess baryons which survived constitute the matter observed now in the universe. By a suitable adjustment of asymmetry in the decay of the X, Y particles and their antiparticles, and introducing in addition the violation of CP symmetry, the observed photon-to-baryon ratio as well as the limits on the matter-to-antimatter ratio can be explained.

In the early universe, super-heavy magnetic monopoles would also have been produced during the symmetry breaking phase transitions. The only way these monopoles could have disappeared is by $M\bar{M}$ annihilation, but a fraction could have survived. It is therefore very important to search for monopoles and measure their flux. Their detection would be of tremendous significance for both physics and astrophysics.

1.4 *Super-Unification—SUSYGUTs*

In the grand unified theory, the strong, electromagnetic and weak forces are unified as also quarks and leptons. A natural extension of this would be to bring gravity also into the same fold. In molecular, atomic and nuclear phenomena, the role of gravity is insignificant compared to that of the other three forces. But as we have seen, extrapolation to the early universe leads us to a situation where the density becomes enormously high even compared to nuclear densities. Thus at less than 10^{-43}s, quantum gravity effects must also come into play and will require a quantum theory of gravity.

One promising attempt in this direction is the supergravity theory. In it, supersymmetry, *i.e.* the symmetry between fermions and bosons, is introduced as a gauge symmetry. As a consequence, boson partners of all fermions must exist and vice-versa. Thus the super-unification of supergravity with a GUT (SUSYGUT) necessarily leads to new spin $\frac{1}{2}$ fermionic partners for all bosons including the gauge and Higgs bosons. It also predicts new spin 0 particles corresponding to quarks and leptons which have been christened squarks and sleptons. In SUSYGUTs the proton decay lifetime may be extended to 10^{32} yr and even beyond. In the simplest supersymmetric SU(5) type theories, the dominant nucleon decay modes will be $p \rightarrow \tilde{\nu}K^+$, $\tilde{\nu}_\tau K^+$, $n \rightarrow \tilde{\nu}_\mu K^0$, $\tilde{\nu}_\tau K^0$, while the decay modes $e^+\pi$, e^+K, $\mu^+\pi$, $\nu\pi$, *etc.* will be suppressed compared to the predictions of the earlier models.

2. Design of proton decay experiments

The design of proton-decay experiments depend on (1) the estimated lifetime, (2) the decay modes and the energy distribution among the secondaries, (3) the background events that stimulate proton decay.

If we take the lifetime as about 10^{30} yr, then one proton will decay in 30 tons of matter per year. Clearly, to cover the predicted range 10^{29}–10^{32} yr the sensitive mass of the proton-decay detector should be several hundred to several thousand tons.

As we have seen in the SU(5) type model the dominant decay mode is $p \rightarrow e^+\pi^0$, with the electron carrying an energy of about 460 Me V and the individual γ rays which result from the π^0 decay, carrying 240 Me V each. The configuration of the event will be a cascade of 460 Me V due to the electron in one direction and two cascades in the

opposite direction, typically of 240 Me V each with an opening of ~ 40°. On the other hand, if we consider a decay mode of the type $p \rightarrow e^+\omega^0$, where $\omega^0 \rightarrow \pi^+\pi^-\pi^0$, then the energies are 145 Me V for the electron, 240 Me V each for π^+ and π^-, and 135 Me V each for 2γ rays that result from $\pi^0 \rightarrow 2\gamma$. The configuration of the event will appear isotropic without any forward-backward peaking of tracks. The π^+ will give rise to the decay chain $\pi^+ \rightarrow \mu^+\nu_\mu$, $\mu^+ \rightarrow e^+\nu_\mu\nu_e$. Thus μ–e decay can be recorded by measurement of delay in the 0.5–10 μs range. The decay mode $p \rightarrow \mu^+\pi^0$ will have the muon of energy of 465 Me V in one direction and a double cascade with an opening angle of 40^0 in the diametrically opposite direction.

The main background in proton decay experiments is from cosmic-ray secondaries produced in the atmosphere. By installing the detectors deep underground, these effects can be minimised. Clearly, the deeper one is able to go the better is the elimination of these effects, especially those due to muons. However, the intensity of the more serious background due to cosmic-ray-produced neutrinos does not decrease by going to large depths. The background problem has been thoroughly discussed in the papers by Krishnaswamy *et al.* (1982a, b). It is shown there that for dominant decay modes such as $p \rightarrow e^+\pi^0$, $e^+\omega^0$, (ρ^0) and $n \rightarrow e^+\pi^-$, $e^+\rho^-$ *etc.* the background will be entirely due to inelastic interactions of neutrinos of a few Ge V energy. The rate of such events is estimated to be one event per year in 60 tons of active detector at equatorial latitudes and a factor of 1.5 higher at higher latitudes ($\lambda > 35°$). This background rate reduces by a factor of 10 if the back-to-back configuration of the events in the decay modes of the type $p \rightarrow e^+\pi^0$ are clearly established.

3. Dedicated proton decay experiments

Two different approaches have been employed in the experiments that are currently operational to detect proton decay.

3.1 *Fine Grain Calorimeters*

In the first method, the source of protons and neutrons are the iron nuclei in stacks of iron plates and iron and other nuclei in concrete blocks interspersed in a matrix of either proportional counters or gas discharge tubes, as in the experimental set-ups of KGF (Krishnaswamy *et al.* 1981, 1982a, b, 1983a, b), NUSEX (Battistoni *et al* 1982, 1983) and Soudan Groups (Peterson *et al* 1983; Bartlet *et al* 1983). In these experiments the individual tracks of the secondaries and the cascade electrons are recorded and a complete re-construction of the configuration of the event to good accuracy is feasible. The total energy is determined from the range of the tracks and in the case of soft cascades from the total track length. One of the disadvantages of high-atomic-number materials like iron is that the decay products—the pions and the kaons—either get absorbed in the nucleus itself or get appreciably scattered. The back-to-back configuration of tracks typical of two-body decay is thus distorted to a certain extent. The deviation from 180° could be as much as ±40°.

The KGF detector set-up at a depth of 7600 m.w.e. (metres water-equivalent) and operating since 1980 November, comprises of 34 layers of proportional counters (1600 in all) with 1.2-cm thick iron plates in between. The total weight of iron is about 140 tons inclusive of the iron walls of the counters. The counters have a cross-sectional area

of 10 cm × 10 cm and are 6m long in the case of counters laid parallel to the walls of the tunnel and 4 m in the case of the alternate layers placed orthogonal to these. From each counter triggered, the ionisation is measured over the range $\frac{1}{3} I_{min}$ to $100 I_{min}$. Since 1982 December, timing information has been available to an accuracy of 0.5 microseconds which enables the identification of μ_e decays associated with the triggered event.

The NUSEX detector operating in the Mont Blanc tunnel since 1981 August is at a depth of 5000 m.w.e. It has 134 horizontal slabs of iron each of thickness 1 cm and of surface area 3.5 m × 3.5 m, interleaved with planes of extruded plastic tubes each 1 cm × 1 cm in cross section and filled with CO_2 and N-pentane operated in the limited streamer mode. The total weight of the assembly is about 160 tons.

The SOUDAN I experiment installed at a depth of 1800 m.w.e. has been operating since 1981 October. It has 3456 proportional counters each of diameter 2.8 cm and length 2.9 m arranged in 48 layers with 4 cm spacing and embedded in a taconite concrete block of dimensions 3m× 3m × 2m. It has a total mass of 31 tons.

3.2 *Water Cerenkov Detectors*

In the second method adopted by one Japanese and two U.S. groups, the source of protons and neutrons is the particle detector itself, a large tank of water. The principle of the method is illustrated in the Fig. 4. Let us consider the decay mode $p \rightarrow e^+\pi^0$. The electron will produce a small cascade which gives rise to a cone of Cerenkov radiation as illustrated. Similarly, the two γ rays that result from the π^0 decay give rise to two cones of radiation in the opposite direction with an angle of ~ 40° between them. From the configuration of the photomultipliers and the amplitudes of light pulses the details of the events such as the location of the vertex, the energy of the secondaries and the relative angles between them can be worked out. One clear advantage of the Cerenkov method is that the direction of the particles can be determined unambiguously. The main disadvantage is that only particles above the threshold velocity will give rise to

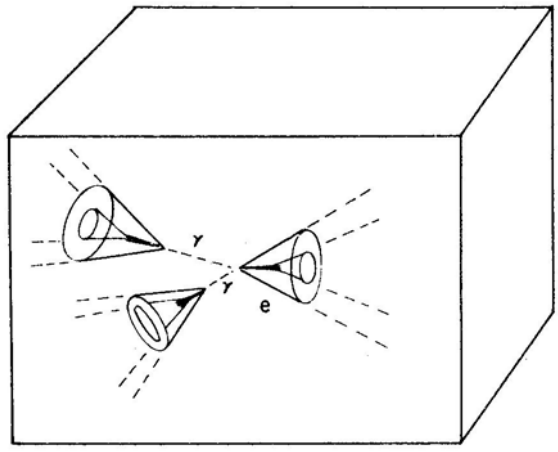


Figure 4. Illustration of the principle of water Cerenkov detector for proton decay. In the decay mode $p \rightarrow e_+ \pi^0$, three cones of light will be produced in the water due to the development of cascades by the positron and the two γ-rays from the decay of π^0.

Cerenkov radiation and therefore the method is not efficient for some of the decay modes. In a medium like water 80 per cent of the decays will take place in oxygen nuclei and they suffer from the same disadvantage as in iron nuclei discussed earlier. However, the remaining 20 per cent of the decays take place in hydrogen and the secondaries do not suffer the nuclear effects.

In the IMB experiment (Bionta *et al.* 1982) set up at a depth of 1570 m.w.e. in the Morton Salt Mine near Cleveland, Ohio, the water tank has dimensions 22.5 m × 17 m × 18 m and holds 8000 tons of water. All the six faces of the tank are lined with photomultipliers of 12.5 cm diameter and spaced 1 metre apart (Fig. 5). In all, there are 2048 photomultipliers. The times of arrival of light at the different photomultipliers are recorded to an accuracy of 11 nanoseconds. The accuracy of vertex position determination is about ± 60 cm, and the opening angle uncertainty is ± 15°. The energy estimate is accurate to ± 10 per cent.

The HPW experiment (Cline 1982) is set up in the Silver King Mine near Park City, Utah, with an overburden of rock 1600 m.w.e. The cylindrical tank (38 ft diameter and 24 ft high) holds 800 tons of water. 704 photomultipliers are distributed throughout the volume of the water. The relative look angle of the photomultipliers are so adjusted that the full 4π solid angle is covered for any event occuring anywhere in the central fiducial

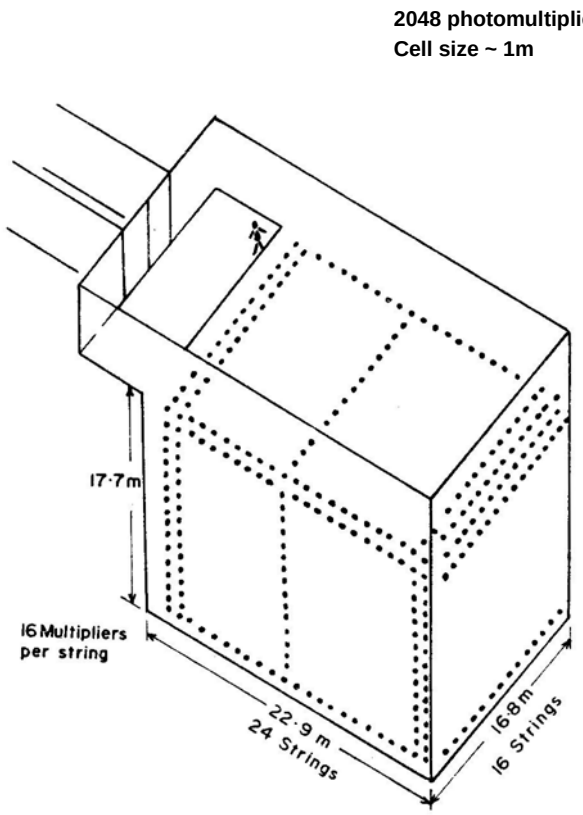


Figure 5. Details of the IMB water Cerenkov detector operating in the Morton salt mines at Fair Port, Ohio at a depth of 1600 m.w.e.

volume. The estimated accuracy of vertex location is 12–40 cm, the decay angle ~ 14° and energy resolution ~ 20 per cent.

The Kamioka proton decay detector (A. Arisaka *et al.* 1984, personal communication) is located in a Zn/Pb mine, 300 km west of Tokyo where the rock overburden is 2700 m.w.e. The capacity of the water tank is 3000 tons, and the fiducial mass 1000 tons. The 1050 photomultipliers with a spherical cathode, specially designed by the Hamamatsu company in Japan are each 20 inches in diameter. A unique feature is that 20 percent of the walls are covered by photomultipliers compared to 1 percent in the IMB experiment. Timing information is available only in the microsecond region and enables the identification of the associated μ-e decay events with 70 per cent efficiency.

4. Results from proton decay experiments

The status of results from the different experiments till 1983 August may be summarised as follows:

The K.G.F. experiment (Krishnaswamy *et al.* 1983a, b) operating since 1980 November has reported three partially confined and three fully confined events, the details of which are given in Table 3. Orthogonal views of the three fully confined and one unconfined event are given in Fig. 6. The cut-away diagrams of events No. 587 (Fig. 7) and 877 (Fig. 8) illustrate the details available for each event.

Among these, event 587 has a high probability of being a neutrino interaction if it is interpreted as a single cascade. However, the detailed features favour the interpretation in terms of two cascades developing in opposite directions in which case the probability for the event to be a neutrino interaction is reduced by a factor of 10.

Event 877 is the strongest candidate for nucleon decay. It has one non-showering track of range 135 g cm $^{-2}$ and in the opposite direction either a shower or a pair of charged particles. In the first case it is consistent with the decay mode $n \rightarrow e^{+}\pi^{-}$ and in the second case with the interpretation $p \rightarrow \mu^{+}K_s^0$ with $K_s^0 \rightarrow \pi^{+}\pi^{-}$.

Event 867 is a single non-showering track suffering a large-angle scattering of 37°. The track could be a pion of 450 Me V corresponding to the decay mode $p \rightarrow \nu\pi^{ch}$ or a kaon of 650 Me V corresponding to $p \rightarrow \nu K^{ch}$ and $K^{ch} \rightarrow \mu^{ch}\nu$. On the basis of these three events, the lifetime is estimated as $\tau / BR \sim 1.7 \times 10^{31}$ yr. These limits are based on 60 tons fiducial mass and 1.9 yr operation. The neutrino background simulating these events is estimated to be less than 0.3, compared to the observed 3 events.

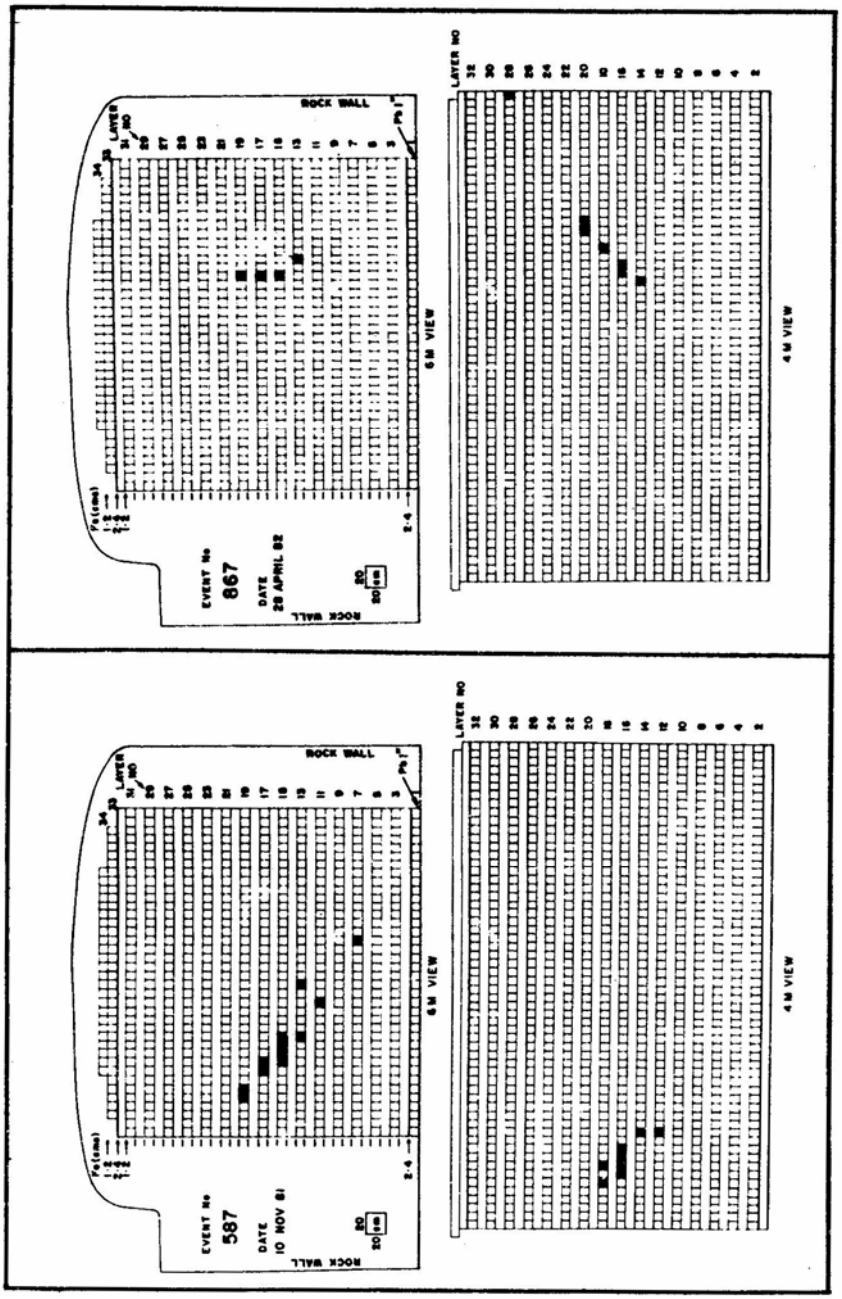
In the Mont Blanc experiment (64 tons fiducial mass, operation time 8 months) 5 fully confined events with a total track length in iron > 10 cm were recorded. The first four events have been interpreted as neutrino interactions. The fifth event cannot easily be interpreted as a neutrino interaction, or as a neutron interaction with any significant probability.

The SOUDAN I experiment with 31.5 tons of iron-loaded concrete has observed one fully confined event in 0.48 years of operation. The estimated total energy of the event is 650 ± 25O Me V. The event appears to have two prongs at about 135° to each other. The absence of hits in counters in certain intermediate layers suggest that part of the event is an electromagnetic cascade. Monte-Carlo simulations by the Soudan group show that the spread in the observation energy of nucleon decay events could be 820–940 Me V depending on the decay mode. The event therefore is consistent with a nucleon decay resulting in two-prong emission.

Table 3. Summary of all the candidate events for nucleon decay recorded in the KGF experiment.

	Event no.	Decay mode	Particle range ($g\,cm^{-2}$)	Energy estimates (MeV)[a]	Other plausible modes	ν-background events/1.3 yr
fully confined events	587	$p \to e^+\pi^0$	$e^+ \sim 115$ $\pi^0 \sim 150$	$U_e \sim 500$ $U_{\pi^0} \sim 500$	—	<0.5[b]
	867	$p \to \bar{\nu}\pi^+$	ABC(π^+) ~ 182	$U_\pi \sim 435$	$p \to \nu K^+$, $K^+ \to \mu^+\nu_\mu$	<0.05
	877	$n \to e^+\pi^-$	AB(e^+): 104	$U_e \sim 400$	$p \to \mu^+ + K^0_s$, $K^0_s \to \pi^+\pi^-$	<0.05
			BC(π^-): 135	$U_\pi \sim 450$	$U_\mu \sim 340$ MeV $U_{\pi^\pm} \sim 320{,}220$ MeV	
partially confined events	87	$n \to e^+\pi^-$	[c]	$U_{e^+} \geqslant 500$ $U_\pi \sim 400$	$p \to e^+\pi^0(2\gamma)$ $U(\gamma_1)$ $U(\gamma_2)$	<0.05
	251	$p \to e^+\rho^0$	[c]	$E_{all} > 540$	$p \to \mu + K^0_s$, $K^0_s \to \pi^+\pi^-$	<0.05
	722	$p \to e^+\omega^0$ $\to 3\pi$	A: ~ 52 B: > 170 C: ~ 15 D(e^+) ~ 10	$E_A \sim 245$ $E_B \sim 335$ $E_C \sim 55$ $E_D \sim 80$	—	<0.2

[a] The symbols E and U refer to visible energy loss (excluding particle mass) and the total energy respectively. [b] Estimated rate for single cascade. It would be much lower for the preferred interpretation of two cascades starting at the mid-point of the event. [c] See Krishnaswamy *et al.* (1981).


ROCK WALL
LAYER NO
EVENT No
867
DATE
28 APRIL 82
ROCK WALL
6 M VIEW
4 M VIEW
ROCK WALL
LAYER NO
EVENT No
587
DATE
10 NOV 81
ROCK WALL
6 M VIEW
4 M VIEW

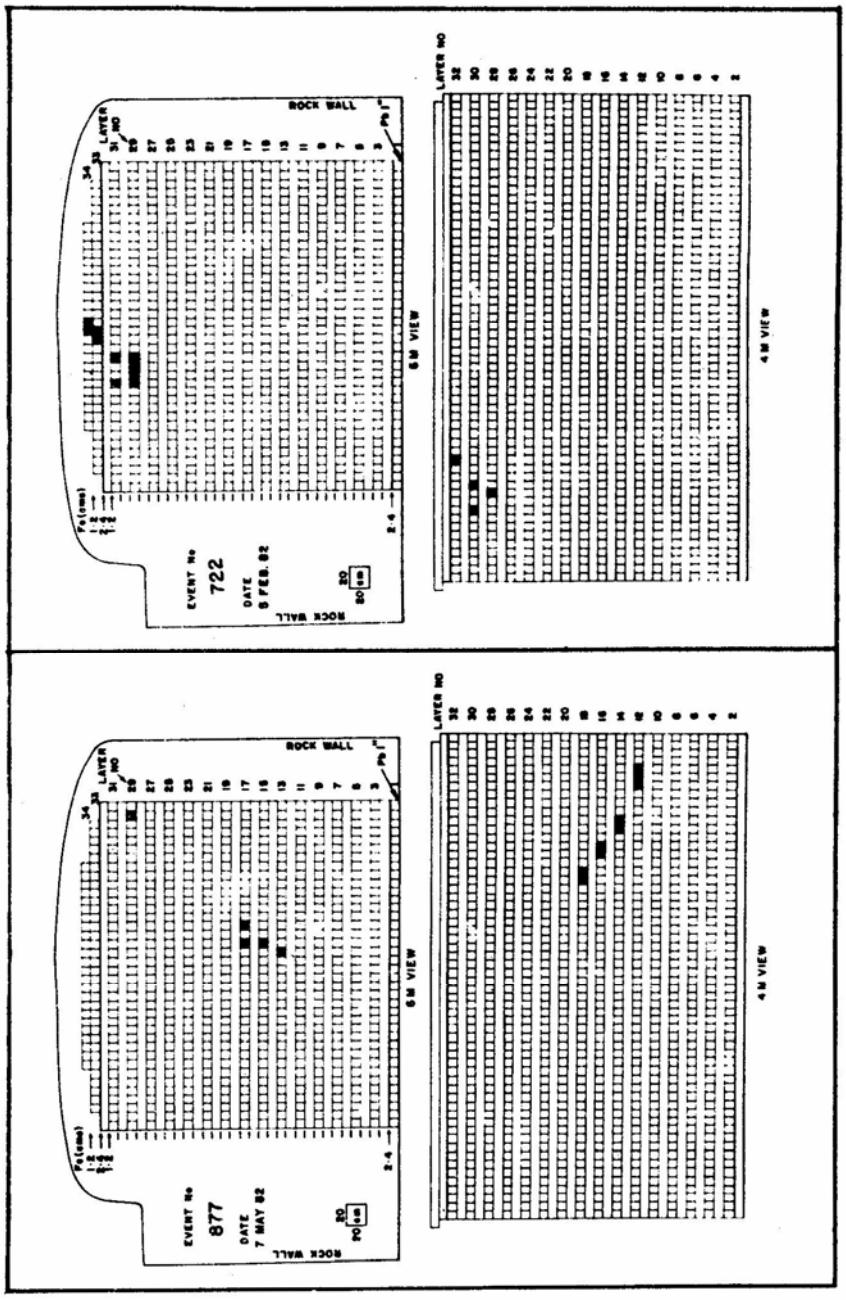


Figure 6. The two orthogonal views of the fully confined events No. 587, 867 and 877 and one partially confined event No. 722 in the KGF proton decay detector. Each black square represents one counter triggered above the threshold value. The full details of the events 587 and 877 are given in Figs 7 and 8.

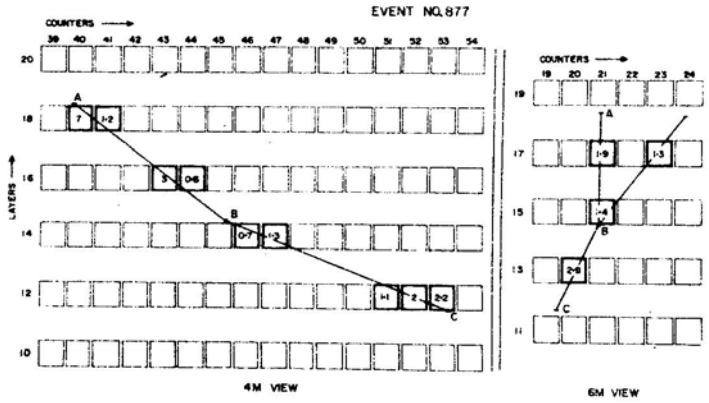


Figure 7. The cut-away diagram of event No. 587. The numbers within the squares give the ionisation recorded in the counters in terms of the ionisation of a vertical minimum ionising particle. The lines indicate the plausible configuration of the event consistent with the interpretation of $p \rightarrow e^{+} \pi^{0}$, $\pi^{0} \rightarrow 2\gamma$. The shower seen in the layers 15–19 corresponds to e^{+} with energy ~ 500 Me V while the two photons with energy ~ 200 and 300 Me V give rise to the shower in the downward direction for layer 15.

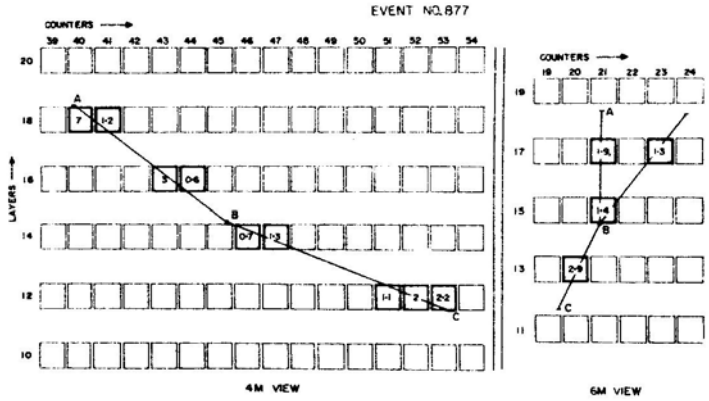


Figure 8. The cut-away view of event No. 877. Two alternative interpretations have been given by the authors (Krishnaswamy *et al.* 1982a, b): (i) $n \rightarrow e^{+}\pi^{-}$. The shower AB is due to the positron e^{+} and the track BC the charged pion, (ii) $p \rightarrow \mu^{+} K_{s}^{0}$. The track BC is due to μ^{+} (350 Me V) and the upward-forked branches correspond to two pions π^{+}, π^{-} (320, 220 Me V) from the decay of K_{s}^{0}.

In the IMB experiment, in the first 80 days of operation with a fiducial mass of 3300 ions, 69 fully confined events were recorded with at least 45 photomultipliers recording above threshold in each event. Within the statistical accuracy, the above events were uniform in vertex position and isotropic in track direction. The fraction of fully contained events with μ-e decay signature was 0.4 ± 0.1 after correcting for detection efficiency.

The characteristics of the 69 confined events have been summarised as follows:

(1) 66 single or multi-track events which do not possess a track lighting more than 40

tubes in the backward hemisphere and hence are outside the angle and energy requirements for $p \rightarrow e^+\pi^0$.
(2) Two wide-angle track events both of which are associated with μ-e decay and hence do not qualify as $p \rightarrow e^+\pi^0$.
(3) One two-track event in which the total number of photomultiplier hit were 340 about a factor of 2 greater than expected for $p \rightarrow e^+\pi^0$. Also the opening angle was 115 ± 15° outside the predicted range (> 140°) for free or bound proton decays in the mode $e^+\pi^0$.

Since there is no candidate, $\tau/BR > 6.5 \times 10^{31}$ yr for (free + bound) and $\tau/BR > 1.9 \times 10^{31}$ yr for free protons only.

Since no candidate events have been seen even in a further analysis of data of 130 days operation, these limits have been pushed to $> 10^{32}$ yr for all and $> 6.5 \times 10^{31}$ yr for protons only.

At the recent ICOBAN 84 (International Colloquium on Baryon Non-conservation) held at Park City, Utah in January 1984, several groups presented candidates for nucleon decay with due reservation and caution.

The first results based on 176 days of operation from 1983 July 6 were presented by the Kamioka group, corresponding to 324 ton-years. Out of a total of 65726 events selected for visual scan and analysis, 57 events were in the fiducial volume, 40 with single rings and 17 with multiple rings (rings of light are produced when the Cerenkov cones intercept the walls of photomultipliers). Out of the 40 single rings, two had the characteristics to be interpreted as $p \rightarrow \nu K^{ch}$ with $K^{ch} \rightarrow \mu^{ch}\nu$. Out of the 17 multiple ring events one three-ring event could be a case of $p \rightarrow \mu^+\eta$ or μK^0 ($K^0 \rightarrow 2\pi^0$) or $n \rightarrow e^+\rho^-$ ($\rho^- \rightarrow \pi^-\pi^0$) and one five-ring event that could be $p \rightarrow e^+\omega^0$ or $n \rightarrow e^+\rho^-$. The experiment sets a lower limit of $\tau/BR = 2.6 \times 10^{31}$ yr for $p \rightarrow e^+\pi^0$ decay mode.

In the HPW experiment, one candidate event with two muons which could be a case of $p \rightarrow \mu^+K^0$, $K^0 \rightarrow \pi^+\pi^-$ and $\pi^+ \rightarrow \mu^+\nu_\mu$ has been seen.

NUSEX have reported a new event which could either be $p \rightarrow e^+\pi^0$ or a neutrino interaction $\nu_e \rightarrow e$.

In the KGF experiment, a fourth confined event has been seen which could be interpreted as either $n \rightarrow \nu\eta^0$, $\eta^0 \rightarrow 3\pi^0$ or $n \rightarrow e^+\rho^-$, $\rho^- \rightarrow \pi^-\pi^0$.

In the light of all this, what seems clear is that the lifetime for the decay of a nucleon is higher than 10^{31} yr and the dominant decay mode is not $e^+\pi^0$. If the lifetime were of the order of 10^{29} to 10^{30} yr and the dominant mode $e^+\pi^0$, then, by now several of the experiments already in operation would have established the decay without any ambiguity. Looking at the candidate events reported by the different groups, it is perhaps possible that decay modes with muon as one of the secondaries may equal or even have a higher probability than electron secondaries.

The reservation and caution in the interpretation of the events as candidates for nucleon decay stem from the fact that there is no unique signature for decay that cannot be simulated by a neutrino interaction. The total energy of the event (940 Me V) and the back-to-back configuration are the two chief characteristics that have been used to reduce the probability of a candidate event from being a neutrino interaction. In these circumstances it has to be recognised that the nucleon decay phenomenon will not be established on the basis of clear-cut individual events that can have no other explanation, as happened in the case of discoveries of various fundamental particles and their decays. The final proof has to come on the basis of a large number of events which, in the energy spectrum plot of neutrino-induced events, will stand out unambiguously

around 1 Ge V energy. From this point of view, it is necessary to have good energy resolution, angular resolution and good rate of candidate events. Consistent with the high rate of potential events, it is unlikely that the energy and angular resolution can be improved beyond what has been achieved with the IMB and Kamioka set ups as far as water Cerenkov detectors are concerned.

The situation with regard to the fine-grain calorimeters is not quite the same. There is considerable scope for improvement in this type of detectors.

The Frejus experiment, which is a collaboration between Orsay, Palaseau, Saclay and Wuppertal (Barloutaud 1982, 1983) and is to go into operation in 1984 in the Modane underground laboratory (4500 m.w.e.) east of Grenoble in France, is a step in this direction. The detector is a fine-grain calorimeter made up of 0.5 cm ×0.5 cm × 6 m flash tubes and 3-mm thick iron plates with Geiger counters of cross-sectional area 1.5 cm × 1.5 cm interspersed in the stack every one metre and will serve as the triggering detectors. The total mass of iron will be 1.5 kilo-tons. Because of the fine-grain nature, the directions of muons for example from kaon decay can be measured by the increase in multiple scattering, and the positrons from μ^+ decay can be seen by using a long HT pulse on the flash chambers. The energy resolution expected is 12–20 per cent for pions of 200–300 Me V. Nucleon decay experiments planned for the future have been reviewed by Grant (1983). The second phase of the KGF experiment, the Soudan II and the Grand Sasso project will all incorporate improved, larger versions of fine-grain calorimeters. Liquid scintillation counters, high-pressure and liquid-argon detectors working in TPC-like modes, and even a 3000-ton liquid-argon bubble with 50 per cent duty cycle have been under consideration to achieve the requisite fine structure and energy resolution.

5. Limits on flux of superheavy magnetic monopoles (GUMs)

Excepting in the experiment of Cabrera (1982) in which the magnetic monopole is detected by the change in magnetic flux as it passes through a superconducting ring, all other experiments that have been carried out recently depend on the velocity of the

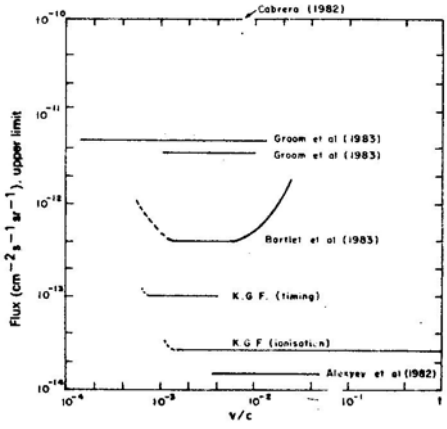


Figure 9. Comparison of the upper limits on the flux of GUT monopoles in various experiments as a function of velocity.

magnetic pole and its ionisation characteristics. They are therefore sensitive only in certain velocity ranges. Cabrera's first experiment carried out with a superconducting loop of area 20 cm^2 gave evidence of one candidate in an operating period of 151 days which corresponds to a velocity independent flux limit of 6.1×10^{-10} cm^{-2} s^{-1} sr^{-1}. In a subsequent experiment Cabrera *et al.* (1983), using a three-loop superconducting device of effective area 476 cm^2, did not find any candidate in 150 days of operation thus setting an upper limit of 3.7×10^{-11} cm^{-2} s^{-1} sr^{-1}.

The limits set by different experiments (Groom *et al.* 1983, Bartlet *et al.* 1983, Krishnaswamy *et al.* 1983a, b, Alexeyev *et al* 1983) sensitive to different velocity ranges are shown in Fig. 9. It is clear that the flux in the velocity range 10^{-3} c to c is less than 2×10^{-14} cm^{-2} s^{-1} sr^{-1} at least three orders of magnitude lower than the limit to the velocity independent flux given by Cabrera *et al.* The experiments will soon be able to reach the bound of 10^{-16} cm^{-2} s^{-1} sr^{-1} set by Parker (1970) on the basis of the existence of a galactic magnetic field.

5.1 *Monopole Catalysis of Nucleon Decay*

Searches for super-heavy magnetic monopoles catalysing nucleon decay have been made with practically all the nucleon-decay experiments in progress. In the IMB experiment, in 100 days of operation, no event was recorded that satisfied the criteria of multiple nucleon decays from the passage of a magnetic monopole. Fig. 10 shows the upper limit (90 per cent confidence) on monopole flux for different velocities as a function of the catalysis cross-section. In the KGF detector, a chain of nucleon decays can be seen if the second and subsequent events occur within 7 μs of the trigger, and the separation between the decays is less than the dimensions of the detector. In 2.52 years

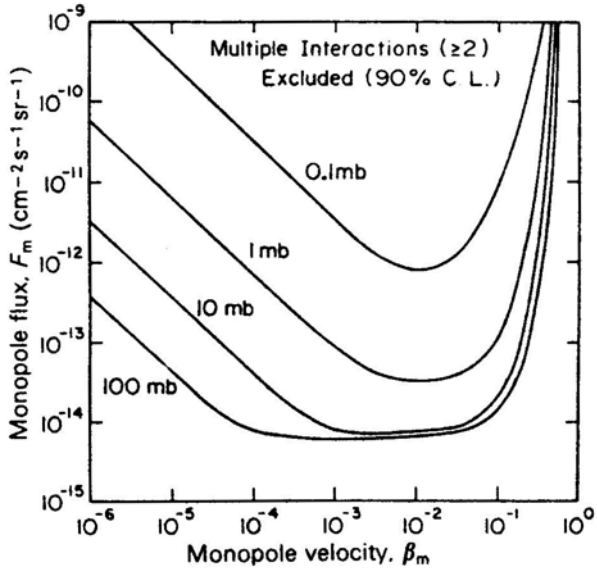


Figure 10. Limits on monopole flux deduced from the IMB experiment on the basis of Rubakov-Callan effect and different cross-sections for induced decay (Bionta *et al.* 1982).

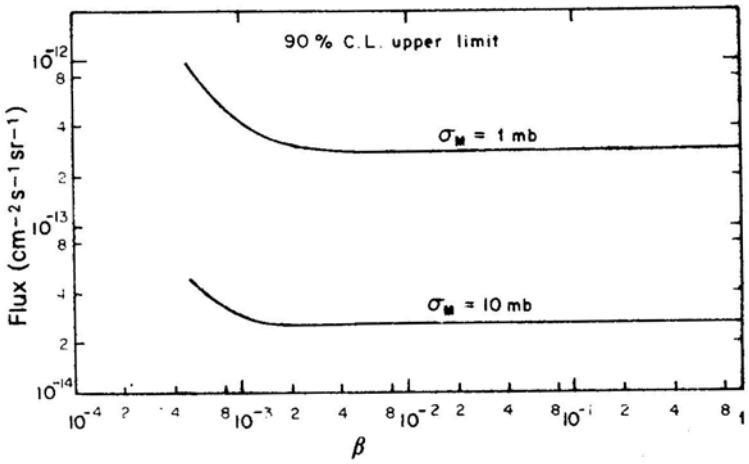


Figure 11. Limits on the monopole flux deduced from the KCF experiment based on two values for induced proton decay (Krishnaswamy *et al.* 1983a, b)

of operation there is no evidence for chains of two or more than two decays. Fig. 11 shows the 90 per cent confidence upper limit on the monopole flux for chain decays as a function of velocity and for two values of catalysis cross-section, 10 mb and 1 mb.

6. Conclusion

In conclusion, as of 1984 January, it may be stated that the existence of super-heavy monopoles and the phenomenon of nucleon decay, both of which are extremely important from the point of view of grand unification theories, are still very open questions. While there has been just one magnetic monopole candidate so far, there have been several as far as nucleon decay is concerned. The first candidates for nucleon decay came from the fine-grain calorimeters of KGF, and NUSEX; recently there have been candidates from the water Cerenkov experiments as well. The experimental situation regarding the other important phenomena of relevance to grand unification which we have not discussed in this article—like the finite mass of neutrinos, neutrino oscillations, and neutron oscillations—continues to be indefinite though many dedicated experiments are in progress.

With the continued operation of the nucleon decay experiments already collecting data and the commissioning of the new generation of experiments over the next few years, the stage is set for a resolution of this problem in a time scale of 5–10 years. The present indication that the dominant decay mode for the proton (even if it decays) is not $p \rightarrow e^{+}\pi^{0}$ and that the lower limit to the lifetime of the nucleon is 10^{31} yr, does not favour the simple SU(5) type models.

The remarkable discoveries of $W^{\pm}$ and Z^{0} with mass values exactly as predicted, have given a boost to the unification based on the gauge theoretical approaches. Whether grand unification can be extended to super-unification, experiment alone can tell. This will be the challenge for the remaining years of this century.

Acknowledgement

I would like to express my thanks to Professor V. S. Narasimham for helpful discussions.

References

Alexeyev, E. N., Boliev, M. M., Chudakov, A. E., Makoev, B. A., Mikheyev, S. P., Sten'kin, Yu, V. 1982, *Lett. Nuovo Cimento*, **35,** 413.
Bartlet, J. *et al.* 1983, *Phys. Rev. Lett.*, **50,** 651.
Barloutaud, R. 1982, in *Int. Coll. Baryon Nonconservation*, Eds V. S. Narasimham, P. Roy, K. V. L. Sarma, B. V. Sreekantan, Indian Academy of Sciences, Bangalore, p. 143.
Barloutaud, R. 1983, in *Int. Coll. Matter Nonconservation*, INFN, Rome, p. 177.
Battistoni, *G. et al* 1982, in *Int. Coll. Baryon Nonconservation*, Eds V. S. Narasimham, P. Roy, K. V. L. Sarma, B. V. Sreekantan, Indian Academy of Sciences, Bangalore, p. 83.
Battistoni, G. *et al.* 1983, in *Int. Coll. Matter Nonconservation*, INFN, Rome, p. 107.
Bionta, R. M. *et al.* 1982, in *Int. Coll. Baryon Nonconservation*, Eds V. S. Narasimham, P. Roy, K. V. L. Sarma, B. V. Sreekantan, Indian Academy of Sciences, Bangalore, p. 137.
Cabrera, B. 1982, *Phys. Rev. Lett.*, **48,** 1378.
Cabrera, B. *et al.* 1983, *Phys. Rev. Lett.*, **51,** 1933.
Callan, C. G. 1982a, *Phys. Rev.*, **D25,** 2141.
Callan, C. G. 1982b, *Phys. Rev.*, **D26,** 2058.
Cline, D. B. 1982, in *Int. Coll. Baryon Nonconservation*, Eds V. S. Narasimham, P. Roy, K. V. L. Sarma, B. V. Sreekantan, Indian Academy of Sciences, Bangalore, p. 99.
Cobb, J. H. 1983, in *Int. Coll. Matter Nonconservation*, INFN, Rome, p. 88.
Dirac, P. A. M. 1931, *Proc. R. Soc. London*, **A133,** 60.
Georgi, H., Glashow, S. L. 1974, *Phys. Rev. Lett*, **32,** 438.
Georgi, H., Quinn, H. R., Weinberg, S. 1974, *Phys. Rev. Lett.*, **33,** 451.
Grant, A. L. 1983, in *Fourth Workshop on Grand Unification*, Birkhäuser, Boston, p. 69.
Groom, D. E. *et al* 1983, *Phys. Rev. Lett.*, **50,** 573.
Krishnaswamy, M. R., Menon, M. G. K., Mondal, N. K., Narasimham, V. S., Sreekantan, B. V., Hayashi, Y., Ito, N., Kawakami, S., Miyake, S. 1981, *Phys. Lett.*, **106B,** 339.
Krishnaswamy, M. R., Menon, M. G. K., Mondal, N. K., Narasimham, V. S., Sreekantan, B. V., Hayashi, Y., Ito, N., Kawakami, S., Miyake, S. **1982a,** *Pramana*, **19,** 525.
Krishnaswamy, M. R., Menon, M. G. K., Mondal, N. K., Narasimham, V. S., Sreekantan, B. V., Hayashi, Y., Ito, N., Kawakami, S., Miyake, S. 1982b, *Phys. Lett.*, **115B,** 349.
Krishnaswamy, M. R., Menon, M. G. K., Mondal, N. K., Narasimham, V. S., Sreekantan, B. V., Hayashi, Y., Ito, N., Kawakami, S., Miyake, S. 1983a, in *Int. Coll. Matter Nonconservation*, INFN, Rome, p. 97.
Krishnaswamy, M. R. *et al.* 1983b, in *Fourth Workshop on Grand Unification*, Birkhäuser, Boston, p. 25.
Langacker, P. 1982, in *Int. Conf. Baryon Nonconservation*, Eds V. S. Narasimham, P. Roy, K. V. L. Sarma, B. V. Sreekantan, Indian Academy of Sciences, Bangalore, p. 34.
Parker, E. N. 1970, *Astrophys. J.*, **160,** 383.
Pati, J. C. 1983, in *Int. Coll. Matter Nonconservation*, INFN, Rome, p. 45.
Pati, J. C, Salam, A, 1973, *Phys. Rev. Let.*, **31,** 661.
Peterson, E. *et al.* 1983, in *Int. Coli. Matter Nonconservation*, INFN, Rome, p. 80.
Polyakov, A. M. 1974, *JETP Lett.*, **20,** 194.
Rubakov, V. A. 1981, *JETP Lett.*, **33,** 644.
Schramm, D. N. 1983, *Physics Today*, **36,** No. 4, 27.
t'Hooft, G. 1974, *Nucl. Phys.*, **B79,** 276.
Wilczek, F. 1982, *Phys. Rev. Lett.*, **48,** 1144.

J. Asrophys. Astr. (1984) **5,** 273–276

A Source for the Reissner-Nordstrøm Metric

Ramesh Tikekar *Department of Mathematics, Sardar Patel University, Vallabh Vidyanagar 388120*

Received 1983 August 23; accepted 1984 March 30

Abstract. An exact solution of the coupled Einstein-Maxwell equations representing the gravitational field in the interior of a sphere of charged incoherent matter in equilibrium is obtained which is a charged analogue of the static perfect fluid sphere solution with spheroidal 3-space obtained by Vaidya & Tikekar.

Key words: charged matter sphere—general relativity

1. Introduction

Vaidya & Tikekar (1982) have shown that the relativistic space-times which have the associated 3-spaces obtained as hypersurfaces t = const., 3-spheroids, are suitable to describe the gravitational field in the interior of superdense spherical stars in which the collapse under gravitational attraction is countered by repulsive fluid pressure. They given an exact solution of Einstein's field equations, hereafter referred to as VTS, representing the gravitational field of a static fluid sphere, choosing a particular spheroidal geometry for the associated 3-space. It is generally suggested that in the absence of fluid pressure, the collapse under gravitational attraction of spherical distributions of incoherent matter to a point singularity can be avoided if the matter is accompanied by some charge. Equilibrium configurations of incoherent matter wherein the equilibrium is maintained by Coulombian electrostatic repulsive force have been studied by Bonnor (1960, 1965), De & Raychaudhari (1968). Cooperstock & Cruz (1978) have obtained the charged analogue of Schwarzschild interior solution, the associated 3-space of which has the geometry of a 3-sphere. The external gravitational field of such spherical distributions is described by the Reissner-Nordstrom metric.

$$\begin{aligned} \mathrm{d}s^2 = & -(1-2m/r+q^2/r^2)^{-1}\,\mathrm{d}r^2 - r^2\,\mathrm{d}\theta^2 - r^2\sin^2\theta\,\mathrm{d}\phi^2 \\ & +(1-2m/r+q^2/r^2)\mathrm{d}t^2. \end{aligned} \tag{1}$$

We have reported here an exact solution of coupled Einstein-Maxwell equations representing an equilibrium configuration of spherical charged distribution of incoherent matter with the geometry of its spatial section identical with that of the VTS, *i.e.*, a charged analogue of VTS, as an interior source for Reissner-Nordstrom metric (1).

2. Charged analogue of VTS

Following Cooperstock & Cruz (1978), for a spherically symmetric charged perfect

fluid distribution in equilibrium with

$$ds^2 = -e^{\lambda(r)}dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta\, d\phi^2 + e^{\nu(r)}dt^2, \tag{2}$$

as the metric of the associated space-time, we adopt the coupled Einstein-Maxwell equations in the form

$$-8\pi p + E^2 = -e^{-\lambda}\left(\frac{\nu'}{r} + \frac{1}{r^2}\right) + \frac{1}{r^2}, \tag{3}$$

$$-8\pi p - E^2 = -\frac{e^{-\lambda}}{2}\left(\nu'' + \frac{\nu'^2}{2} + \frac{\nu' - \lambda'}{r} - \frac{\lambda'\nu'}{2}\right), \tag{4}$$

$$8\pi\rho + E^2 = e^{-\lambda}\left(\frac{\lambda'}{r} - \frac{1}{r^2}\right) + \frac{1}{r^2}. \tag{5}$$

Here ρ and p respectively denote the matter density and the fluid pressure of the charged fluid. The 'electric field intensity' E is determined by the defining equation

$$E(r) = \frac{q(r)}{r^2} = \frac{1}{r^2}\int_0^r 4\pi r^2 e^{\lambda/2}\sigma\, dr, \tag{6}$$

where σ denotes the charge density of the fluid and $q(r)$ denotes the charge enclosed within the sphere of radius r.

The system of equations (3)–(5) admits VTS as an exact solution when $E = 0$, which is obtained by setting

$$e^{\lambda} = \frac{(1 + 2r^2/R^2)}{(1 - r^2/R^2)}, \quad r < R, \tag{7}$$

in the space-time metric (2). Vaidya & Tikekar have shown that the hypersurfaces t = const, of the space-time of metric (2) with e^{λ} as given by (7) have the geometry of a 3-spheroid immersed in a four-dimensional Euclidean space.

It is found that the metric (2) with e^{λ} as given by (7) and

$$e^{\nu(r)} = A^2 \frac{\exp\left[-2\sqrt{2}\tan^{-1}\{2(1 - r^2/R^2)/(1 + 2r^2/R^2)\}^{1/2}\right]}{\left[(1 + 2r^2/R^2)^{1/2} + (1 - r^2/R^2)^{1/2}\right]^2}, \tag{8}$$

where A is an arbitrary constant, provides an analytic solution of the system of equations (3)–(5) with $p = 0$, representing a spherical distribution of charged incoherent matter in equilibrium.

The matter density, the charge density and the electric field intensity of this distribution have the following expressions:

$$8\pi\rho = \frac{2}{r^2}\frac{(1 - r^2/R^2)^{1/2}}{(1 + 2r^2/R^2)^{1/2}}\left[1 - \frac{(1 - r^2/R^2)^{1/2}}{(1 + 2r^2/R^2)^{1/2}}\right] + \frac{6/R^2}{(1 + 2r^2/R^2)^2}, \tag{9}$$

$$8\pi\sigma = \pm 8\pi\rho \tag{10}$$

$$E^2 = \frac{\left[(1 + 2r^2/R^2)^{1/2} - (1 - r^2/R^2)^{1/2}\right]^2}{r^2(1 + 2r^2/R^2)} \tag{11}$$

The matter density decreases radially out ward from its central value

$$8\pi\rho(0) = \frac{9}{R^2}, \tag{12}$$

obtained from Equation (9) in the limit $r \to 0$. Equation (11) implies that the electric field strength E is zero at the origin in the limit $r \to 0$ and increases steadily outward as one moves away from the centre. Hence Equations (7) and (8) provide an exact solution of (3)–(5) with $p = 0$, which is regular everywhere in the region $r < R$. The constant R is determined by Equation (12) in terms of the central value of the matter density as in the case of VTS.

3. Discussion

We consider the space-time metric (2) with Equations (7) and (8), as describing the field in the interior of a charged dust sphere in equilibrium at radius $a < R$. The field in the exterior region will be represented by the Reissner-Nordstrøm metric (1). Since the electric field intensity E is continuous at all points $r < R$ the appropriate boundary conditions to be satisfied at the boundary $r = a$, state

$$e^{\nu(a)} = \frac{1-a^2/R^2}{1+2a^2/R^2} = 1 - \frac{2m}{a} + \frac{q^2}{a^2}, \tag{13}$$

where m and q respectively denote the total gravitational mass and the total charge of the distribution. Equation (6) determines the total charge of the distribution q at radius a as

$$q = \left[1 - \frac{(1-a^2/R^2)^{1/2}}{(1+2a^2/R^2)^{1/2}}\right] a. \tag{14}$$

Subsequently, Equation (13) determines the constants A and m as

$$A^2 = \frac{(1-a^2/R^2)}{(1+2a^2/R^2)} \left[(1+2a^2/R^2)^{1/2} + (1-a^2/R^2)^{1/2}\right]^2 \times \exp\left[2\sqrt{2}\tan^{-1}\left\{\frac{2(1-a^2/R^2)}{(1+2a^2/R^2)}\right\}^{1/2}\right], \tag{15}$$

$$m = \frac{3}{2}\frac{a^2/R^2}{(1+2a^2/R^2)}\, a + \frac{q^2}{2a}. \tag{16}$$

Substitution for q from Equation (14) into (16) gives

$$m = \frac{3}{2}\frac{a^2/R^2}{(1+2a^2/R^2)}\, a + \frac{a^2/R^2 + 2\left[1-(1+a^2/R^2-2a^4/R^4)^{1/2}\right]}{2(1+2a^2/R^2)}\, a. \tag{17}$$

The second term here represents the contribution to the total gravitational mass of the charged matter sphere from the electrostatic field energy. We will denote this term by m_{el}.

The static fluid sphere of VTS has the total gravitational mass at radius a.

$$m_{\mathrm{VTS}} = \frac{3}{2}\frac{a^2/R^2}{(1+2a^2/R^2)}\, a. \tag{18}$$

Accordingly, adopting the same values for a and a/R for the charged matter sphere as in VTS, we find that the charged matter sphere has

$$m = m_{\mathrm{VTS}} + m_{\mathrm{el}}. \tag{19}$$

The contribution m_{el} from the electrostatic field energy

$$m_{el} \simeq \frac{9}{8} \frac{a^4/R^4}{(1+2a^2/R^2)} a, \tag{20}$$

will be significant for higher values of a/R only, since $a/R < 1$.

The boundary radius of the fluid sphere of VTS is restricted by the condition

$$a/R \leqslant 0.5567, \tag{21}$$

arising out of physical requirements such as $\rho > 0$, $p \geqslant 0$, $\rho - 3p \geqslant 0$. However, the solution of Einstein-Maxwell equations obtained here permits the charged matter sphere to have boundary radius exceeding the limit imposed by the inequality (21). Nevertheless, the boundary radius is restricted by the condition

$$a/R \leqslant 0.8940, \tag{22}$$

(accurate up to four decimal places) which arises from the physical requirement $\rho > 0$ to be satisfied throughout the distribution.

Hence, for the values of a/R restricted in accordance with (21), the solution of Einstein-Maxwell equations representing spherical distribution of charged incoherent matter can be considered as the charged analogue of VTS.

Acknowledgement

The author would like to thank the University Grants Commission of India, for an assistance grant, supporting the work.

References

Bonnor, W. B. 1960, Z. *Phys.,* **160,** 59.
Bonnor, W. B. 1965, *Mon. Not. R. astr. Soc,* **129,** 443.
Cooperstock, F. I., de la Cruz, V. 1978, *Gen. Relativ. Gravitation,* **9,** 835.
De, U. K., Raychaudhari, A. K., 1968, *Proc. R. Soc. London Ser. A*, **303,** 97.
Vaidya, P. C., Tikekar, R. 1982, *J. Astrophys. Astr*., **3,** 325.

J. Astrophys. Astr. (1984) **5,** 277–283

Astrophysical Boosters

W. Kundt *Institut für Astrophysik der Universität Bonn, Auf dem Hügel 71, 5300 Bonn, FRG*

Received 1984 January 30; accepted 1984 April 27

Abstract. Constraints are derived on the acceleration of charges in shocks to highly relativistic energies. When applied to the extended extragalactic radio sources and to the cosmic rays, they cast doubt on the mechanism of 'in-situ acceleration', both for energy, entropy and statistical mechanics reasons.

Key words: radio galaxies—shock acceleration—cosmic rays

1. Sources of high-energy particles

Our cosmic neighbourhood is capable of accelerating protons and/or ions to energies in excess of 10^{20} eV — corresponding to. Lorentz factors γ larger than 10^{11} —and of accelerating electrons to Lorentz factors in excess of 10^6, perhaps even 10^8. The former are observed directly as 'cosmic rays' by particle detectors, or indirectly via scintillation events and air showers. The latter are often inferred from the upper cutoff frequency ν_u in non-thermal spectra:

$$\nu_u = \mathrm{e}\, B_\perp\, \gamma^2/\pi m_e c = 5.6 \times 10^2 \gamma^{\,2}\, B_{-4}\ \mathrm{Hz}, \tag{1}$$

where $B_\perp$ is the transverse magnetic field strength, with $B^{\,-4} := B/10^{-4}$ G. ν_u reaches optical frequencies in the jets of some extragalactic radio sources, like Cen A, M87, 3C273, NGC 1097 and several others, (Schreier, Gorenstein & Feigelson 1982), and also in young pulsars. What engines can achieve such high energies at reasonable efficiencies?

In a microscopic description, a charge e can only gain energy ΔE with respect to an inertial frame by falling through an electric field **E**:

$$\Delta E = e \int \mathbf{E} \cdot \mathrm{dx} \tag{2}$$

We are not aware of large electrostatic fields in the interplanetary, interstellar or intergalactic medium, and do not expect such to exist because of the high prevailing conductivities. But **E** may belong to the outgoing wave radiated by a rotating magnet, as envisaged for pulsars, or E may be the electric field $-\boldsymbol{\beta} \times \mathbf{B}$ seen by a charge when a magnetic field **B** is convected across it at velocity $\mathrm{v} = c\,\boldsymbol{\beta}$:

$$\Delta E = e \int (\boldsymbol{\beta} \times \mathbf{B}) \cdot \mathrm{dx} \tag{3}$$

For pulsars, the highest available voltages are probably those across a certain fraction of the speed-of-light cylinder, of radius c/Ω, whence (*cf* Kundt 1981a)

$$E_{\mathrm{PSR}} \lesssim e\, B_s\, R\, (\Omega R/c)^2 = 10^{17}\, B_{13} \Omega^2{}_2\ \mathrm{eV}. \tag{4}$$

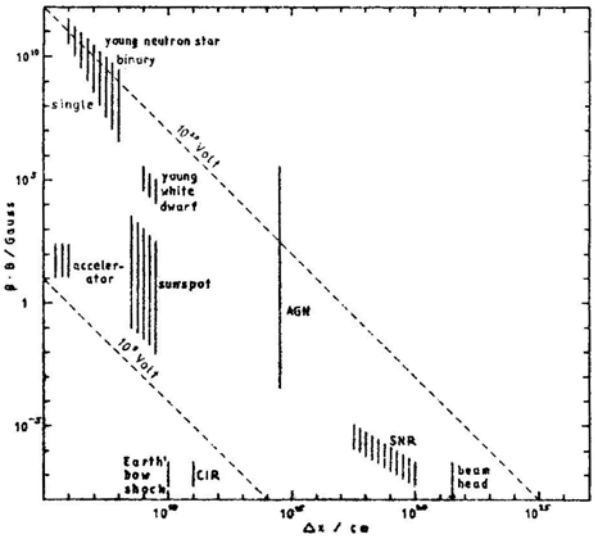


Figure 1. Estimated values of $\beta B/$ Gauss versus Δx for astrophysical boosters (*cf.* Equations 3–5 and 19). $\beta = v/c$ ranges from 1 (for wavelike phenomena) through $\Omega R/c$ (for rotators) down to some 10^{-3} (for old SNRs) and is often quite uncertain (*e.g.* for sunspots). The largest uncertainty rests in the distances Δx which particles can traverse during the acceleration event; for instance, the entries 'SNR' and 'beam head' may have to he moved to the left by four and seven scale units respectively (factors 10^4 and 10^7 in Δx). For a repeated number Z of independent accelerations, an entry should be 'raised' by the factor $\sqrt{Z}$.

The broken lines show voltages $\Delta E/e$. A similar diagram is contained in Hillas (1983).

For interplanetary shocks, field strengths $B_\perp$ of $\lesssim 10^{-4}$ G and velocities $\lesssim 10^3$ km s $^{-1}$ lead to maximal energy gains per particle

$$E_{max}^{shock} \simeq e\beta B_\perp \Delta x \simeq 10^6 (\beta B)_{-7} (\Delta x)_{10} \text{ eV}, \tag{5}$$

in agreement with measurements of both electrons and protons near the Earth's bowshock [$(\Delta x)_{10} \lesssim 1$] and: in the corotating interaction regions of the solar wind [$(\Delta x)_{10} \lesssim 10$], (*cf* Kundt 1983a). In both cases, the spectra peak at energies ~ 50 times below the maximum and have exponential tails. Interstellar shocks can be more than ten times faster; they may or may not extend through much larger distances Δx, depending on the internal geometry of a supernova shell (Kundt 1983b). Even in the most optimistic case, however, Equation (5) does not yield energies anywhere near the maximal ones of cosmic rays, not even anywhere near the maximal energy for galactic containment (= 10^{19} eV).

It has been suggested that the highest-energy cosmic rays are of extragalactic origin, perhaps from the nuclei of active galaxies (AGNs) (Shapiro & Silberberg 1983). Difficulties of this interpretation are (1) the huge overall energy requirement, unless one restricts the claim to energies above 10^{19} eV which are comparatively rare, (2) a monotonic increase and reorientation of the anisotropy (in arrival directions) with energy (Hillas 1983), even though one would sample over volumes which extend far beyond the Virgo cluster, and whose sizes change rapidly with energy, (3) a likely rareness of high-energy ions in AGNs: the total energy radiated by high-energy electrons is already so large that ions cannot store a much larger fraction thereof; the

jets may predominantly consist of electrons and positrons (Kundt & Gopal-Krishna 1980; Kundt 1982a). They may share this property with pulsars whose high-energy wind is thought to be mainly leptonic (Sturrock 1971; Cheng & Ruderman 1980; Kundt & Krotscheck 1980; Kundt 1980).

So far, the above discussion has not thrown suspicion upon any astrophysical source as the booster to the highest observed energies. This absence of clues has led theorists to revive Fermi's idea of accelerating charges in many small steps, by bouncing them off approaching walls (Axford, Leer & Skadron 1977; Bell 1978; Blandford & Ostriker 1978; Krymsky 1977; Drury 1983). Such walls, or shocks, are believed to exist in the form of supernova shells, strong stellar winds and the like. The idea of tapping strong shocks—often called '*in-situ*-acceleration'—has become so widespread that researchers speak of little synchrotrons drifting down the jets of extragalactic radio sources', which are conceived capable of upgrading the bulk energy of beam motion into extremely relativistic electrons *(cf.* Rees 1982). In this way, the kinetic energy of marginally relativistic or even non-relativistic protons is assumed to be converted into a power-law distribution of electrons, with Lorentz factors exceeding the value of 10^6, with the electron power distributed almost evenly up to the high-energy cutoff, at efficiencies reaching 30 per cent, and possibly even in an anisotropic fashion such that we can only see the approaching jet. The existence of similar mechanisms has been claimed for laboratory plasmas.

If such were true, we should wonder why engineers build intricate machines to convert 1 MW of electric power into $\lesssim$ 1 kW of electrons with Lorentz factors $\gamma \lesssim 10^4$ where on thermodynamic grounds we expect a near-l00-per-cent efficiency. All they should do is blow a plasma beam into plasma in order to transform the kinetic energy of ordered proton motion into a power-law-distribution of relativistic electrons, with part of the electron power stored at several hundred times the proton streaming energy.

2. Energy estimates

Let us look in somewhat more detail at the constraints imposed upon a strong shock in the head of an extragalactic radio source. We can estimate the injected power and spectrum $\dot{N}_{\mathrm{e,E}}\mathrm{d}E$ of the relativistic electrons from their synchrotron radiation. This power is often $\gtrsim$ 1 per cent of the total radiation from the active nucleus, *i.e.* corresponds to an efficiency $\eta \gtrsim 1$ per cent of the central engine which generates the beams (Kundt & Gopal-Krishna 1980). Clearly, there is not much room in the overall power budget for energetic protons or ions: their power $\dot{N}_{\mathrm{p}}E_{\mathrm{p}}$ should be at most comparable. *If* $c\beta_{\mp}$ is their bulk velocity $\left\{\begin{matrix}\text{before}\\ \text{after}\end{matrix}\right\}$ the shock, conservation of kinetic energy implies for an incoming 1-temperature hydrogen plasma assumed to segregate (completely) into thermal protons and relativistic electrons:

$$\dot{N}_{\mathrm{p}}[m_{\mathrm{p}}c^2(\gamma_- - 1) + 3kT_-] \simeq \dot{N}_{\mathrm{p}}[m_{\mathrm{p}}c^2(\gamma_+ - 1) + 3kT_+/2] + \int E\dot{N}_{\mathrm{e},E}\mathrm{d}E. \qquad (6)$$

Entropy problems are least for an incoming cold beam. In this case, *i.e.* for $3kT_- \ll m_{\mathrm{p}}c^2(\gamma_- - 1)$, and in the absence of positrons we infer

$$\dot{N}m_{\mathrm{p}}c^2(\gamma_- - \gamma_+ - 3kT_+/2m_{\mathrm{p}}c^2) \simeq \int E\dot{N}_E\mathrm{d}E, \qquad (7)$$

With

$$\dot{N} = \dot{N}_{\rm p} = \dot{N}_{\rm e} = \int_{E_{\rm min}}^{E_{\rm max}} \dot{N}_E \, {\rm d}E.$$

Observations give $\dot{N}_{\rm E} \sim E^{-g}$ for $E_{\rm min} \leqslant E \leqslant E_{\rm max}$ with g = 2.2 ± 0.2, $E_{\rm min} \simeq 10^2$ Me V, $E_{\rm max} \gtrsim 10^2$ GeV, so that the eadic* $E^2\dot{N}_E$ peaks at the lower cutoff energy, and

$$\frac{\int E \dot{N}_E \, {\rm d}E}{\dot{N}} = E_{\rm min} \frac{(g-1)}{(g-2)} \frac{[1-(E_{\rm min}/E_{\rm max})^{g-2}]}{[1-(E_{\rm min}/E_{\rm max})^{g-1}]} \simeq \frac{E_{\rm min}}{g-2}, \tag{8}$$

whence from Equation (7)

$$E_{\rm min} \simeq (g-2) m_{\rm p} c^2 (\gamma_- - 1) \left[1 - \frac{\gamma_+ - 1 + 3kT_+/2m_{\rm p}c^2}{\gamma_- - 1} \right]. \tag{9}$$

In particular, for a non-relativistic incoming bulk velocity $c\beta_-$ this expression implies the inequality

$$E_{\rm min} \lesssim (g/2-1) m_{\rm p} c^2 \beta_-^2 \simeq 10^2 \beta_-^2 \ {\rm MeV}. \tag{10}$$

Comparison with the 'observed' $E_{\rm min}$ shows that either $\beta_- \simeq 1$ must hold, in mild conflict with the assumption of anon-relativistic incoming flow, or else there must be a large number of low-energy electrons in addition to the high-energy power-law. In both cases, most of the electrons would have lower energies than the (shocked) protons, which can cause problems with fundamental constraints by nonequilibrium thermodynamics, in particular with the second law.

3. Entropy estimates

In order to see this, remember that the μ-space entropy S of a homogeneous system of N classical particles reads

$$S/Nk = 1 - \int f \ln f \, {\rm d}^3 p / nh^3, \tag{11}$$

where the phase space density $f = f(\mathbf{p})$ is normalized according to $\int f {\rm d}^3 p = nh^3$, and n: = N/volume. Call s: = S/Nk the (dimensionless) 'entropy per particle'. A Maxwell distribution f (of temperature T) has

$$s_{\rm th} = 5/2 - \ln(n\lambda^3) \quad \text{with} \quad \lambda := h/(2\pi mkT)^{1/2}, \tag{12}$$

where for relativistic temperatures, m has to be replaced roughly by $3\,kT/c^2$. A supersonic bi-Maxwellian beam has

$$s_{\rm be} = 5/2 - \ln(n\lambda_\parallel \lambda_\perp^2) \quad \text{with} \quad \lambda_j := h/(2\pi mkT_j)^{1/2}, \tag{13}$$

and an isotropic truncated relativistic power-law distribution with $f \sim p^{-g-2}$ for $P_{\rm min} \leqslant p \leqslant p_{\rm max}$, $E = pc$, $g > 1$:

$$s_{\rm po} \simeq (2g+1)/(g-1) - \ln[nl^3(g-1)/4\pi] \quad \text{with} \quad l := hc/E_{\rm min}. \tag{14}$$

The particle density n in the beam of an extragalactic radio source is of order $n = L/A\beta cE \simeq 10^{-7}$ cm$^{-3}/\beta$ where L = power, A = beam cross section and E = typical particle energy; $\lambda = 10^{-13}$ cm$/T_{12.5}^{1/2}$ for protons,

* power per energy *e*-folding interval (*eade*).

$l = 10^{-12}$ cm (10^2 Me V/E). We thus get:

$$\begin{aligned} s^{\mathrm{p}}_{\mathrm{th}} &\simeq \ln\left(10^{46} T_{12.5}{}^{3/2}\right) + 5/2 = 108, \\ s^{\mathrm{p}}_{\mathrm{be}} &\simeq \ln\left(10^{43} T_{10.5}{}^{3/2}\right) + 5/2 = 102, \\ s_{\mathrm{po}} &\simeq \ln\left(10^{42} E_{-4.5}{}^{3}\right) + 9/2 = 101, \end{aligned} \tag{15}$$

for the respective entropies per particle (an upper 'p' standing for 'proton', $T_{\|} T_{\perp}^2 =: T^3$, $\beta \simeq 1$). These numbers depend somewhat on one's preferred values for the densities, temperatures and velocities before and after shocking, but satisfy the strict inequalities $s_{\mathrm{be}}, s_{\mathrm{po}} < s_{\mathrm{th}}$. The comparatively low entropy of a beam is due to its low temperature, and that of a soft ($g > 2$) power-law distribution (of comparable total power, Equations 9, 10) is due to the preponderance of cold particles ($E \simeq E_{\min}$).

For the joint system of protons and electrons, the average entropy per particle $s_{\mp} \left\{ \begin{matrix} \text{before} \\ \text{after} \end{matrix} \right\}$ shocking depends on the relative numbers ν of thermalized (ν_{th}), power-law (ν_{po}) and 'cold' (ν_{co}) particles. Somewhat more generally than above, I assume a mixture of incoming thermal protons and electrons ($T = T_-$) plus relativistic power-law electrons, and an outgoing shocked composition of thermal protons ($T = T_+$) plus relativistic power-law electrons and protons plus cold electrons (demanded by Equation 10), and obtain

$$\begin{aligned} 2s_- &= s^{\mathrm{p}}_{\mathrm{be}} + \nu^{\mathrm{e}}_{\mathrm{th},-} s^{\mathrm{e}}_{\mathrm{th}} + (1 - \nu^{\mathrm{e}}_{\mathrm{th},-}) s^{\mathrm{e}}_{\mathrm{po}} \\ 2s_+ &= \nu_{\mathrm{th}} s^{\mathrm{p}}_{\mathrm{th}} + (1 - \nu_{\mathrm{th}}) s^{\mathrm{p}}_{\mathrm{po}} + \nu^{\mathrm{e}}_{\mathrm{po}} s^{\mathrm{e}}_{\mathrm{po}} + (1 - \nu^{\mathrm{e}}_{\mathrm{po}}) s^{\mathrm{e}}_{\mathrm{co}} \end{aligned} \tag{16}$$

with $\nu^{\mathrm{e}}_{\mathrm{po}} := \nu_{\mathrm{po}}$ for electrons, $0 < \nu_j < 1$, $s_{\mathrm{co}} \lesssim s_{\mathrm{po}}$. Proponents of cosmic-ray production via shocks want $\nu^{\mathrm{p}}_{\mathrm{po}} \simeq \nu^{\mathrm{e}}_{\mathrm{po}}$. The power budget in extragalactic radio sources wants the energy in protons to be small. The second law of thermodynamics demands $s_- < s_+$ and is not automatically satisfied: it needs a sufficient number of thermalized protons (ν_{th} not too small) and/or a sufficiently low temperature T_- of the incoming beam:

$$\nu_{\mathrm{th}} (s^{\mathrm{p}}_{\mathrm{th}} - s^{\mathrm{p}}_{\mathrm{be}}) > (1 - \nu_{\mathrm{th}})(s^{\mathrm{p}}_{\mathrm{be}} - s^{\mathrm{p}}_{\mathrm{po}}) + \nu^{\mathrm{e}}_{\mathrm{th},-} (s^{\mathrm{e}}_{\mathrm{th}} - s^{\mathrm{e}}_{\mathrm{po}}) + (1 - \nu^{\mathrm{e}}_{\mathrm{po}})(s^{\mathrm{e}}_{\mathrm{po}} - s^{\mathrm{e}}_{\mathrm{co}}). \tag{17}$$

Note that all terms in this inequality are positive. Clearly, any proof of efficient shock acceleration (of electrons) that does not make use of a low enough pre-shock temperature ($s^{\mathrm{p}}_{\mathrm{be}}$ small) would violate the second law of thermodynamics, and hence must be inconclusive.

4. Further constraints

However, a growth of entropy is not the only condition imposed by non-equilibrium thermodynamics: The joint N-particle distribution function $f(\mathrm{x}_i, \mathbf{p}_i, t)$ of a closed system obeys a master equation which describes its evolution in time, as a consequence of which its spectral measures (phase-space integrals)

$$\int_{f > \lambda} [f(t) - \lambda] \prod_i \mathrm{d}^3 x_i \, \mathrm{d}^3 p_i \tag{18}$$

cannot grow with time for any positive λ (Schlögl 1980, Equation 4.3.13). This condition goes beyond the entropy theorem, and limits severely the (post-shock) number of 'cold' particles.

I therefore question results obtained on the shock-acceleration to extremely relativistic energies whenever significant efficiencies ($\eta \gtrsim 1$ per cent) are claimed, *i.e.* whenever the boosted charges leave the test particle regime, or, in other words, whenever the (joint) inertia of the boosted charges grows comparable to that of the scatterers.

5. Neutron stars and AGNs

It is my conviction that nature has found different solutions to the relativistic acceleration problem (*cf.* Fig. 1). We know of the existence of fast-spinning magnetized neutron stars. Their strong outgoing waves involve potential drops which are a factor of order 10^{10} above those of interstellar shocks (*cf* Equations 4 and 5) and which may well be responsible for most of the high-energy cosmic-ray electrons and positrons. Quite likely, the central engines in the active galactic nuclei are likewise magnetized rotators and act like giant pulsars (Morrison 1969; Ozernoy & Usov 1977; Kundt 1979, 1982a). If they can generate the observed extremely relativistic $e^{\pm}$ -flows on the innermost scale of $10^{15\pm1}$ cm, there would be no need for any downstream (hot-spot) post-acceleration.

We also know of the existence of magnetized neutron stars in binary systems where the weak but nevertheless heavy wind of a companion star can quench the pulsar mechanism, impinge upon the corotating magnetosphere and be flung out like sparks from a grindstone. When confining the magnetosphere deep inside the speed-of-light cylinder, the falling matter can (in principle) generate voltages of order

$$E_{\rm grind} \leqslant 2e\,Br = 2e\,B_{\rm s}R^3/r^2 = 10^{20}\,(B_{13}/r_7^2)\,{\rm eV}, \tag{19}$$

which reach high enough to explain the observed cosmic-ray spectrum. This potential cosmic-ray booster has been recently discussed in (Kundt 1983a). Note that whereas shock acceleration models derive a power-law distribution in energy from a balance between an exponentially increasing particle energy and an exponentially decreasing survival probability with time, the magnetic grindstone can explain a power-law output in energy $\dot{N}_{\rm E}\,{\rm d}E \sim E^{-2.7\pm0.3}\,{\rm d}E$ as a number ratio of particle orbits traversing different electric potentials: For a distance-dependence $B \sim r^{-n}$, and $\Delta x \sim r$, β = const, Equation (3) implies $E \sim B\,\Delta x \sim r^{1-n}$; hence

$$E\dot{N}_E \sim r^3 \sim E^{-3/(n-1)}, \tag{20}$$

which agrees with the primary cosmic-ray spectrum for $n = 2.8 \pm 0.3$, *i.e.,* for an almost-dipole behaviour of the magnetic field.

A final word concerns SS 433. Its ultraviolet output is difficult to estimate, but both the observed fluxes and the Eddington limit suggest that the central engine radiates less than 10^{39} erg s^{-1}. At an efficiency η of order 1 per cent, this central engine should not be able to produce beams more powerful than some 10^{37} erg s^{-1}. This consideration stands against the hydrogen beam model, in favour of extremely relativistic $e^{\pm}$ –beams (Kundt 1981b).I even maintain that details of the radio spiral of Hjellming &Johnston (1981) and X-ray map of Seward *et al* (1980) can be understood better in terms of the 'soft beam' model (hair-drier, Kundt 1982b) than in terms of the 'hard beam' model (lawn sprinkler). If so, SS 433 would follow a pattern similar to that of the AGNs.

Acknowledgements

I thank Axel Jessner and Hajo Leschke for discussions, Reinhard Schlickeiser for critical remarks, and an anonymous referee for very constructive criticism.

References

Axford, W. I., Leer, E., Skadron, G. 1977, *Proc. Int. Cosmic Ray Conf.,* **11,** 132.
Bell, A. R. 1978, *M on. Not. R. astr. Soc,* **182,** 147.
Blandford, R. D., Ostriker, J. P. 1978, *Astrophys. J.,* **221,** L29.
Cheng, A., Ruderman, M. A. 1980, *Astrophys. J.,* **235,** 576.
Drury, L. O'C. 1983, *Rep. Prog. Phys.,* **46,** 973.
Hillas, A. M. 1983, in *Composition and Origin of Cosmic Rays,* Ed. M. M. Shapiro, D. Reidel, Dordrecht, p. 125.
Hjellming, R. M., Johnston, K. J. 1981, *Astrophys. J.*, **246,** L141.
Krymsky, G. F. 1977, *Dokl. Akad. Nauk SSSR*, **234,** 1306.
Kundt, W. 1979, *Astrophys. Space Sci*, **62,** 335.
Kundt, W. 1980, *Ann. N.Y. Acad. Sci*., **336,** 429.
Kundt, W. 1981a, *Astr. Astrophys*., **98,** 207.
Kundt, W. 1981b, *Vistas Astr.*, **25,** 153.
Kundt, W. 1982a, in *IAU Symp.* 97: *Extragalactic Radio Sources*, Eds D. S. Heeschen & C. M. Wade, D. Reidel, Dordrecht p.265.
Kundt, W. 1982b, *Mitt. astr. Ges.*, **57,** 65.
Kundt, W. 1983a, *Astrophys. Space Sci*., **90,** 59.
Kundt, W. 1983b, *Astr. Astrophys.*, **121,** L15.
Kundt, W., Gopal-Krishna 1980, *Astrophys. Space Sci.*, **75,** 257.
Kundt, W., Krotscheck, E. 1980, *Astr. Astrophys.*, **83,** 1.
Morrison, P. 1969, *Astrophys. J.*, **157,** L73.
Ozernoy, L. M., Usov, V. V. 1977 *Astr. Astrophys.*, **56,** 163.
Rees, M. J. 1982, in *IAU Symp.* 97: *Extragalactic Radio Sources*, Eds. D. S. Heeschen & C. M. Wade, D. Reidel, Dordrecht p.211.
Schlögl, F. 1980, *Phys. Rep.*, **62,** 267.
Schreier, E. J., Gorenstein, P., Feigelson, E. D. 1982, *Astrophys. J.*, **261,** 42.
Seward, F., Grindlay, J., Seaquist, E., Gilmore, W. 1980, *Nature*, **287,** 806.
Shapiro, M. M., Silberberg, R. 1983, *Astrophys. J.*, 265, 570.
Sturrock, P. A. 1971, *Astrophys. J.*, **164,** 529.

J. Astrophys. Astr. (1984) **5,** 285–306

Stability of a Finite Disc under the Influence of a Spherical Halo

Ashok Ambastha *Udaipur Solar Observatory, 11 Vidya Marg, Udaipur 313001*

Received 1984 March 2: accepted 1984 May 1

Abstract. We have studied the stability of finite gaseous discs, against large-scale perturbations, under the influence of spherical, massive haloes. A surface-density distribution consistent with the observed spiral-tracer profiles in disc galaxies is considered for the disc. We find that growing eigenmodes with both 'trailing' and 'leading' spirals exist in 'cold' discs for a wide range of values of the halo mass and its radius. The amplification rates of the unstable modes reduce as the ratio of the mass of the halo to the mass of the disc is increased. A uniform halo is not very effective towards stabilizing the disc against these modes. The results from the present study are considered *vis-a-vis* previous studies on the global modes of self-gravitating discs.

Key words: galaxies, spiral structure—galaxies, haloes—density waves

1. Introduction

The explanation of the origin and maintenance of various structural components of flat galaxies, *viz.*, spiral arms, bars and rings has been an important problem in astrophysics. Based on the idea that these structures originate in waves and instabilities in self-gravitating systems, Lin and his co-workers (Lin & Shu 1964; Lin, Yuan & Shu 1969) showed by a 'local' analysis that neutral, spiral density waves exist in infinitely extended rotating galactic discs. However, since the gravitational force is a long-range one and there is no shielding effect in galactic discs, a local theory is severely limited when applied to such systems, particularly when one considers wavelengths comparable to the characteristic dimensions of the system. Consequently, a 'global' analysis of waves in self-gravitating discs is called for (lye 1978; Aoki, Noguchi & lye 1979; Ambastha & Varma 1983).

The asymptotic theory has been extended further by Lau, Lin & Mark (1976) and Bertin & Mark (1978) who obtained discrete unstable spiral modes from their numerical studies. Pannatoni & Lau (1979) have relaxed the restrictions of the asymptotic approach. Bertin (1980) and Toomre (1981) have reviewed the subject very extensively.

However, most numerical calculations of large-scale spiral modes (Kalnajs 1972; Bardeen 1975) and N-body computer simulations (Ostriker & Peebles 1973; Hohl 1970) often yielded spiral disturbances which are 'explosively' unstable. These strong instabilities tend to prohibit truely long-lived spiral structures in the galactic model under investigation. Even if the condition of local stability with a minimum dispersion (Toomre 1964) is obeyed in the model disc, it is found unstable against global

perturbations and a complete stability could be attained only by unacceptably large thermal dispersions (Ambastha & Varma 1983). It is observed that these explosive instabilities could alternatively, be suppressed to some extent by having massive central bulge component in the system (Ambastha & Varma 1982).

The result emanating from the computer experiments on the evolution of disc models suggest that a massive 'halo' surrounding the disc could inhibit the formation of bar instabilities (Berman, Brownrigg & Hockney 1978; Efstathiou, Lake & Negroponte 1982). On the other hand, recent observations on the rotational profiles of the Milky Way and other external galaxies give evidence of nearly flat rotation curves at large galactocentric distances (Rubin, Ford, Thonard 1978; Bosma 1978; Krumm & Salpeter 1979). This would indicate the existence of an extensive halo enveloping the optically visible disc (Bok 1981). Theoretical possibilities regarding this massive constituent are varied—to mention a few among many, black holes (Truran & Cameron 1971), comets (Tinsley & Cameron 1974), a population of jovian planets (Salpeter 1977), frozen hydrogen snowballs (Reddish 1968) and faint main-sequence stars (Ostriker, Peebles & Yahil 1974). With the exception of the last possibility, all of these objects are essentially undetectable by any current or foreseeable observational technique. The halo could be ten to twenty times as massive as the disc which would have a significant role to play in the structure and stability of galactic discs.

Using the approach of the global linear mode analysis, we wish here to study the consequences of a fixed spherical halo on the eigenmodes of oscillation of a flat gaseous disc. Thus, rather than following an overall evolution of the system, as in a computer experiment, our interest here is to investigate the effect of various parameters characterizing the halo, on the normal modes of oscillation of the gaseous disc.

The spiral structure of disc galaxies is most significantly traced by its constituent components with low thermal dispersion, *e.g.*, dust, gas (particularly, HII regions), O and B populations.* These spiral-tracers are observed to be distributed in a ring within 2 kpc $< r <$ 16 kpc with a peak in density at $r \sim$ 4–6 kpc in our galaxy (Gordon & Burton 1976; Stecker 1976; Hart & Pedlar 1976; Kodaira 1974). However, minor quantities of gas exist even closer to the galactic nuclei. Keeping such a distribution of spiral-tracers in mind, we adopt here a gaseous disc with a surface-density distribution which vanishes both at the centre of the disc and at its boundary. Such a model is significantly different from that of Takahara (1978).

We shall consider a static halo, as the objects constituting galactic halos usually have almost no tendency to concentrate towards the flat disc and have very large thermal dispersions. The earlier studies have demonstrated that the effect of small perturbations on such 'hot' components is indeed very small (Kato 1974).

2. Basic formulation

Let us consider a flat self-gravitating disc, rotating about an axis perpendicular to its plane and passing through its centre. The fluid-dynamical equations, governing the

* One may note here that a significant number of galaxies possess very fragmentary spiral patterns with no grand two-armed design, but rather a system of multiple and branched arms which can only be individually followed over short distances. Seiden & Gerola (1979) have deveoped computer models using the concept of stochastic, supernova-induced star formation which simulate such structures and show that it may have very little at all to do with a global density-wave pattern. Such a model may play some part in influencing the morphology of spiral patterns even when a large-scale density wave is operative.

dynamics of a pressureless gaseous disc under the static gravitational field of a spherically symmetric halo, are

$$\frac{\partial \sigma}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}[r\sigma u] + \frac{1}{r}\frac{\partial}{\partial \theta}[\sigma v] = 0, \tag{2.1}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial r} + \frac{v}{r}\frac{\partial u}{\partial \theta} - \frac{v^2}{r} = \frac{\partial \psi_g}{\partial r} + \frac{\partial \psi_h}{\partial r}, \tag{2.2}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial r} + \frac{v}{r}\frac{\partial v}{\partial \theta} + \frac{uv}{r} = \frac{1}{r}\frac{\partial \psi_g}{\partial \theta}, \tag{2.3}$$

where σ, u, v and y ψ_g, ψ_h represent, at an instant of time t, the surface-density, radial and azimuthal velocities and the gravitational potentials at a point (r, θ) produced by the distribution of matter in the disc and by the halo respectively, on the plane of the disc.

The surface-density and potential of the self-gravitating gaseous disc are related by

$$\nabla^2 \psi_g = -4\pi G\delta(z)\sigma(r, \theta, t). \tag{2.4}$$

3. Equilibrium of the system

We now consider the equilibrium, *i.e.*, the time-independent, axisymmetric state of the disc under the external influence of the spherical halo. The azimuthal velocity, $V_g(r)$, of the gaseous component which constitutes the disc is obtained from the radial component of the momentum conservation equation. Here, we assume the absence of any radial flows in the disc (*i.e.*, $U_g(r) = 0$). Then one obtains from Equation (2.2)

$$-\frac{V_g^2(r)}{r} = \frac{d\Psi_0(r)}{dr} \tag{3.1}$$

where $\Psi_O(r)$ is the net gravitational potential at a point r in the disc, and is given by

$$\Psi(r) = \Psi_g(r, z = 0) + \Psi_h(r)$$

since the gravitational potential is additive. Here Ψ_g is the self-consistent gravitational potential of the gaseous disc and Ψ_h is the potential exerted due to the mass distribution in the spherical halo surrounding the disc.

In what follows we would consider a gaseous disc with surface-density distribution σ_g of the form

$$\begin{aligned} \sigma_g(r) &= \sigma_0 J_2(\lambda_0 r) \quad ; \quad r \leqslant R_d \\ &= 0 \qquad\qquad\quad ; \quad r > R_d \end{aligned} \tag{3.2}$$

where R_d defines the radius of the disc and σ_0 is a constant. Also λ_0 is the first zero of the Bessel function of second order, *i.e.*, the first root of the transcendental equation $J_2(\lambda R_d) = 0$. The surface density represented by Equation (3.2) vanishes both at the disc-centre, *i.e.* $r = 0$, and at the disc-boundary, *i.e.*, at $r = R_d$ (see Fig. 1). With the adoption of the above form, the gradient of the surface density is free from any discontinuity as $r \to 0$ since $\sigma_g'(r) \sim$ (0) (r); however, at $r = R_d$ there appears a dis-

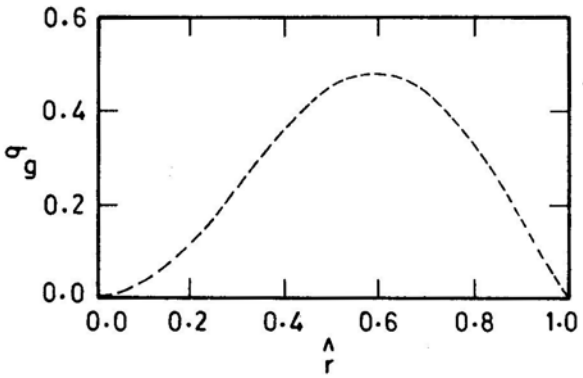


Figure 1. Surface-density profile of the gaseous disc, normalized to (δ_g/A).

continuity since $\sigma'_g \neq 0$. Most finite radius disc models suffer from such a discontinuity at the edge of the disc* (Hunter 1963).

Now, the gravitational potential, corresponding to a general density distribution of the form

$$\begin{aligned} \sigma(r) &= \sum_{j=0}^{\infty} a_j^{(m)} J_m(\lambda_j^{(m)} r) \quad ; \quad r \leqslant R \\ &= 0 \qquad\qquad\qquad\qquad\quad ; \quad r > R \end{aligned} \tag{3.3a}$$

is given by (*cf.* Yabushita 1969):

$$\begin{aligned} \Psi(r) &= 2\pi G \sum_{j=0}^{\infty} a_j^{(m)} \left[\frac{J_m(\lambda_j^{(m)} r)}{\lambda_j^{(m)}} + \frac{4}{\pi^2 Y_m(\lambda_j^{(m)} R)} \int_0^{\infty} \frac{\mathrm{d}k' I_m(k'r) K_m(k'R)}{(\lambda_j^{(m)})^2 + k'^2} \right]; \quad r \leqslant R \\ &= 2\pi G \sum_{j=0}^{\infty} a_j^{(m)} \left[\frac{4}{\pi^2 Y_m(\lambda_j^{(m)} R)} \int_0^{\infty} \frac{\mathrm{d}k' I_m(k'R) K_m(k'r)}{(\lambda_j^{(m)})^2 + k'^2} \right]; \quad r > R. \end{aligned} \tag{3.3b}$$

Hence, for the density distribution given by Equation (3.2), the potential at a point r on the plane of the disc would be

$$\Psi_g(r) = 2\pi G \sigma_0 \left[\frac{J_2(\lambda_0 r)}{\lambda_0} + \frac{4}{\pi^2 Y_2(\lambda_0 R_d)} \int_0^{\infty} \frac{I_2(k'r) K_2(k'R_d)}{\lambda_0^2 + k'^2} \mathrm{d}k' \right]; \quad r \leqslant R_d. \tag{3.4}$$

Having discussed the gaseous component, we now consider a spherically symmetric halo with a volume-density, $\rho_h(r)$, defined by

$$\begin{aligned} \rho_h(r) &= \rho_0 (1 + r^2/R_h^2)^{-1} \quad ; \quad r \leqslant R_h \\ &= 0 \qquad\qquad\qquad\quad ; \quad r > R_h \end{aligned} \tag{3.5}$$

where ρ and R_h are the central density and the core-radius of the halo distribution respectively. The halo matter outside the disc radius, R_d, would not influence the dynamics of the disc since a spherical halo has been assumed. The gravitational

* One may circumvent this problem by defining a new surface-density distribution. $\sigma_{g,\,new}$ constructed as

$$\sigma_{g,new} = (R_d - r)\sigma_g(r)$$

whose derivative vanishes both at $r = 0$ and $r = R_d$ (G. Contopoulos 1984, personal communication).

potential corresponding to the halo distribution given by Equation (3.5) at a point r in the plane of the gaseous disc is

$$\Psi_h(r) = -\frac{4\pi G\rho_0}{2} R_h^2 [\ln(1 + r^2/R_h^2) + \frac{2R_h}{r} \tan^{-1}(r/R_h) - 2], \tag{3.6}$$

which yields a gravitational force

$$\frac{d\Psi_h}{dr} = -\frac{4\pi G\rho_0}{r^2} R_h^3 [r/R_h - \tan^{-1}(r/R_h)]. \tag{3.7}$$

3.1 *Normalization of the Equations*

We normalize the various physical quantities, *e.g.*, the surface-density $\sigma(r)$, the potential $\Psi(r)$, the velocities $u(r)$ and $v(r)$ as follows:

$$\begin{aligned} \sigma(r) &= (M/2\pi R_d^2)\hat{\sigma}(\hat{r}) \\ \Psi(r) &= (\pi GM/R_d)\hat{\Psi}(\hat{r}) \\ \begin{Bmatrix} u(r) \\ v(r) \end{Bmatrix} &= \left(\frac{\pi GM}{R_d}\right)^{1/2} \begin{Bmatrix} \hat{u}(\hat{r}) \\ \hat{v}(\hat{r}) \end{Bmatrix} \end{aligned} \tag{3.8}$$

where $\hat{r} = r/R_d$ and $M = M_g + M_h \equiv M_{total}$.

The mass of the gaseous disc, M_g, is given by

$$M_g = \int_0^{R_d} 2\pi r \sigma_g(r)\, dr, \tag{3.9}$$

from which we obtain

$$\sigma_0 = \frac{M_g}{2\pi R_d^2} \cdot \frac{1}{A}. \tag{3.10}$$

Hence,

$$\hat{\sigma}_g(\hat{r}) = (\delta_g/A) J_2(\hat{\lambda}_0 \hat{r}), \tag{3.11}$$

with $\delta_g = (M_g/M)$ and $A = \int_0^1 \hat{r} J_2(\hat{\lambda}_0 \hat{r}) d\hat{r}$, $\hat{\lambda}_0 = \lambda_0 R_d$.

The gravitational potential of the gaseous component in normalized form, consequently, turns out to be

$$\hat{\Psi}_g(\hat{r}) = (\delta_g/\pi A)\left[\frac{J_2(\hat{\lambda}_0\hat{r})}{\hat{\lambda}_0} + \frac{4}{\pi^2 Y_2(\hat{\lambda}_0)} \int_0^\infty d\hat{k}' I_2(\hat{k}'\hat{r}) \frac{K_2(\hat{k}')}{\hat{\lambda}_0^2 + \hat{k}'^2}\right]. \tag{3.12}$$

Similarly, the halo mass within the disc radius, R_d is

$$\begin{aligned} M_h(R_d) &= \int_0^{R_d} 4\pi r^2 \rho_h(r)\, dr \\ &= 4\pi R_h^3 \rho_0 [R_d/R_h - \tan^{-1}(R_d/R_h)]. \end{aligned} \tag{3.13}$$

Thus, we define the central density, ρ_0, such that

$$\rho_0 = M_h(R_d)[4\pi R_h^3\{R_d/R_h - \tan^{-1}(R_d/R_h)\}]^{-1}. \tag{3.14}$$

However, one would note here that the total halo mass is

$$M_h(R_h) = 4(1-\pi/4)\pi\rho_0 R_h^3.$$

The non-dimensional potential of the halo and its gradient are thus

$$\hat{\Psi}_h(\hat{r}) = -(\delta_h/2\pi\eta)\left[\ln(1+\hat{r}^2/\eta^2) + \frac{2\eta}{\hat{r}}\tan^{-1}(\hat{r}/\eta) - 2\right][\eta^{-1} - \tan^{-1}(\eta^{-1})]^{-1} \tag{3.15}$$

and

$$\hat{\Psi}'_h(\hat{r}) = -(\delta_h/\pi\hat{r}^2\eta)[\hat{r} - \eta\tan^{-1}(\hat{r}\eta^{-1})][\eta^{-1} - \tan^{-1}(\eta^{-1})]^{-1}. \tag{3.16}$$

Here,

$$\delta_h = M_h(R_d)/M \quad \text{and} \quad \eta = (R_h/R_d).$$

In the present investigation, our main interest is to understand the influence of the halo distribution on the stability of the disc; hence, we have assumed for simplification a 'cold' or pressureless disc, *i.e.*, $P_g = 0$. The effect of thermal pressure has been studied earlier (Aoki *et al.* 1979; Ambastha & Varma 1983). By substituting Equations (3.12) and (3.15) in Equation (3.1) one can now evaluate the circular-velocity $V_g(r)$ of the

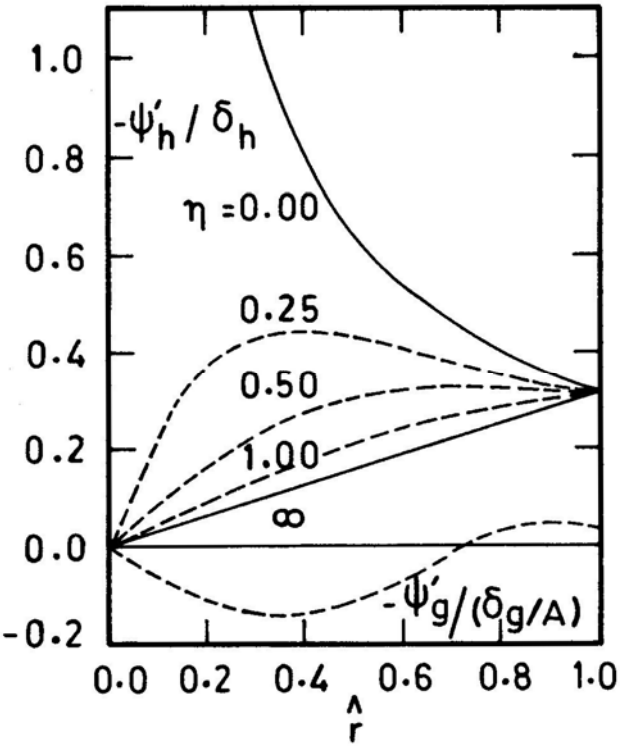


Figure 2. Profile of $-\Psi'_h/\delta_h$ for various values of η. The profile of $-\Psi'/(\delta_g/A)$ is also shown.

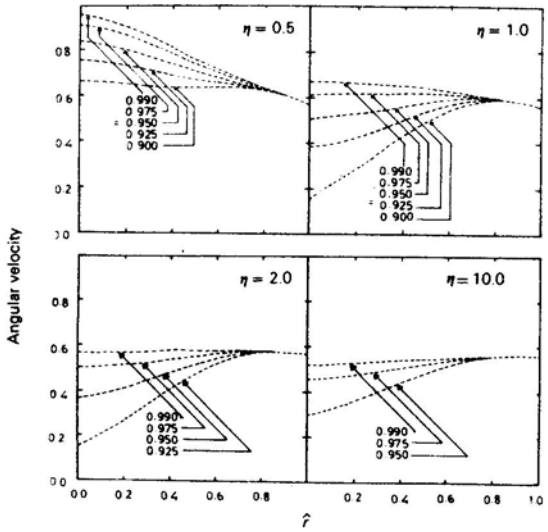


Figure 3. Variation of angular velocities with distance for different values of η and δ_h. Each curve is labelled by the corresponding value of δ_h.

gaseous disc. Note that the gaseous disc given by Equation (3.1) is unphysical in the absence of a contribution from an external halo or some other source, as $-\Psi'_g(r)$ is negative in the range $0 < r < 0.7\ R_d$ and consequently $V_g^2(r) < 0$ in this interval according to Equation (3.1). Fig. 2 shows profiles of $-\Psi'_h/\delta_h$ for various values of η and of $-\Psi'_g$ (δ_g/A). Fig. 3 gives the profiles of the angular velocities for various values of η and $(\delta_h$ ($\equiv M_h/M$). The angular velocity, $\Omega(r)$ increases towards the boundary of the disc in most of the cases shown in Fig. 3. However, for $\eta = 0.5$, $\Omega'(r) \lesssim 0$ over the entire disc.

4. Results and discussion

We wish to investigate the stability and eigenmodes of the axisymmetric, equilibrium disc surrounded by a halo, discussed in Section 3, against small amplitude perturbation of the form

$$\tilde{A}(\hat{r}, \theta, \tau) = \hat{A}(\hat{r}) \exp\left[i(\omega\tau + m\theta)\right], \tag{4.1}$$

where $\tilde{A}$ symbolizes the perturbations in σ, u, v and Ψ. Here $\omega(\equiv \omega + i\omega_i)$ is the frequency and m is the azimuthal wavenumber of the perturbations. The perturbations grow exponentially, damp, or are oscillatory according as $\omega_i < 0$, > 0, or $=0$, respectively. Thus the net quantities now become

$$A(\hat{r}, \theta, \tau) = A_0(\hat{r}) + \varepsilon \tilde{A}(\hat{r}, \theta, \tau), \tag{4.2}$$

where ε is a smallness parameter. We can linearize the set of Equations (2.1)–(2.4) to obtain the hydrodynamic equations governing the perturbations.

The method followed here is essentially on the lines as described in Ambastha & Varma (1983); hence the details in the analysis are avoided. However in what follows, an outline of the method is described for continuity.

The radial parts of the perturbation $\hat{A}$ ($\hat{r}$) in the linearized equations for the perturbed quantities, $\hat{A}(\hat{r}, \theta, \tau)$ can be expressed as infinite expansions in terms of suitable Bessel functions. Using the orthogonality of the Bessel functions, the linear equations are then integrated over the disc, *i.e.*, in the interval $0 \leqslant \hat{r} \leqslant 1$. As a result, an infinite set of algebraic equations is derived which could more conveniently be expressed in the form of an eigenvalue problem

$$MZ = \omega z \tag{4.3}$$

where M is a $3\infty \times 3\infty$ matrix, nonsymmetric in general and Z is a column matrix constituted by the basis vectors of $\hat{\sigma}(\hat{r})$, $\hat{u}(\hat{r})$ and $\hat{v}(\hat{r})$. The eigenvalue problem expressed by Equation (4.3) cannot be solved analytically in general, and we use a numerical method involving elementary similarity transformation of a suitably truncated matrix M to evaluate the eigenvalues ω, and the corresponding associated eigenvectors Z. In what follows, we have considered only bisymmetric, *i.e.*, $m = 2$ perturbations. The results are, however, qualitatively similar for other perturbations as found previously (Ambastha & Varma 1983).

The unstable eigenmodes are listed in Table 1 for $m = 2$ (bisymmetric) perturbations. The results have been obtained for $\eta = 0.5$, 1.0, 2.0, 5.0, 10.0 and 50.0 and $\delta_h = 0.9$, 0.925, 0.95, 0.975 and 0.99, the net mass of the composite system being normalized to unity. There exist also a number of oscillatory ($\omega_i = 0$), and damped ($\omega_i > 0$) modes (complex-conjugate to the unstable modes) in each cases; but we do not discuss them here.

A number of unstable modes are allowed when $\eta = 0.5$, $\delta_h = 0.9$, ranging from those with growth-rates comparable to the pattern-frequencies ($\equiv -\omega_r/m$) to those with $\omega_i \ll \omega_r$. As the mass ratio δ_h / δ_g is increased, a few unstable modes are suppressed and eventually none exists when $\delta_h > 0.95$. It should be noted that $\Omega'(r) < 0$ over almost the entire disc when $\eta = 0.5$.

As the halo increases in radius, the stabilizing effect on the unstable modes is found to diminish (see the column under $\eta = 1.0$ for various values of δ_h in Table 1). For all values of δ_h, in the case of $\eta = 1.0$, $\Omega'(r) \gtrsim 0$ over almost the enitre disc.

The angular velocity$\Omega(r)$ of the disc for $\eta = 2.0$, $\delta_h = 0.9$; $\eta = 5.0$, $\delta_h = 0.9$ and 0.925 *etc.* becomes imaginary as $r \to 0$ and hence these discs are physically forbidden. We do not consider such a disc. The modes are found only moderately affected by increasing η beyond 5.0.

4.1 *The Patterns of Oscillation*

One can construct eigenpatterns associated with any of the perturbations, *i.e.*, $\tilde{\sigma}$, $\tilde{u}$ and $\tilde{v}$ from the eigenfunctions corresponding to each eigenfrequency discussed earlier. Figs 4–7 show the eigenpatterns constructed for $\tilde{\sigma}(\hat{r}, \theta, \tau = 0)$. The digits along the patterns denote the amplitudes of the perturbation at those points on the disc, normalized to 'F' in the hexadecimal system. Only positive values of $\tilde{\sigma}$ (*i.e.*, density enhancement) have been printed. The sense of rotation is anticlockwise in all figures, as indicated by an arrow in the first frame of Fig. 4. The results are discussed here in the light of earlier studies on the global modes.

It has been found previously that the unstable eigenmodes are 'leading' in nature in 'cold' discs, *i.e.*, the patterns open in the direction of rotation. These leading patterns gradually turn to 'trailing' patterns when sufficient nonzero pressures are introduced (Ambastha & Varma 1983).

Table 1. Unstable eigenmodes ($m = 2$) for various values of η and δ_h.

η \ δ_h	0.900		0.925		0.950		0.975		0.990	
	$-\omega_r$	$-\omega_i$	$-\omega_r$	$-\omega_i$	$-\omega_r$	$-\omega_i$	$-\omega_r$	$-\omega_i$	$-\omega_r$	$-\omega_i$
0.5	1.2787	0.8166								
	1.2861	0.6252	1.3042	0.2840	No unstable modes		No unstable modes		No unstable modes	
	1.5220	0.0889								
	1.2642	0.0599	1.5866	0.0995						
	1.1492	0.0167								
	1.4327	0.0065								
1.0	0.9935	0.9480								
	0.9801	0.8213	1.0924	0.5794						
	0.8925	0.5363	1.1099	0.3873			No unstable modes		No unstable modes	
	0.7839	0.3556	1.0719	0.0967	1.1868	0.0045				
	0.9764	0.3074	1.0816	0.0395						
	1.0639	0.0610	1.1346	0.0060						
	0.7333	0.0364								
2.0	Physically forbidden		0.9040	0.6950						
			0.8943	0.5682	1.0441	0.0830			No unstable modes	
			0.7338	0.2824	1.3611	0.0464	1.1314	0.0014		
			0.8371	0.2803						
			0.9299	0.0759						
			0.5511	0.0199						
			1.0838	0.0065						
5.0	Physically forbidden		Physically forbidden		0.9579	0.1903	No unstable modes		No unstable modes	
					1.1979	0.1055				
10.0	Physically forbidden		Physically forbidden		0.9433	0.2099	No unstable modes		No unstable modes	
					1.1739	0.1129				
50.0	Physically forbidden		Physically forbidden		0.9386	0.2163	No unstable modes		No unstable modes	
					1.1661	0.1151				

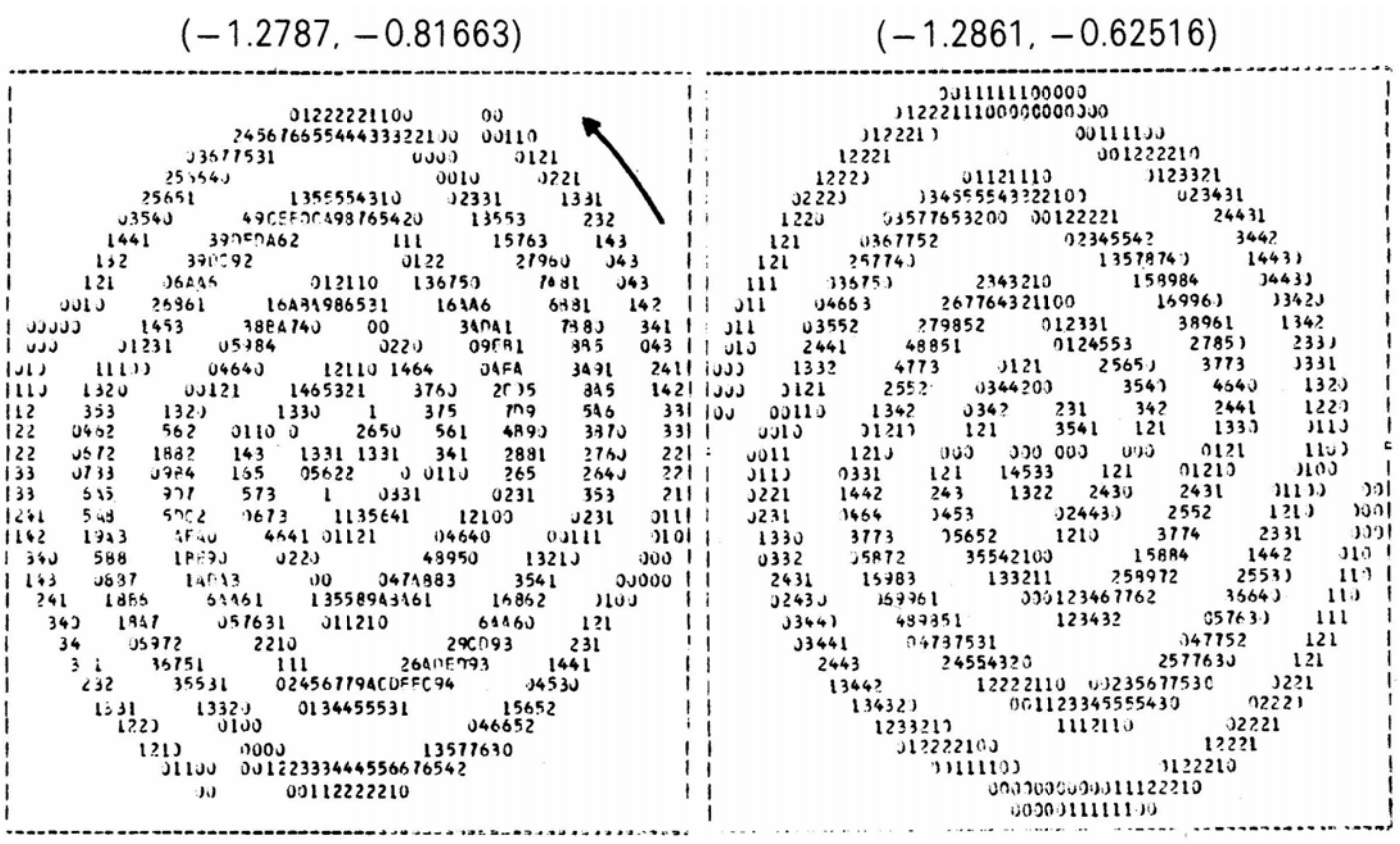


Figure 4. Eigenpatterns of oscillation in $\tilde{\sigma}(\hat{r}, \theta, \tau = 0)$ for $\eta = 0.5$, $\delta_h = 0.9$. The sense of rotation is anticlockwise. The values (ω_r, ω_i) appear on top of each frame.

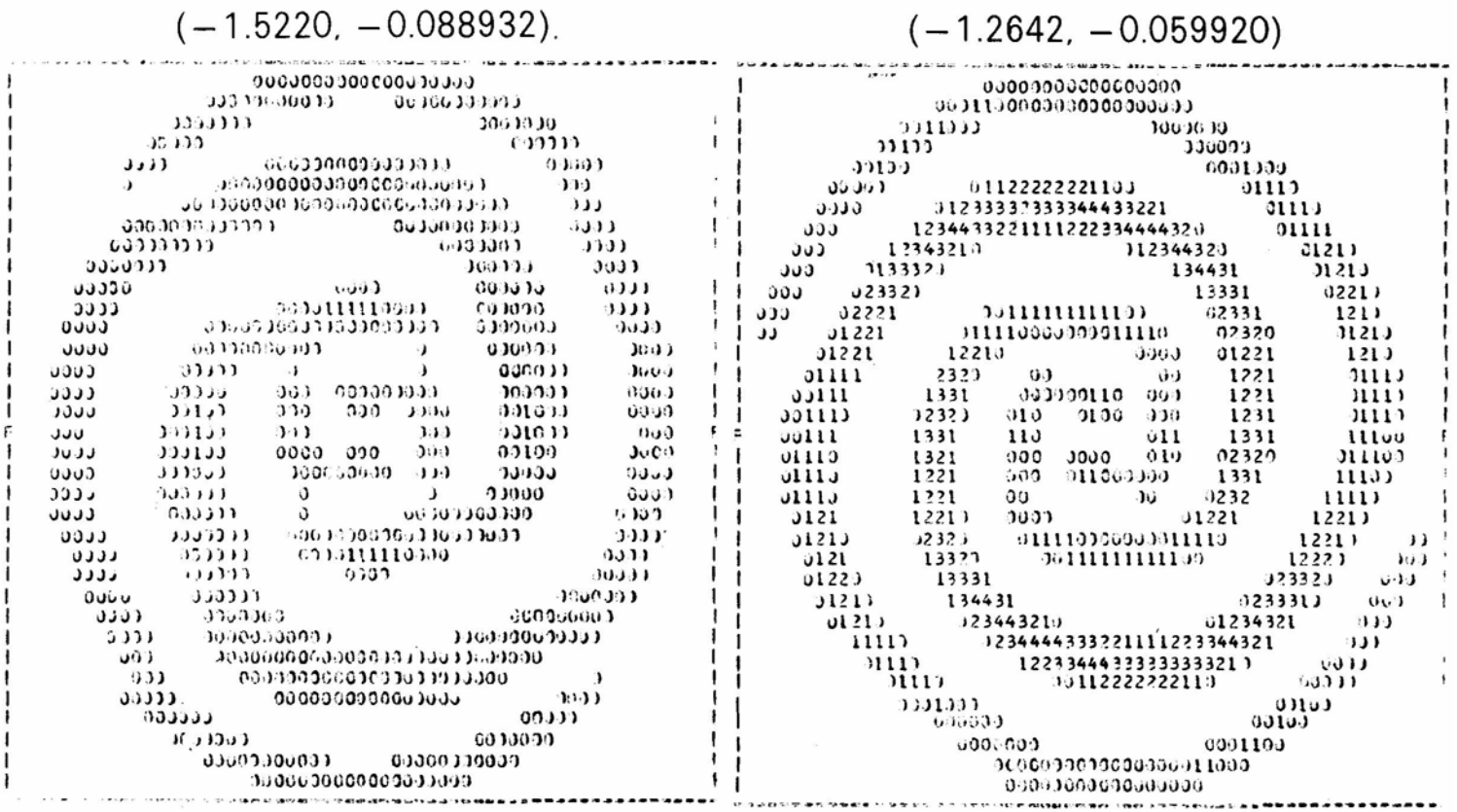


Figure 4. Continued.

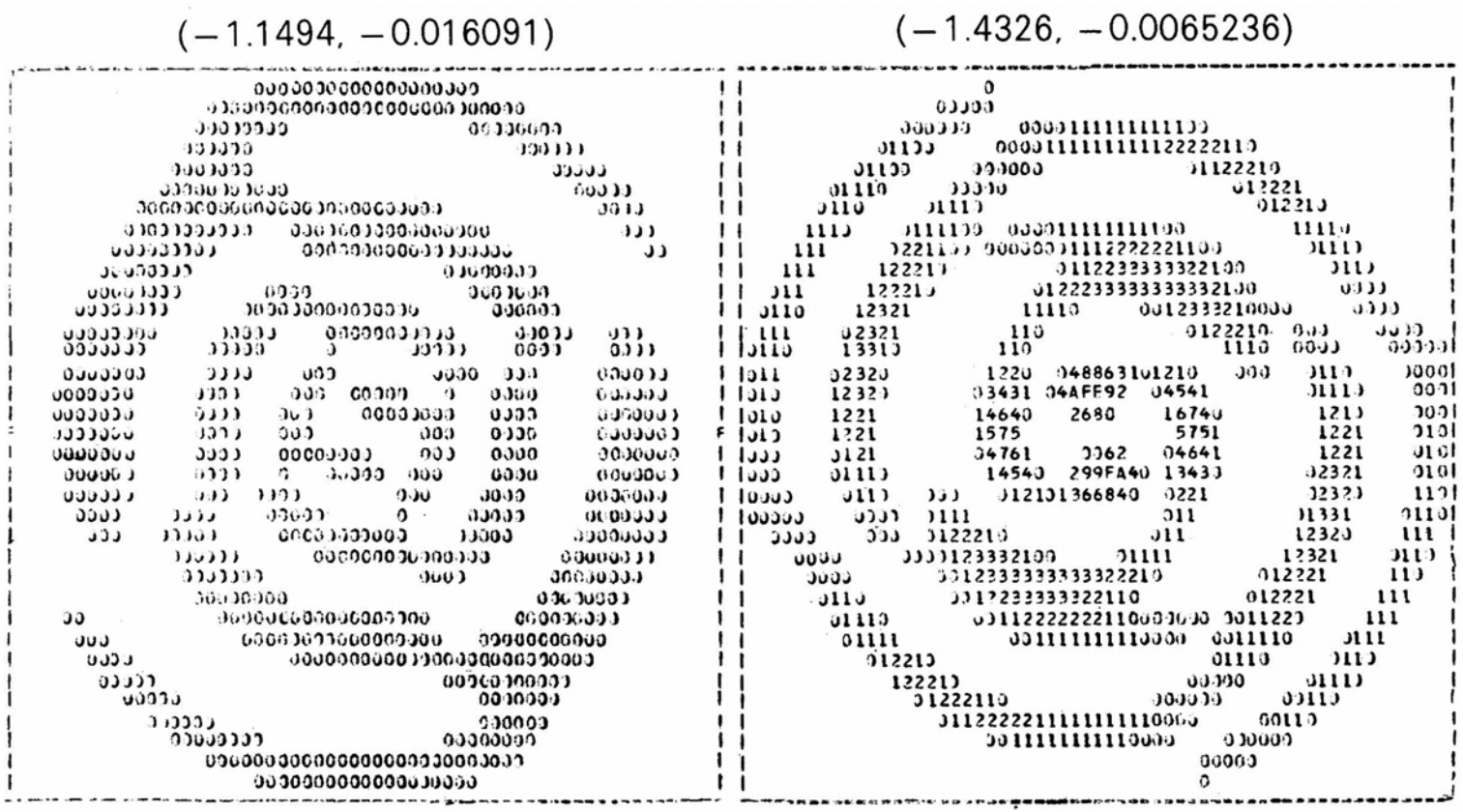


Figure 4. Continued.

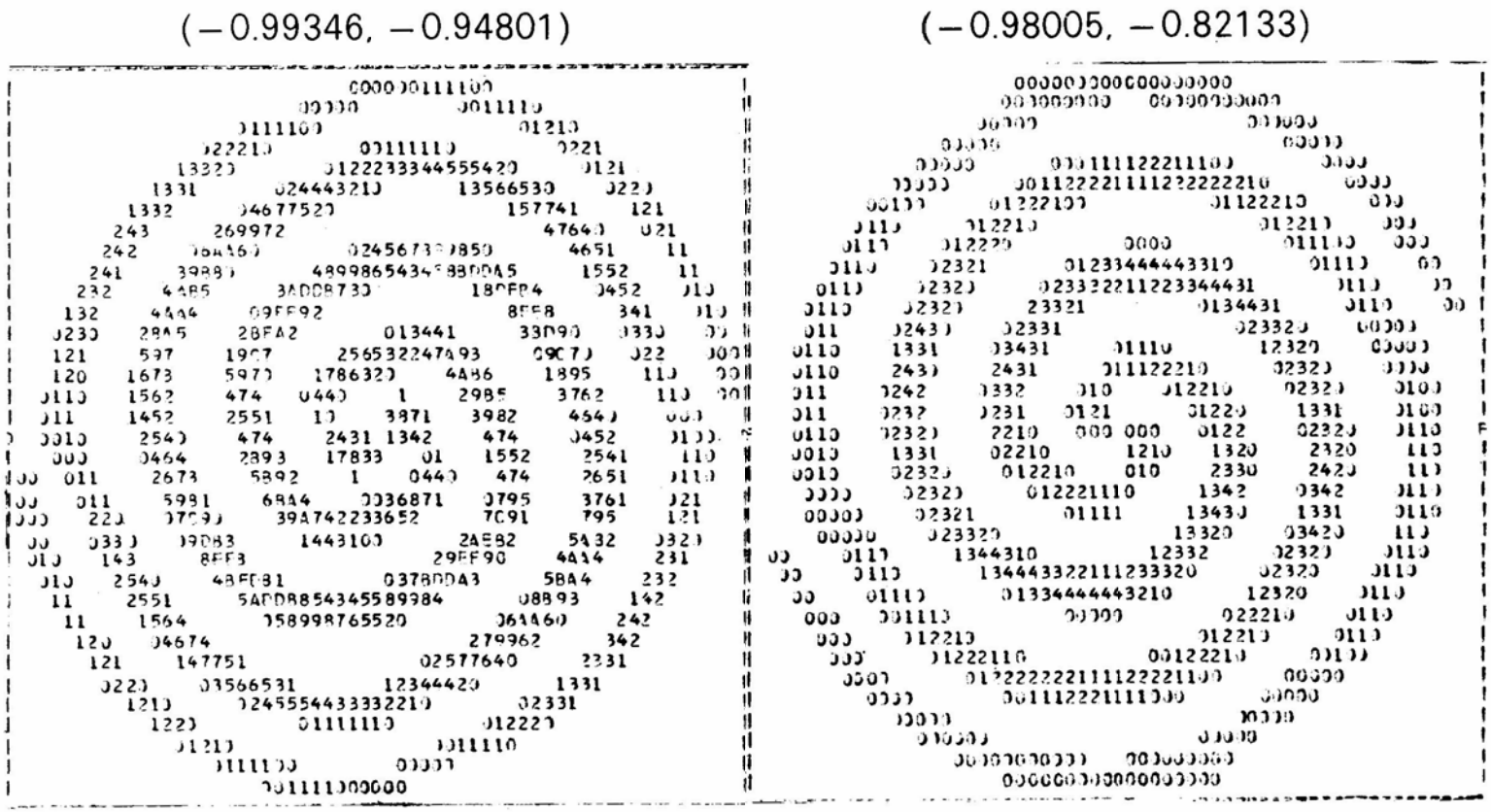


Figure 5. Eigenpatterns of oscillation in $\tilde{\sigma}(\hat{r}, \theta, \tau = 0)$ for $\eta = 1.0$, $\delta_h = 0.9$.

(−0.89248, −0.53631) (−0.78394, −0.35563)

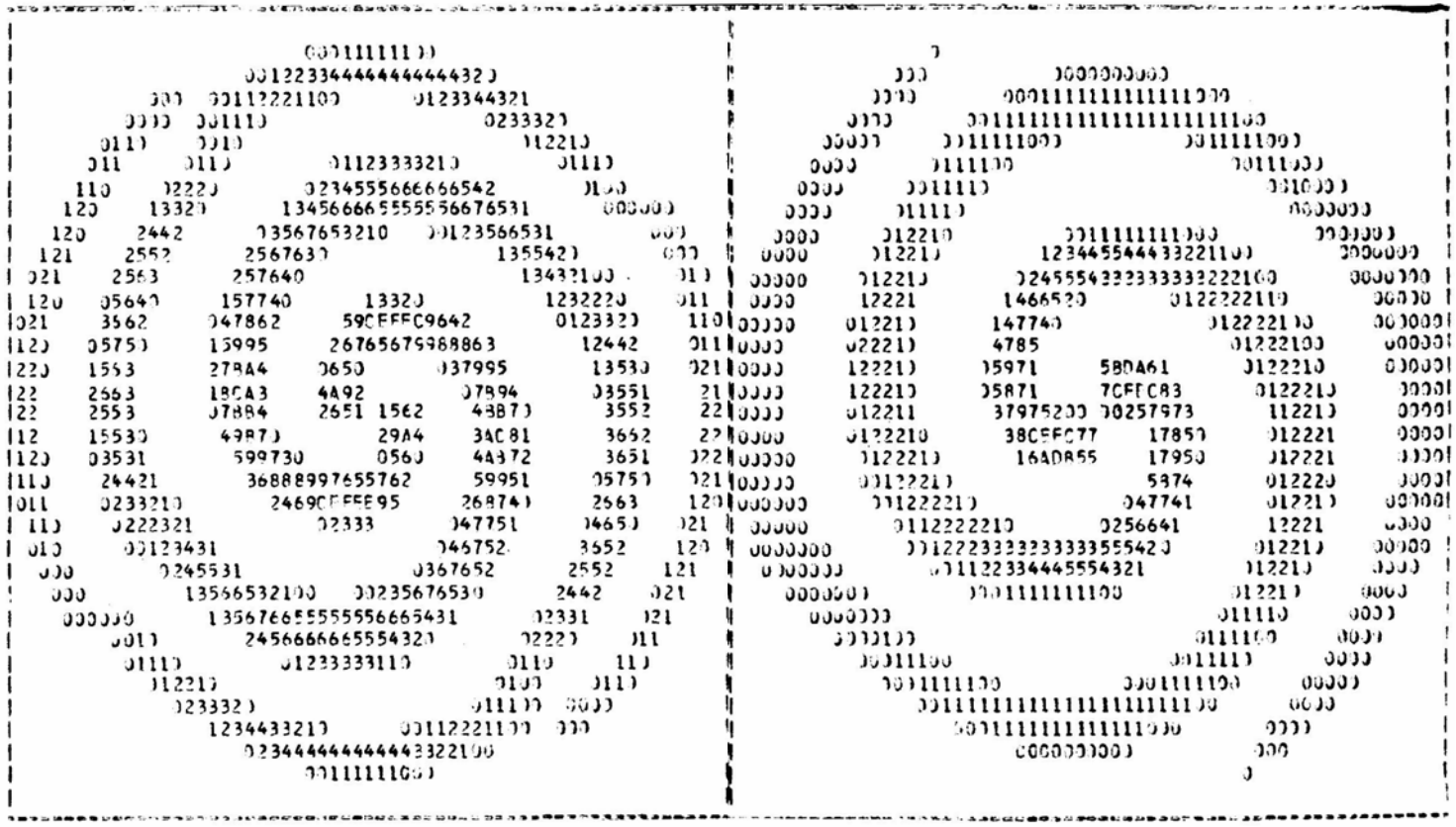


Figure 5. Continued.

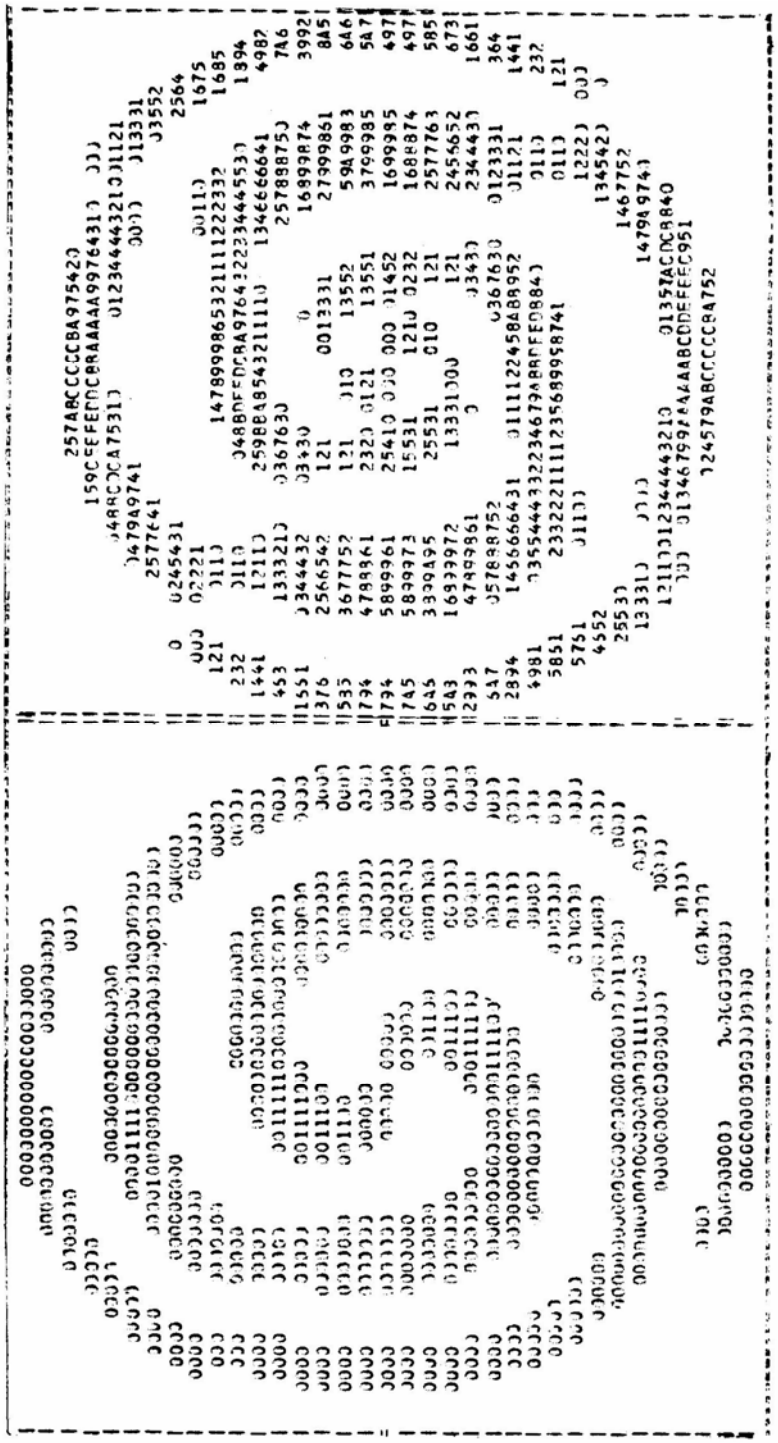


Figure 5. Continued.

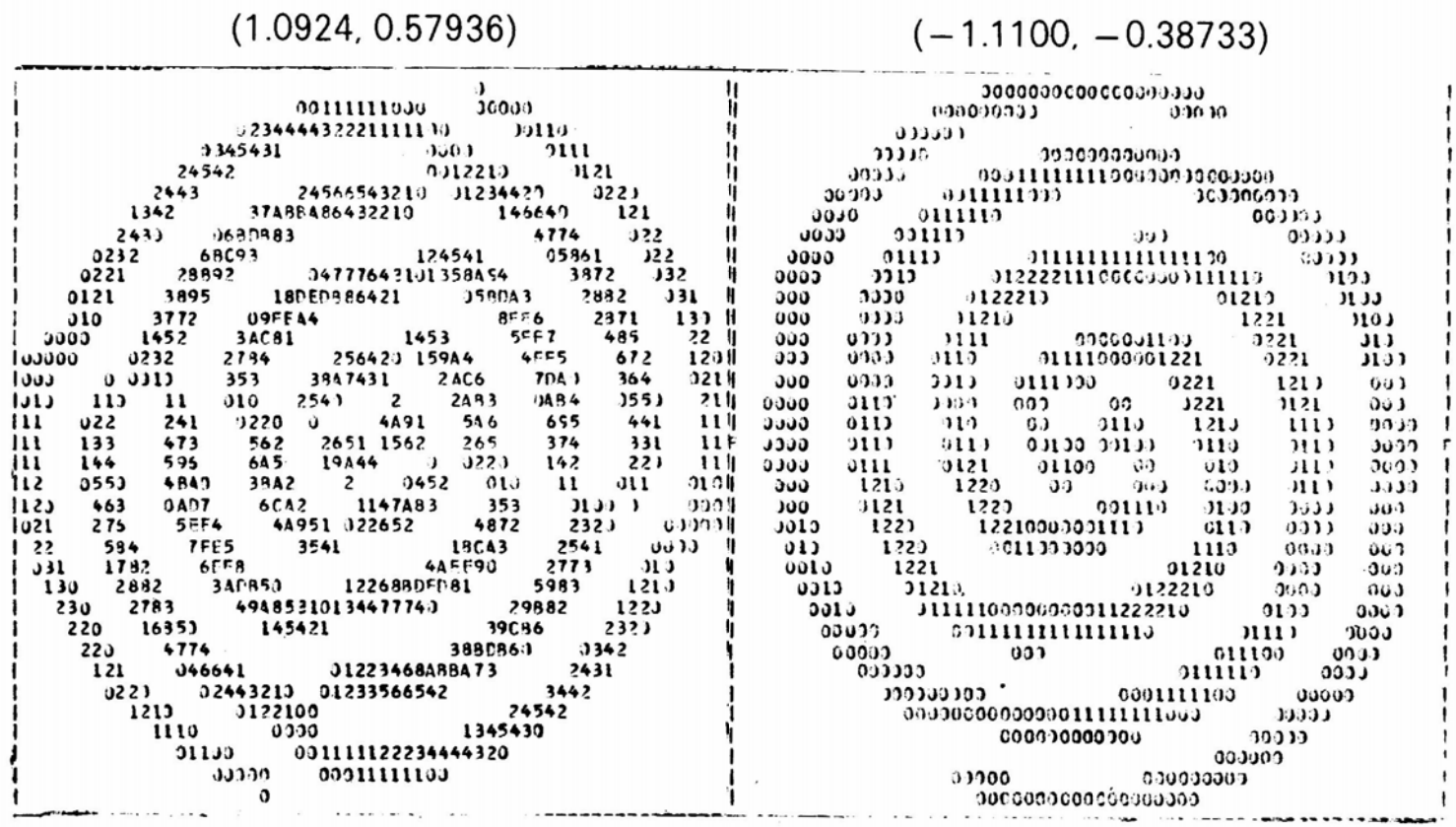


Figure 6. Eigenpatterns of oscillation in $\tilde{\sigma}(\hat{r}, \theta, \tau = 0)$ for $\eta = 1.0$, $\delta_h = 0.925$.

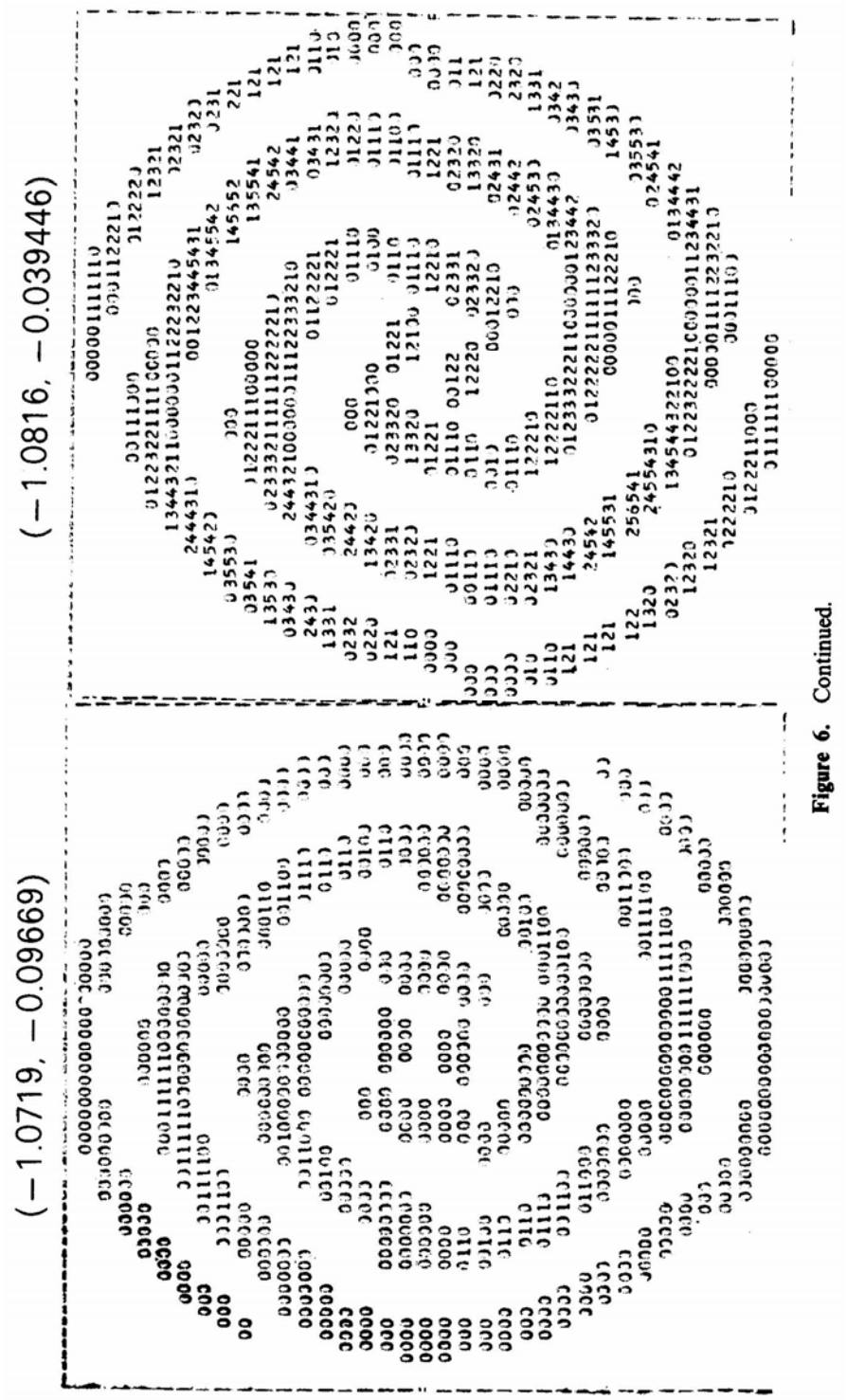


Figure 6. Continued.

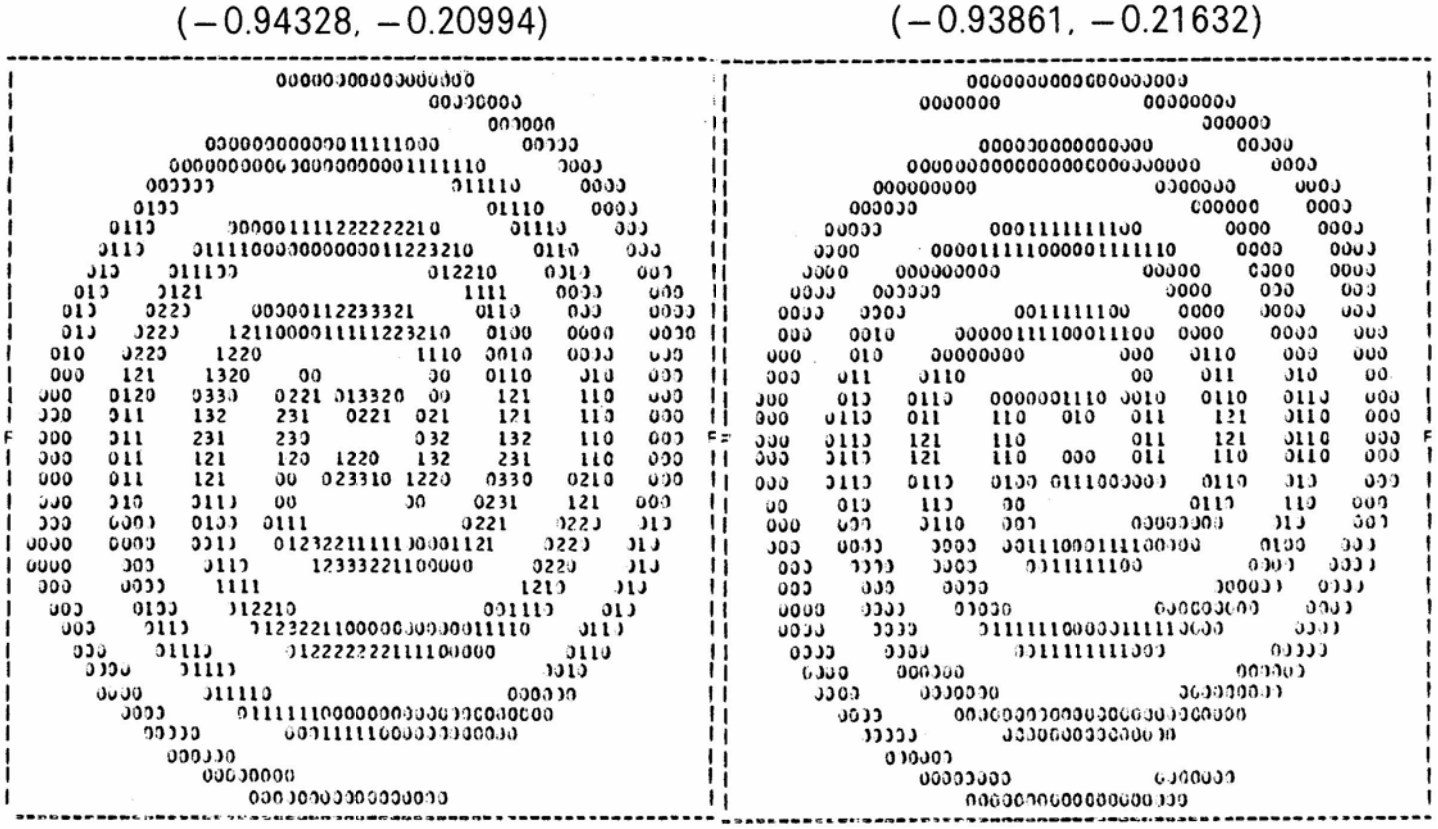


Figure 7. Eigenpatterns of oscillation in $\tilde{\sigma}(\hat{r}, \theta, \tau = 0)$ for $\eta = 10$ (left) and 50 (right), and $\delta_h = 0.950$.

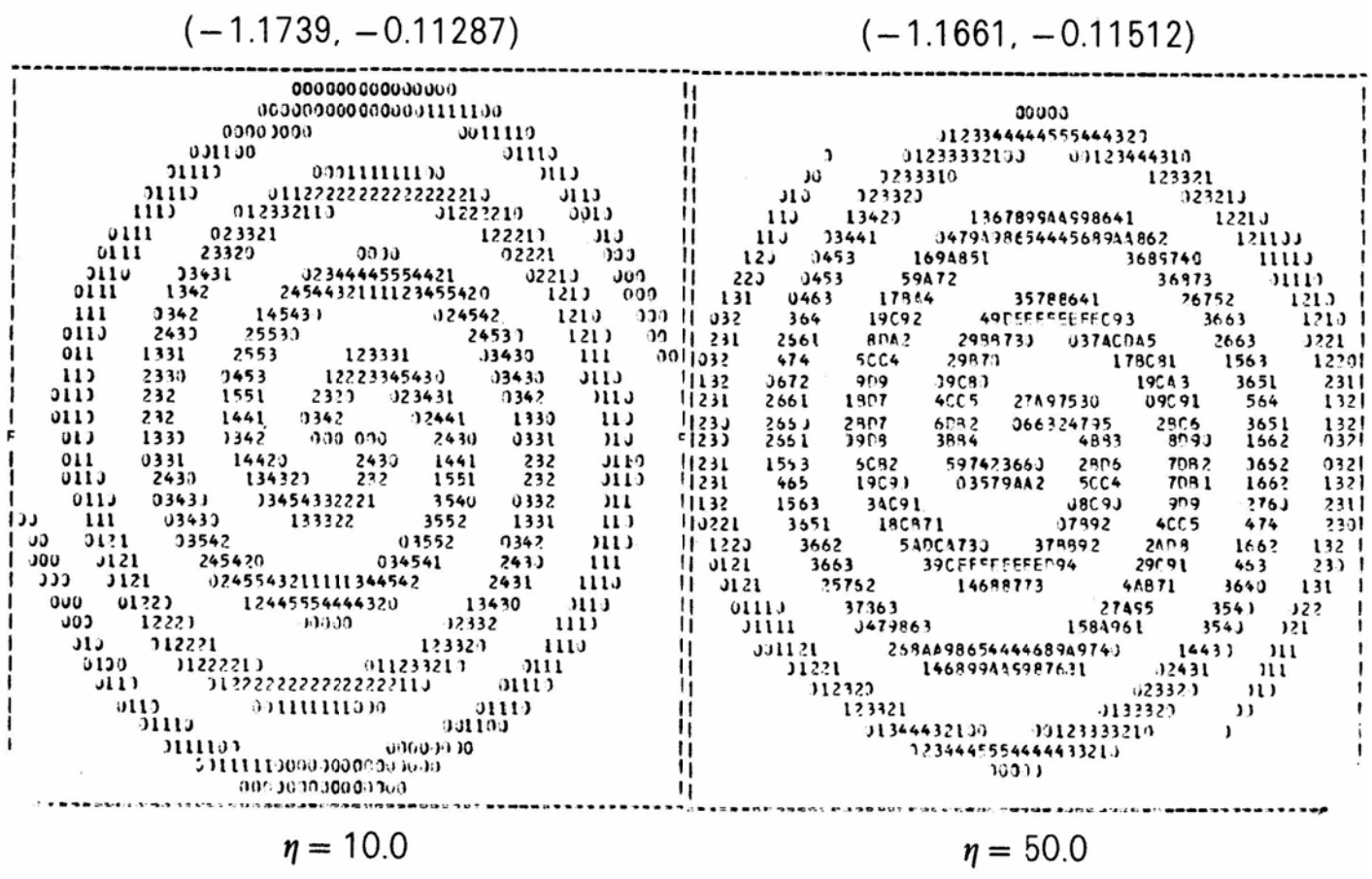


Figure 7. Continued.

Now the eigenpatterns are examined in the presence of a halo. For a disc with $\delta_h/\delta_g = 9$ and $\eta = 0.5$, we notice that many unstable modes exist; the fastest-growing mode attains an *e*-folding in less than a complete rotation of the mode around the centre of the disc, while the slowest-growing mode requires around a hundred revolutions. Almost all the modes exhibit 'leading' patterns (Fig. 4). A spherical halo with a radius smaller than the radius of the gaseous disc could be thought of as a 'bulge' in the disc as studied earlier (Ambastha & Varma 1982), wherein similar unstable leading modes were found to exist. However, the gaseous disc here continues inside the halo (or the bulge) rather than originating at $R_{\rm bulge}$ (radius of the bulge). The growth-rates of the modes decrease as the ratio δ_h/δg is increased and only two unstable leading modes are left unsuppressed for $\delta_h = 0.925$, $\delta_g = 0.075$, $\eta = 0.5$ (Table 1).

There is a complete difference in the nature of the eigenpattern for the case $\eta = 1.0$, $\delta_h/\delta_g = 9$ (Fig. 5).Here, all the modes show smooth trailing patterns. The radius of the halo in this case is the same as the radius of the gaseous disc, *i.e.*, the halo is not as strongly concentrated as in the previous case when $\eta = 0.5$. The stabilizing effect of the halo appears to have reduced as the growth-rates have increased for the corresponding eigenmodes. The fastest growing mode shows a comparatively tighter spiral pattern. This result agrees with that of Ambastha & Varma (1983).

There is a basic difference in the discs with halos when $\eta = 0.5$ and 1.0. From Fig. 3 one notices that $d\Omega/dr > 0$ over most of the disc when $\eta = 1.0$, $\delta_h = 0.9$. However when $\eta = 0.5$, $\delta_h \geqslant 0.9$, $d\Omega/dr \lesssim 0$ as is the case with the disc models in previous studies.

Fig. 6 shows the eigenpatterns of $\tilde{\sigma}$ associated with the unstable modes of the disc with $\delta_h = 0.925$, $\eta = 1.0$. The pattern-frequencies, Ω_ρ, increase and the growth-rates decrease—as is the case with the modes for $\eta = 0.5$—when halo mass δ_h is increased. Some modes have been stabilized completely. Thus, the trailing modes behave in the same manner as do the leading modes in the presence of the halo. Interestingly, one unstable mode (—1.1100, 0.38733), has become entirely trailing and another (—1.0924, 0.57936) a mixed one; trailing in the central regions and 'leading' outwards. This is an interesting case of a mode in transition.

Fig. 7 shows the eigenmodes of oscillation in the case of very diffused halos with almost uniform densities for $\delta_h = 0.95$. Most of the unstable modes of lower values of η have now been suppressed completely. However, the stabilizing effect of the halo could be seen to have diminished (Table 1). The modes with the larger growth-rates are 'leading' and the smaller ones are very smooth, trailing modes. One would notice that the modes become slightly tighter when η is increased.

5. Conclusions

Earlier studies of unstable global density waves have shown that very large pressures, such that the thermal energy is comparable to the rotational energy, are required in order to stabilize all the unstable modes (Ostriker & Peebles 1973; Ambastha & Varma 1983). In fact, it is rather difficult to construct stable disc models with a reasonable pressure distribution when no halo is present. Flat galaxies consist mainly of objects with small velocity dispersions compared to the rotation velocity. The global modes of gravitational instability would grow in a few rotation periods, unless there exists a massive enough halo in such systems.

The gaseous disc, with the surface-density profile considered here, is physically prohibited since $\Omega^2 (r) < 0$ as $r \to 0$. A massive halo helps it become kinematically stable. However, dynamically there exist a number of unstable modes, both 'trailing' and 'leading' in nature even when the halo is around nine times heavier than the disc. These modes are successively suppressed by further increasing the mass of the halo. A uniform halo is found not very effective in stabilizing the disc under consideration against the unstable modes. This does not agree with Takahara (1978). However, one would note that Takahara (1978) considered a qualitatively different disc model. It was found earlier that cold (pressureless) discs allow 'leading' modes which gradually turn into trailing modes when pressure is increased (Ambastha & Varma 1983). However, unstable trailing modes exist in the present study even in cold discs when enveloped by massive halos. Thus, it appears that 'hotness' is not a unique parameter in determining the nature of the patterns of the unstable modes. Other causes, such as the rotation law of the equilibrium disc, must be invoked. The present study on disc-halo composite systems include discs both with $d\Omega/dr > 0$ as well as, $d\Omega/dr \lesssim 0$ to facilitate some conclusions in this regard.

It is not clear as to how one would relate the results with the concepts such as the inner- and outer-Lindblad, and corotation resonances. It is basically due to the approach adopted here that the existence of resonances is not apparent. We have used an expansion in terms of Bessel functions and truncated the infinite-dimensional matrix to finite dimension. These approximations are justified by the results obtained: yet a more accurate treatment is certainly called for.

Acknowledgements

It is a pleasure to acknowledge the help provided by the staff, computer centre, Physical Research Laboratory, Ahmedabad. Thanks are due to Dr Arvind Bhatnagar, Director, Udaipur Solar Observatory for the encouragement to complete this work and to Professor G. Contopoulos for making suggestions which helped in improving the manuscript. The discussions with Professor R. K. Varma are also acknowledged.

References

Ambastha, A., Varma, R. K. 1982, *J. Astrophys. Astr.*, **3,** 125.
Ambastha, A., Varma, R. K. 1983, *AstrophysJ.* , **264,** 413.
Aoki, S., Noguchi, M., lye, M. 1979, *Publ. astr. Soc Japan*, **31,** 737.
Bardeen, J. M. 1975, in *IAU Symp.* 69: *Dynamics of Stellar Systems*, Ed. A. Hayli, D. Reidel, Dordrecht, p. 297.
Berman, R. H., Brownrigg, D. R. K., Hockney, R. W. 1978, *Mon. Not. R. astr. Soc*, **185,** 861.
Bertin, G. 1980, *Phys. Rep.*, **61,** 1.
Bertin, G., Mark, J. W. -K. 1978, *Astr. Astrophys.*, **64,** 389.
Bok, B.1981, *Mercury*, **10,** 130.
Bosma, A. 1978, *PhD Thesis*, Rijksuniv., Groningen.
Efstathiou, G., Lake, G., Negroponte, J. 1982, *Mon. Not. R. astr. Soc.*, **199,** 1069.
Gordon, M. A., Burton, W. B. 1976, *Astrophys. J.*, **208,** 346.
Hart, L., Pedlar, A. 1976, *Mon. Not. R, astr. Soc.*, **176,** 547.
Hohl, F. 1970, *NASA Tech. Rep.* **TRR** 343.
Hunter, C.1963, *Mon. Not. R. astr. Soc.*, **126,** 299.

lye, M. 1978, *Publ. astr. Soc. Japan*, **30,** 223.
Kalnajs, A. J. 1972, *Astrophys. J.*, **175,** 63.
Kato, S. 1974, *Publ. astr. Soc. Japan*, **26,** 207.
Kodaira, K. 1974, *Publ. astr. Soc. Japan*, **26,** 255.
Krumm, N., Salpeter, E. E. 1979, *Astr. J.*, **84,** 1138.
Lau, Y. Y. Lin, C. C., Mark, J. W. -K.1976, *Proc. nat. Acad. Sci. Am.*, **73,** 1379.
Lin, C. C., Shu, F. H. 1964, *Astrophys. J.*, **140,** 646.
Lin, C. C., Yuan, C, Shu, F. H. 1969, *Astrophys. J.*, **155,** 721.
Ostriker, J. P., Peebles, P. J. E. 1973, *Astrophys. J.*, **186,** 467.
Ostriker, J. P., Peebles, P. J. E., Yahil, A. 1974 *Astrophys. J.*, **193,** L1.
Pannatoni, R. F., Lau, Y. Y. 1979, *Proc. Nat. Acad. Sci. USA*, **76,** 4.
Reddish, V. C. 1968, *Q. J. R. astr. Soc.*, **9,** 409.
Rubin, V. C, Ford, W. K., Thonnard, N. 1978, *Astrophys. J.*, **225,** L107.
Salpeter, E. E.1977, *Ann. N. Y. Acad. Sci.*, **302,** 681.
Seiden, P. E., Gerola, H. 1979, *Astrophys. J.*, **233,** 56.
Stecker, F. W.1976, *GSFC Rep. X-662-76-154*, 357.
Takahara, F. 1978, *Publ. astr. Soc. Japan*, **30,** 253.
Tinsley, B. M., Cameron, A. G. W. 1974, *Astrophys. Space Sci.*, **31,** 31.
Truran, J. W., Cameron, A. G. W. 1971, *Astrophys. Space Sci.*, **14,** 179.
Toomre, A. 1964, *Astrophys. J.*, **139,** 1217.
Toomre, A. 1981, *in The Structure and Evolution of Normal Galaxies*, Eds S. M. Fall & D. Lynden-Bell, Cambridge Univ. Press, p. 111.
Yabushita, S.1969, *Mon. Not. R. astr. Soc.*, **142**, 201.

J. Astrophys. Astr. (1984) **5,** 307–316

Braking Index Diagnostics of Pulsars. I. Alignment, Counteralignment and Slowing-down Noise

Pranab Ghosh *Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400005*

Received 1984 February 19; accepted 1984 May 25

Abstract. We show that the strong correlation observed between the braking indices (n) and the slowing-down ages (τ) of pulsars is inconsistent with counteralignment between their rotation and magnetic axes, but that the data on pulsars with positive braking indices is consistent with alignment. Alternatively, slowing-down noise can quantitatively account for the data on all pulsars except the Crab and the Vela, and so for the apparent $|n| \sim \tau^2$ correlation observed for the older pulsars.

Key words: pulsars—braking index—alignment—noise

1. Introduction

Whether the magnetic and rotation axes of pulsars align or counteralign with age has been a much debated but unsettled point. The many theoretical calculations performed to date give widely different results. Davis & Goldstein (1970) and Michel & Goldwire (1970) showed that, for a perfectly conducting spherical star, the two axes align on a braking timescale, a result that can be generalized to a fluid body whose rotation deformation follows the instantaneous rotation axis (Macy 1974). In Goldreich's (1970) analysis of an imperfectly rigid, rotationally and/or magnetically distorted neutron star, the magnetic axis remains fixed relative to the symmetry axis, while the latter precesses about the instantaneous rotation axis with an amplitude which is increased or decreased by the radiation-reaction torque depending on whether the angle between the magnetic and rotation axes is large or small. Cracking and creeping of the crust, as well as internal friction, can cause alignment or counteralignment (Goldreich 1970; Chau & Henriksen 1971). Macy's (1974) analysis suggests that alignment or counteralignment can occur when a magnetic distortion moves through the star due to the kind of instability which is believed to cause polar wandering on the Earth, there being alignment or counteralignment according as the external polodial magnetic field of the pulsar is larger or smaller than $1/\sqrt{3}$ times the (volume-weighted-averaged) internal torodial magnetic field. Detailed considerations of the properties of neutron-star matter have led to contradictory results. Jones (1975, 1976) argues that alignment occurs after the star has cooled sufficiently to make the temperature-dependent dissipative torque negligible. On the other hand, Flowers & Ruderman (1977) suggest that, through internal fluid motions, the magnetic field relaxes to a minimum energy configuration of counteralignment.

The observational situation is uncertain. The presence of a strong interpulse in the Crab was used to argue that rotation and magnetic axes were nearly orthogonal in this

pulsar (Radhakrishnan & Cooke 1969). Several authors (Lyne, Ritchings & Smith 1975; Jones 1976) have argued that the strong increase in the downward slope of evolutionary tracks in the P–$\dot{P}$ plane which is caused by alignment is in better agreement with the P–$\dot{P}$ data than the Standard ($\dot{P} \propto P^{-1}$) fixed-inclination slope. Further, both Jones (1976) and Candy & Blair (1983) have argued that the observed widening of the beam with age argues in favour of alignment. On the other hand, Macy (1974) identifies type D pulsars as those which are aligning and type C pulsars as those which are counteraligning, by relating their general properties to his theory. Flowers & Ruderman (1977) have argued in favour of counteralignment by showing that, among pulsars with drifting subpulses, the drift direction reverses with age, as is predicted by their model.

Measurements of the second derivative of the pulsar period, $\ddot{P}$, (or, equivalently, the frequency second derivative $\ddot{\nu}$, or the braking index, $n = \nu\ddot{\nu}/\dot{\nu}^2$) provide us with another potential diagnostic tool for testing alignment or counteralignment. In this paper, we introduce the plot of the braking index, n, *versus* the slowing-down age, $\tau = P/\dot{P}$, as a powerful way of using the second derivative data. The available data on 19 pulsars shows that 11 of them have positive braking indices and 8 of them have negative indices and that *there are strong*, *essentially identical*, *positive correlations between* $|n|$ *and* τ *in the two classes*. We show that aligning and counteraligning pulsars have distinctly different evolutionary tracks on this plot. For those pulsars which have positive braking indices, the data are consistent with alignment on timescales $\sim 10^3$–10^5 yr, and inconsistent with counteralignment. The data on pulsars with negative braking indices are inconsistent with either alignment or counteralignment. Hence there is no evidence in the present braking index data for a general counteralignment in pulsars, contrary to Nowakowski's (1983b) suggestion.

We then investigate the alternative hypothesis that the observed braking indices are dominated by noise processes. We show that, with the exception of the young Crab and Vela pulsars, the observed correlation between $|n|$ and τ for both classes of older pulsars is consistent with the hypothesis that most of the apparent $\ddot{\nu}$ arises from a noise in the slowing-down process, and the observed $|n| \propto \tau^2$ correlation for the older ($\tau \gtrsim 10^5$ yr) pulsars is thus adequately explained by this hypothesis.

2. Alignment and counteralignment

The braking index of a pulsar slowing down via magnetic dipole radiation can be written (Macy 1974) as

$$n = 3 - \frac{25}{2}\frac{I_0\Omega^2 R}{GM^2} - \tau\left(\frac{\mathrm{d}\ln\sin^2\chi}{\mathrm{d}t} - \frac{2}{\tau_B}\right). \tag{1}$$

Here I_0 and R are respectively the moment of inertia and radius of the (rotationally) undistorted star, M is the stellar mass, Ω is the angular frequency, χ is the angle between the rotation and magnetic axes and τ_B is the timescale for magnetic-field decay. The first term on the right-hand side of Equation (1) is that given by the 'standard' model: a perfect sphere with constant magnetic field inclined at a constant angle to the rotation axis. The second term comes from rotational distortion, the effects of which are discussed in detail in Cowsik, Ghosh & Melvin (1983). The first part of the third term

describes the effects of alignment/counteralignment, and the second part those of magnetic-field decay. Rotational distortion and counteralignment decrease n from its standard value, while field decay and alignment increase it above this value. Since n has been accurately measured only for pulsars with $P \gtrsim 33$ ms (indeed, $P \sim 1$ s for most of the sample), the effects of rotational distortion are negligible for this sample (Cowsik, Ghosh & Melvin 1983), and we shall neglect these effects in what follows. For alignment, we adopt Jones' model, which gives

$$\sin\chi = \sin\chi_0 \exp(-t/\tau_a), \tag{2}$$

τ_a being the alignment timescale. This gives

$$n_a = 3 + 2\tau(\tau_B^{-1} + \tau_a^{-1}). \tag{3}$$

For counteralignment, we adopt (Flowers & Ruderman 1977; Nowakowski 1983b) the form

$$\chi = \frac{\pi}{2} - \left(\frac{\pi}{2} - \chi_0\right) e^{-t/\tau_c}, \tag{4}$$

τ_c being the counteralignment timescale. This gives (Nowakowski 1983b)

$$n_c = 3 + 2\tau\left[\tau_B^{-1} - \tau_c^{-1}(e^{-2t/\tau_c})\, 2\left(\frac{\pi}{2} - \chi_0\right) \tan\left\{\left(\frac{\pi}{2} - \chi_0\right) e^{-2t/\tau_c}\right\}\right], \tag{5}$$

the relation between the actual age and the slowing-down age being

$$\tau = e^{2t/\tau_B} \sec^2\left\{\left(\frac{\pi}{2} - \chi_0\right) e^{-2t/\tau_c}\right\}\left[\tau_0 \sin^2\chi_0 + \frac{\tau_B}{2}\{1 - e^{-2t/\tau_B}\}\right.$$
$$\left. + \tau_c\tau_B \sum_j^{0,\infty} \frac{(-1)^j(\pi - 2\chi_0)^{2j}}{2j!\,2(\tau_c + 2j\tau_B)}\{1 - \exp(-2t\tau_B^{-1} - 4jt\tau_c^{-1})\}\right]. \tag{6}$$

Here τ_0 is the value of τ at $t = 0$. The variations of n_a and n_c with τ are shown in Fig. 1. We have chosen various values of τ_B in the range 10^6–10^7 yr, in agreement with current understanding of the magnetic-field decay processes, and have tried various values of τ_a and τ_c in the range 10^3–10^7 yr. We see that, due to alignment and field decay (particularly alignment), n can become orders of magnitude larger than its standard value of 3 and thus explain the observed large positive values of n (and so the large negative values of $\ddot{P}$: see Gullahorn & Rankin 1982; Manchester *et al.* 1983; Nowakowski 1983a, b). Counteralignment does reduce n below its standard value (and can make $\ddot{P} > 0$), but does not lead to negative values of n. The counteralignment curves shown in Fig. 1 correspond to $\chi_0 = 0$ (alignment at $t = 0$), which maximizes the counteralignment effects. At large τ, the field-decay effects dominate, and the counteralignment curves also rise to large positive values of n. For the alignment curves shown in Fig. 1, different values of τ_B in the 10^6–10^7 yr range make no visible difference. Similarly, for the counteralignment curves shown, different values of τ_c in the 10^3–10^7 yr range make little visible difference. Hence, the former curves are labelled by values of τ_a, and the latter, by those of τ_B. The data on 19 pulsars for which $\ddot{\nu}$ has been measured are shown in Fig. 1 (Gullahorn & Rankin 1978a, 1982; Demiański & Prószyński 1979 and the references therein; Downs 1981) and displayed in Table 1. We have *not* included those pulsars for which only an upper limit to $\ddot{\nu}$ is available. The data fall into two distinct classes, eleven pulsars with $n > 0$ and eight with $n < 0$ (Table 1). In

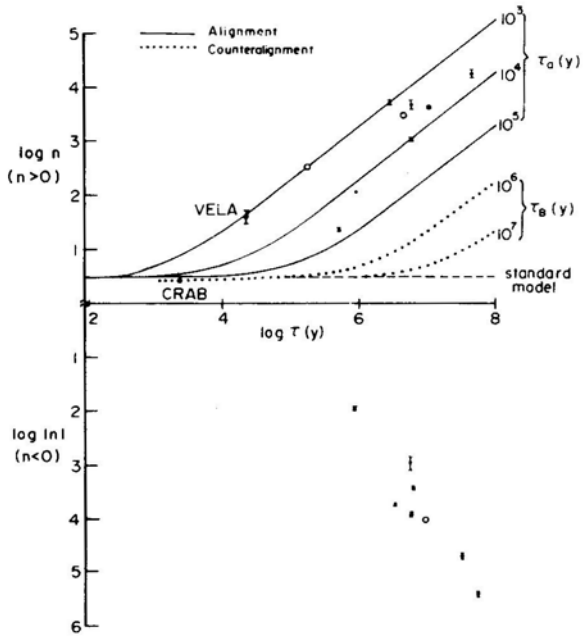


Figure 1. Braking index–slowing-down-age correlation. Plotted is the logarithm of the *absolute value of the* braking index against the logarithm of age. Also shown are the theoretical curves for alignment (solid lines), those for counteralignment (dotted lines) and that given by the Standard $n = 3$ model (dashed line). The alignment curves are labelled by alignment timescales, τ_a; different values of the field decay timescale, τ_B, in the range 10^6 yr–10^7 yr make no visible difference for these curves. The counteralignment curves are labelled by τ_B; different values of the counteralignment timescale, τ_c, in the range 10^3 yr–10^7 yr make little visible difference for these curves. Filled circles: measurements quoted with error bars as shown. Open circles: measurements quoted without error bars.

each class, there is a strong correlation (indeed an almost obvious linear relation) between $\log |n|$ and $\log \tau$. The theoretical curves for alignment are consistent with the $n > 0$ data for $\tau_a \sim 10^3$–10^5 yr (the effect on these curves of varying τ_B in the range 10^6–10^7 yr is negligible). The curves for counteralignment are inconsistent with the $n > 0$ data. (The case of Crab pulsar, which has $n \simeq 2.5$, is a special one, and is discussed later). *The $n < 0$ data are in complete disagreement with either alignment or counteralignment.* It is thus clear from the braking index data available at the present time that *there is no evidence for general counteralignment in pulsars.* This is in contradiction with the conclusion of Nowakowski (1983b), who did not attempt a quantitative comparison of theory with observation.

3. Slowing-down noise

The secular nature of the apparent second derivatives of pulsar periods has been questioned before (see Cordes & Helfand 1980 and references therein). The facts that almost half the known sample shows unexplained large negative braking indices, and that both halves show essentially the same strong correlation between the absolute

Table 1. Braking indices.

PSR	Frequency ν (Hz)	$\dot{\nu}$ (Hz s^{-1})	$\ddot{\nu}$ (Hz s^{-2})	Braking index $n = \nu\ddot{\nu}/\dot{\nu}^2$	Slowing-down age $\tau = -\nu/\dot{\nu}$	Reference
0329+54	1.400	−4.02 E−15	...	(4.81 ±0.18)E3	1.11E7	3
0531+21	30.18	−3.848E−10	(1.234±0.002)E−20	(2.515±0.005)E0	2.49E3	2
0540+23	4.066	−2.550E−13	(3.953±0.082)E−25	(2.5 ±0.05)E1	5.16E5	1
0611+22	2.986	−5.31 E−13	...	(3.5)E2	1.78E5	4
0823+26	1.884	−6.06 E−15	...	(−1)E4	9.87E6	4
0833−45	11.21	−1.57 E−11	(9.2 ±2.6)E−22	(4.2 +1.3)E1	2.27E4	5
0950+08	3.951	−3.578E−15	(−1.704±0.124)E−25	(−5.2 ±0.4)E4	3.47E7	1
1508+55	1.352	−9.20 E−15	...	(3.25)E3	4.66E6	6
1541+09	1.336	−7.682E−16	(−1.116±0.091)E−25	(−2.5 ±0.2)E5	5.52E7	1
1604−00	2.371	−1.720E−15	(2.310±0.448)E−26	(1.8 ±0.4)E4	4.38E7	1
1859+03	1.526	−1.743E−14	(1.100±0.050)E−24	(5.5 ±0.05)E3	2.78E6	1
1900+01	1.371	−7.58 E−15	(2.1 ±0.5)E−25	(5.0 ±1.2)E3	5.74E6	1
1907+00	0.983	−5.33 E−15	(−2.5 ±0.2)E−25	(−8.7 ±0.7)E3	5.85E6	1
1907+02	2.021	−1.13 E−14	(−6.0 ±2.0)E−26	(−9.5 ±3.1)E2	5.68E6	1
1907+10	3.526	−3.277E−14	(−1.689±0.061)E−24	(−5.5 ±0.2)E3	3.42E6	1
1915+13	5.138	−1.901E−13	(−6.440±0.505)E−25	(−4.2 ±0.3)E1	8.58E5	1
1929+10	4.415	−2.254E−14	(−3.180±0.150)E−25	(−2.8 ±0.1)E3	6.23E6	1
2002+31	0.4737	−1.673E−14	(7.126±0.268)E−26	(1.2 ±0.05)E2	8.99E5	1
2020+28	2.912	−1.608E−14	(1.028±0.080)E−25	(1.2 ±0.1)E3	5.75E6	1

References:
1. Gullahorn & Rankin (1978a, 1982)
2. Groth (1975)
3. Demiański & Prószyński (1979)
4. Helfand *et al.* (1980)
5. Downs (1981)
6. Manchester & Taylor (1977), p. 121

value of the braking index and the slowing-down age, make us reconsider this point. Random walks in phase, frequency, or the first derivative of frequency can give rise to apparent second derivatives. For the last case, that of the so-called slowing-down noise, the apparent second derivative, $\langle \ddot{\nu}_R \rangle$, is related to the rms residual, $\sigma_R(2,T)$, of a least-squares second-order polynomial fit to the timing data over an interval of length T through the inequality (Cordes & Helfand 1980)

$$\langle \ddot{\nu}_R \rangle \lesssim \frac{120\sqrt{7}}{T^3 P} \sigma_R(2,T) \tag{7}$$

This immediately gives

$$\langle |n_R| \rangle \lesssim (\tau/\tau_R)^2 \tag{8}$$

where

$$\tau_R = \left\{ \frac{120\sqrt{7}\sigma_R(2,T)}{T^3} \right\}^{-1/2} \tag{9}$$

is the 'random-noise age'. τ_R is determined by the nature of the noise processes and by the observation interval T. $\sigma_R(2, T)$ and T have been given by Cordes & Helfand (1980) for 16 of the 19 pulsars considered here. In Table 2, we give the values of τ_R and $\langle |n_R| \rangle$, and compare $\langle |n_R| \rangle$ with the observed $|n|$ in Fig. 2 and Table 2. For all but the Crab and Vela pulsars, there is excellent agreement, $|n|_{obs}$ always lying below the upper limit on $\langle |n_R| \rangle$ given by Equation (8) with two slight exceptions. For Crab and Vela, $\langle |n_R| \rangle_{max}$ are far below the observed values. We thus arrive at the conclusion that for aging pulsars ($\tau \gtrsim 10^5$ yr) which show apparent braking indices whose magnitudes are very much larger than 3, the second derivative may very well be dominated by slowing-down noise, whereas for the young Crab and Vela pulsars, the second derivative cannot be so dominated.

We stress that Equation (8) does not imply $\langle |n_R| \rangle \propto \tau^2$, since the random-noise age

Table 2. Slowing-down noise.

PSR	T (d)	$\sigma_R(2,T)$ (ms)	τ_R (yr)	$\langle \lvert n_R \rvert \rangle$	Observed $\lvert n \rvert$
0329+54	2859	2.31	1.44 E 5	5.95 E 3	(4.81 ±0.18) E 3
0531+21	1628	11.97	2.72 E 4	8.4 E −3	(2.515±0.005) E 0
0540+23	1368	0.52	1.00 E 5	2.64 E 1	(2.5 ±0.05) E 1
0611+22	1586	107.00	8.74 E 3	4.15 E 2	(3.5) E 2
0823+26	1834	12.55	3.17 E 4	9.68 E 4	(1) E 4
0833−45	1000	4.2 − 32.0	2.21 E 4	1.06 − 8.05 E 0	(4.2 ±1.3) E 1
0950+08	1563	0.45	1.32 E 5	6.93 E 4	(5.2 ±0.4) E 4
1508+55	2865	5.67	9.21 E 4	2.56 E 3	(3.25) E 3
1541+09	1669	0.95	1.00 E 5	3.04 E 5	(2.5 ±0.2) E 5
1604−00	2177	0.33	2.53 E 5	3.00 E 4	(1.8 ±0.4) E 4
1859+03	1227	3.98	3.08 E 4	8.13 E 3	(5.5 ±0.05) E 3
1907+10	1107	1.40	4.45 E 4	5.9 E 3	(5.5 ±0.2) E 3
1915+13	1603	1.30	8.05 E 4	1.13 E 2	(4.2 ±0.3) E 1
1929+10	1334	0.38	1.13 E 5	3.03 E 3	(2.8 ±0.1) E 3
2002+31	1607	1.39	7.82 E 4	1.32 E 2	(1.2 ±0.05) E 2
2020+28	2047	0.51	1.86 E 5	9.6 E 2	(1.2 ±0.1) E 3

Data in columns 2 and 3 are taken from Cordes & Helfand (1980)

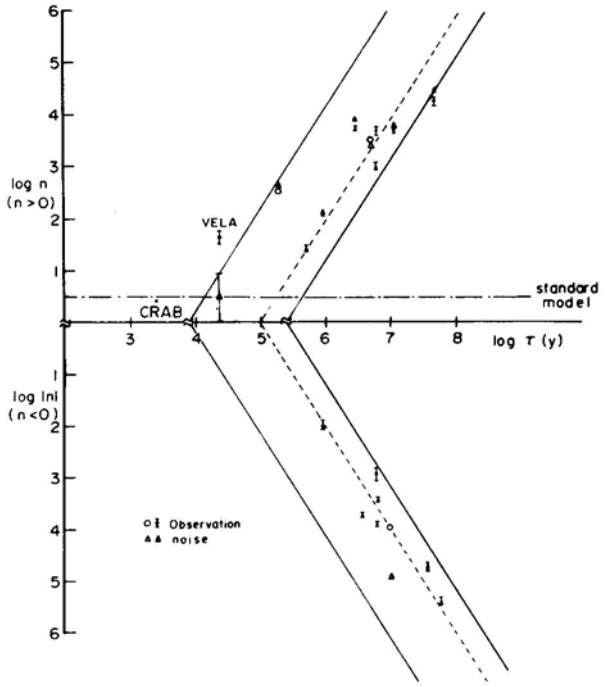


Figure 2. Same as Fig. 1, but comparing the observed braking indices (circles) with those expected from slowing-down noise (triangles). Filled circles: measurements quoted with error bars as shown. Open circles: measurements quoted without error bars. For vela, the slowing down noise component itself has an uncertainty as shown (Downs 1981). Also shown are the 'noise-band' (the area between the two solid lines which corresponding to the maximum and minimum values of τ_R in Table 2) and the relation $|n| — (\tau/10^5 \text{ yr})^2$ (dashed line) which describes most pulsars adequately.

τ_R depends on the observation interval T, on the properties of neutron stars and on those of pulsar magnetospheres. However, for most of the pulsar sample considered here, $\tau_R \sim 10^5$ yr (see Table 2), so that $|n| \simeq (\tau/10^5 \text{ yr})^2$ gives a good description to 16 of the 19 pulsars (the exceptions being Crab, Vela and PSR 0611 + 22) in Fig. 2, where the 'noise band' corresponds to the range of τ_R given in Table 2. On the basis of the slowing-down noise hypothesis, we expect future data on older pulsars to fall in the 'noise band' of Fig. 2, and to cluster around the line $|n| = (\tau/10^8 \text{ yr})^2$.

This point is brought out again by Fig. 3, where the observables $|n_{ohs}|/\tau^2$ and τ_R are plotted against each other. The shaded region, which corresponds to slowing-down noise, essentially accounts for all pulsars except Crab and Vela, which stand out far above noise. Also, most points cluster around $\tau_R = 10^5$ yr, thus giving the apparent $|n| \sim \tau^2$ correlation as before. Plots like Fig. 3 should be useful in quickly separating the slowing-down-noise component in future data on pulsars.

4. Discussion

Our main conclusions are:

1. *There is no evidence for general counteralignment in the present data on the braking indices of pulsars.*

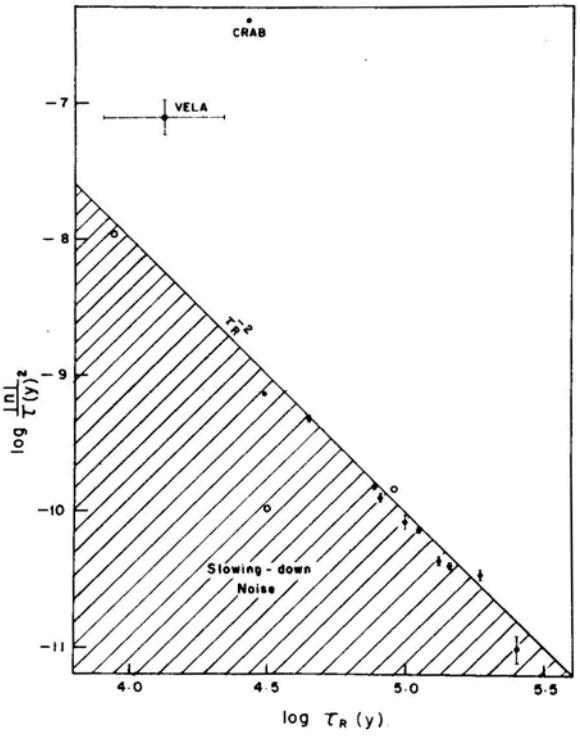


Figure 3. $|n|/\tau^2$ *vs* τ_R, both on logarithmic scales, for 16 pulsars (see Table 2). Filled circles: measurements quoted with error bars as shown. Open circles: measurements quoted without error bars. Shaded region: slowing-down noise contribution, bounded by the line $|n|\tau^{-2} = \tau_R^{-2}$. Note the particularly large uncertainties for Vela (Downs 1981).

2. Pulsars with $n > 0$ are consistent with alignment on timescales $\sim 10^3$–10^5 yr and field decay on timescales $\sim 10^6$–10^7 yr.
3. For all but the Crab and Vela pulsars, the data can be accounted for by the hypothesis that slowing-down noise dominates the measured second derivative of frequency. We predict that future data on older pulsars will fall in the 'noise band' of Fig. 2 and cluster around the $|n| = (\tau/10^5 \text{ yr})^2$ line.

In view of the last conclusion we do not give consideration to the short alignment timescales found from braking indices ($\sim 10^3$–10^5 yr). However, it is now clear from Equation (3) and Fig. 1 why short field-decay timescales $\sim 10^3$ yr were inferred by Nowakowski (1983a) on the hypothesis that the large values of n were entirely caused by field-decay effects. It is also disturbing to note that short alignment timescales imply a near alignment, and therefore a difficulty in the pulse generation, in the older pulsars. Thus, even for $n > 0$, alignment is not a particularly attractive mechanism for producing large braking indices.

Braking-index diagnostics can become most useful in studying the basic processes underlying pulsar braking torques when the random noise age τ_R can be made greater than the age of the pulsar (see Equation 8), since different physical processes will give characteristically different 'tracks' on Fig. I. Increasing the observation interval T increases τ_R (see Equation 9).

Of the two pulsars (Crab and Vela) in which slowing-down noise *cannot* account for most of the second derivative, Crab is easily understandable in terms of the standard model, since magnetospheric (Roberts & Sturrock 1972) or other effects, including those of counteralignment, can easily account for the small ($\simeq 0.5$) deviation from the standard model value of $n = 3$. Vela remains a problem, since none of the host of processes which are ordinarily thought to be able to affect the braking index (see Manchester & Taylor 1977 for a list of these processes and their effects) will increase it to $n \gg 3$. These processes include multipole electromagnetic radiation, gravitational quadrupole radiation, magnetospheric effects, rotational distortion, and pulsar proper motion (for a discussion of the inadequacy of the last process, see Gullahorn & Rankin 1982 and references therein). Alignment on a short timescale (see Fig. 1) is implausible in view of Vela's age (see above).

We note that other explanations for large braking indices suggested so far are tentative. Gullahorn & Rankin's (1978b) braking torque variations are formally similar to slowing-down noise. Doppler effects due to planets with long ($P_b \sim 50$ yr) orbital periods (Demiański & Prószyński 1979) *will* produce an apparent correlation $|n| \sim \tau^2$ over observation intervals $T \ll P_b$ for a collection of pulsars with similar planetary masses, orbital periods and inclination angles, but n will eventually show periodic variations. Finally, even if a neutron star could sustain triaxial distortions to the moment of inertia (due, for example, to misaligned rotation and magnetic axes), to achieve distortions of sufficient magnitude to explain the observed braking indices (Tademaru 1981, personal communication to Gullahorn & Rankin) is problematic, particularly for the older pulsars, in view of the known results for spheroidal distortions (see above and Cowsik, Ghosh & Melvin 1983).

What other evidence is available on the alignment/counteralignment question of our pulsar sample? As pointed out by Lyne, Ritchings & Smith (1975), a traditional 'alignment' plot of $P\dot{P}$ *vs* $P/\dot{P}$ would be unreliable because of the narrow range of P values considered in our sample. However, we do note that the two classes ($n > 0$ and $n < 0$) of pulsars seem to be very similar in other properties, including their pulse-widths, W(Manchester & Taylor 1981). It is known that $WP^{29/42}$ goes like cosec χ and so should increase with age (Jones 1976) for aligning pulsars, and that W itself should show a characteristic rise with age for old aligning pulsars (Candy & Blair 1983). We have performed these tests on our sample, from which we draw the tentative conclusions that the two classes behave essentially identically in these tests, and that both seem to be undergoing alignment on the canonical timescale $\sim 10^6$–10^7 yr.

Acknowledgements

It is a pleasure to thank Professor R. Cowsik for stimulating discussions and Mr. P. K. Ghosh for vital encouragement. An anonymous referee called my attention to an instructive plot in Lyne, Ritchings & Smith (1975).

References

Candy, B. N., Blair, D. G. 1983, *M on. Not. R. astr. Soc*, **205,** 281.
Chau, W. Y., Henriksen, R. N. 1971, *Astrophys. Lett.*, **8,** 49.

Cordes, J. M., Helfand, D. J. 1980, *Astrophys. J.*, **239,** 640.
Cowsik, R., Ghosh, P., Melvin, M. A. 1983, *Nature*, **303,** 308.
Davis, L., Goldstein, M. 1970, *Astrophys. J.*, **159,** L81.
Demiański, M., Prószyński, M. 1979, *Nature*, **282,** 383.
Downs, G. S. 1981, *Astrophys. J.*, **249,** 687.
Flowers, E., Ruderman, M. A. 1977, *Astrophys. J.*, **215,** 302.
Goldreich, P. 1970, *Astrophys. J.*, **160,** L11.
Groth, E. J. 1975, *Astrophys. J. Suppl. Ser.*, **29,** 453.
Gullahorn, G. E., Rankin, J. M. 1978a, *Astrophys. J.*, **83,** 1219.
Gullahorn, G. E., Rankin, J. M. 1978b, *Bull. am. astr. Soc.*, 9, 562.
Gullahorn, G. E., Rankin, J. M. 1982, *Astrophys. J.*, **260,** 520.
Helfand, D. J., Taylor, J. H., Backus, P. R., Cordes, J. M. 1980, *Astrophys. J.*, **237,** 206.
Jones, P. B. 1975, *Astrophys. Space Sci.*, **33,** 215.
Jones, P. B. 1976, *Nature*, **262,** 120.
Lyne, A. G., Ritchings, R. T, Smith, F. G. 1975, *Mon. Not. R. astr. Soc*, **171,** 579.
Macy, W. W. 1974, *Astrophys. J.*, **190,** 153.
Manchester, R. N., Newton, L. M., Hamilton, P. A., Goss, W. M. 1983, *Mon. Not. R. astr. Soc*, **202,** 269.
Manchester, R. N., Taylor, J. H. 1977, *Pulsars*, W. H. Freeman, San Francisco, pp. 121, 188.
Manchester, R. N., Taylor, J. H. 1981, *Astrophys. J.*, **86,** 1953.
Michel, F. C, Goldwire, H. C. 1970, *Astrophys. Lett.*, **5,** 21.
Nowakowski, L. A. 1983a, *Astr. Astrophys.*, **118,** 29.
Nowakowski, L. A. 1983b, *Astr. Astrophys.*, **127,** 259.
Radhakrishnan, V., Cooke, D. J. 1969, *Astrophys. Lett.*, **3,** 225.
Roberts, D. H., Sturrock, P. A. 1972, *Astrophys. J.*, **173,** L33.

J. Astrophys. Astr. (1984) **5,** 317–322

Cosmological Solution with an Energy Flux

Bijan Modak *Department of Physics, Presidency College, Calcutta 700073*

Received 1984 February 25; accepted 1984 June 4

Abstract. The paper presents some spherically symmetric cosmological Solutions in which the velocity field is shear-free but there is a flux of energy. The solutions are believed to be new and the previous known solutions of this class due to Bergmann and Maiti may be obtained as special cases of our metrics.

Key words: heat flow—cosmology

1. Introduction

Recently, Bergmann (1981) and Maiti (1982) have given spherically symmetric metrics corresponding to distributions of fluid with a heat flux. Bergmann imposed the condition that the fluid velocity vector is both geodetic and shear-free and then integrating the condition of isotropy of fluid pressure he obtained the metric

$$ds^2 = e^{\nu}dt^2 - e^{\mu}[dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2] \tag{1}$$

with

$$e^{\nu} = 1, \quad e^{\mu} = \frac{R^2}{[1 + kr^2/4]^2} \tag{2}$$

where R and k are undetermined functions of t alone. Obviously the essential difference between the Bergmann metric and the Friedmann metric of the isotropic universe is that in the latter the parameter k is a constant and hence, with a suitable choice of the radial variable, can be reduced to any of the values 0, +1, – 1.

Maiti, on the other hand, started with the condition that the space-time is conformally flat. Plugging in the requirements of spherical symmetry and absence of shear, Maiti was led to the solution

$$e^{\nu} = \left[A + \frac{a}{1 + kr^2/4}\right]^2, \quad e^{\mu} = \frac{R^2}{[1 + kr^2/4]^2}. \tag{3}$$

Here k, as in isotropic cosmology, is a constant (reducible to 0, +1, — 1) and A, a and R are three arbitrary functions of time. It may be noted that (as Maiti also noticed) one of the functions, A, may be removed by a transformation of the time variable:

$$t \to \int A\, dt, \quad \text{if } A \neq 0.$$

There are thus essentially two arbitrary functions of time as in the Bergmann solution.

This paper presents a number of such solutions some of which contain Bergmann's and maiti's solutions as special cases but we claim them to be more general and new.

Besides assuming spherical symmetry and shear-free velocity field, some additional ad-hoc mathematical relations are introduced to have the solution in simple closed form. The procedure is similar to that adopted by Tolman (1939) in his classical paper on static spherically symmetric fluid spheres.

2. The field equation and their integration

Consider the spherically symmetric metric (1) where v and μ are functions of radial coordinate r and time t. With the velocity vector $v^\mu = e^{-\nu/2}\,\delta^\mu_0$ this ensures that the velocity field is shear-free. Allowing for the possibility of a heat flux the energy-momentum tensor may be written as

$$T^\mu_\nu = (\rho + p)v^\mu v_\nu - p\delta^\mu_\nu + q^\mu v_\nu + q_\nu v^\mu. \tag{4}$$

Here, the coordinates t, r, θ, ϕ are numbered 0,1,2,3 respectively, and q^μ is the heat flux vector. With the metric (1), the non-trivial field equations are

$$8\pi T^0_0 = 8\pi\rho = -e^{-\mu}\left[\mu'' + \frac{2\mu'}{r} + \frac{\mu'^2}{4}\right] + \frac{3}{4}e^{-\nu}\dot{\mu}^2, \tag{5}$$

$$8\pi T^1_1 = -8\pi p = -e^{-\mu}\left[\frac{\mu'^2}{4} + \frac{\mu'\nu'}{2} + \frac{\mu' + \nu'}{r}\right] + e^{-\nu}\left[\ddot{\mu} + \frac{3}{4}\dot{\mu}^2 - \frac{\dot{\mu}\dot{\nu}}{2}\right], \tag{6}$$

$$8\pi T^2_2 = -8\pi p = -e^{-\mu}\left[\frac{\mu'' + \nu''}{2} + \frac{\nu'^2}{4} + \frac{\mu' + \nu'}{2r}\right] + e^{-\nu}\left[\ddot{\mu} + \frac{3}{4}\dot{\mu}^2 - \frac{\dot{\mu}\dot{\nu}}{2}\right], \tag{7}$$

$$8\pi T^1_0 = 8\pi q^1 v_0 = e^{-\mu}\left[\dot{\mu}' - \frac{\dot{\mu}\nu'}{2}\right], \tag{8}$$

where primes and dots indicate partial differentiation with respect to r and t respectively. Introducing the variable $\xi = [e^\mu r^2]^{1/4}$ and $x = \ln r$ we may reduce Equation (5) to the form (*cf.* Raychaudhuri 1953)

$$\frac{\partial^2\xi}{\partial x^2} = \frac{1}{4}\xi + \frac{3}{16}e^{-\nu}\dot{\mu}^2\xi^5 - 2\pi\rho\xi^5. \tag{9}$$

We obtain a family of solutions by assuming

$$8\pi\rho = \frac{3}{4}e^{-\nu}\dot{\mu}^2 + \frac{1}{2}h^2\xi^n \tag{10}$$

in an ad-hoc manner so that Equation (9) may be integrated in closed form. We then get

$$e^\mu = \frac{R_n^2}{[1 + kr^\alpha/4]^{4/\alpha}} \tag{11}$$

where $R_n = [(n + 6)k/h^2]^{4/(n+4)}$, $\alpha = (n + 4)/2$, h and k are arbitrary functions of t alone and $n \geqslant 0$, R_n, a function of t through k, h and n respectively. Eliminating p from Equations (6) and (7) we get

$$\frac{\partial Z}{\partial r} + \frac{rZ^2}{[1 + kr^\alpha/4]^{4/\alpha}} = \frac{nkr^{\alpha-3}}{4[1 + kr^\alpha/4]^{n/\alpha}} \tag{12}$$

where $Z == (\nu'/2r)[1 + kr^{\alpha}/4]^{4/\alpha}$. An integral of (12) in closed form with arbitrary n, k has not been found. Therefore we discuss the solution (12) with particular values of n and functional dependence of k.

When $n = 0$, the solution becomes

$$e^{\nu} = \left[1 + \frac{a}{1 + kr^2/4}\right]^2, \quad e^{\mu} = \frac{R_0^2}{[1 + kr^2/4]^2} \tag{13}$$

where a is an arbitrary function of t alone. The solution (13) differs from Maiti's solution only in that k is an arbitrary function of time rather than a constant. Bergmann's solution may also be obtained from Equation (13) by putting $a = 0$, which makes the t-lines geodetic. The solution (13) is again conformally flat, as is the case in both Bergmann's and Maiti's metric. When both $k = 0$ and $a = 0$, the solution (13) goes over to the standard homogeneous and isotropic model. Considered in generality, Equation (13) is thus a new solution.

We write down the expressions for ρ, p, q^1, and the expansion θ for our solution (13)

$$8\pi\rho = \frac{3k}{R_0^2} + 3\left[\frac{\dot{R}_0}{R_0} - \frac{\dot{k}r^2/4}{1 + kr^2/4}\right]^2\left[1 + \frac{a}{1 + kr^2/4}\right]^{-2}, \tag{14a}$$

$$\begin{aligned} 8\pi p = & -\frac{k}{R_0^2} - \frac{ka}{R_0^2}\left[\frac{1 - kr^2/4}{1 + kr^2/4}\right]\left[1 + \frac{a}{1 + kr^2/4}\right]^{-1} \\ & - 3\left[\frac{\dot{R}_0}{R_0} - \frac{\dot{k}r^2/4}{1 + kr^2/4}\right]^2\left[1 + \frac{a}{1 + kr^2/4}\right]^{-2} \\ & - \left[\frac{\dot{R}_0}{R_0} - \frac{\dot{k}r^2/4}{1 + kr^2/4}\right]^{-1}\frac{\partial}{\partial t}\left[\left(\frac{\dot{R}_0}{R_0} - \frac{\dot{k}r^2/4}{1 + kr^2/4}\right)^2\left(1 + \frac{a}{1 + kr^2/4}\right)^{-2}\right], \end{aligned} \tag{14b}$$

$$8\pi q^1 = -\frac{\dot{k}r}{R_0^2}\left[1 + \frac{a}{1 + kr^2/4}\right]^{-1} + \frac{kar}{R_0^2}\left[\frac{\dot{R}_0}{R_0} - \frac{\dot{k}r^2/4}{1 + kr^2/4}\right]\left[1 + \frac{a}{1 + kr^2/4}\right]^{-2}, \tag{14c}$$

$$\theta = 3\left[\frac{\dot{R}_0}{R_0} - \frac{\dot{k}r^2/4}{1 + kr^2/4}\right]\left[1 + \frac{a}{1 + kr^2/4}\right]^{-1}, \tag{14d}$$

and the nonzero component of acceleration vector

$$\dot{v}^1 = -\frac{akr}{R_0^2}\left[1 + \frac{a}{1 + kr^2/4}\right]^{-1} \tag{15a}$$

and the 3-space curvature is

$$R^* = 2(8\pi\rho - \theta^2/3) = 6k/R_0^2. \tag{15b}$$

The parameter k is thus associated with the scalar curvature R^* and the parameter a measures the non-geodesicity of the space-time.

Another new class of simple solutions may be obtained by letting $k = 0$ in Equation (11). Equation (12) then leads to

$$e^{\nu} = [1 + br^2]^2, \quad e^{\mu} = R^2, \tag{16}$$

where R and b are arbitrary functions of t alone. It is interesting to note that solution (16) does not belong to Maiti's form. Maiti's solution goes over to the Friedmann metric with flat space-section if one puts $k = 0$ and this is obtained from (16) only if

$b = 0$. We give below the expressions for ρ, p, q_1 and θ with the metric (16).

$$8\pi\rho = \frac{3\dot{R}^2}{R^2[1+br^2]^2}, \quad 8\pi p = \frac{4b - \dot{R}^2 - 2R\ddot{R}}{R^2[1+br^2]^2} + \frac{2\dot{R}}{R}\frac{br^2}{[1+br^2]^3} \tag{17a, b}$$

$$8\pi q_1 = \frac{4\dot{R}br}{R[1+br^2]^2}, \qquad \theta = \frac{3\dot{R}}{R[1+br^2]} \tag{17c, d}$$

and the nonzero component of acceleration vector

$$v_1 = -\frac{2br}{(1+br^2)} \quad \text{and} \quad R^* = 0.$$

The metric (16) apparently has a singularity as $r \to \infty$, because e^{ν} blows up, however ρ, p, q_1 and θ all vanish as $r \to \infty$. It can be shown that the scalar curvature $R_{\mu\nu\rho\sigma} R^{\mu\nu\rho\sigma}$ vanishes as $r \to \infty$. So the space-time seems asymptotically flat.

3. Solution by the method of Glass and Bergmann

In the present section we shall follow the procedure of Bergmann to obtain a solution more general than that given by him. Glass (1979) shows that for a spherically symmetric fluid in shear-free motion, the condition of isotropy of pressure gives

$$\frac{\partial^2 A}{\partial x^2} + \frac{2}{F}\frac{\partial F}{\partial x}\frac{\partial A}{\partial x} - \frac{1}{F}\frac{\partial^2 F}{\partial x^2} = 0 \tag{18}$$

where $A^2 = e^{\nu} F^2 = e^{\mu}$ and $x = r^2$. The above equation holds irrespective of whether q^{μ} vanishes or not. The integration is easiest if one puts $e^{\nu} = A^2 = 1$ (in which case the motion becomes geodetic) and this was indeed done by Bergmann. However one may integrate the equations easily by plugging in other functional forms of A (or F). We exhibit only one case. Assuming

$$A^2 = e^{\nu} = [1 + br^2]^2,$$

we readily find

$$e^{\mu} = R^2\left[1 + kr^2 + kbr^4 + \frac{kb^2}{3}r^6\right]^{-2} \tag{19}$$

where R, k, b are arbitrary functions of t alone. The role of R as the curvature parameter no longer holds now, although it does have that significance in the neighbourhood of the centre of symmetry. We go over to the Bergmann solution by putting $b = 0$, while $k = 0$ leads to our solution (16). The general solution gives non-geodetic metric as $(e^{\nu})' \neq 0$. The solutions (13) and (16) are conformally flat but with nonzero value of k and b, whereas (19) is not. It is interesting that in this non-conformally flat solution, the physical variables like energy density *etc.* diverge as $r \to \infty$.

4. Concluding remarks

Maiti started with the condition of conformal flatness. One may therefore wonder why he did not obtain our general solutions (13) and (16). The reason is that on feeling that

the three space metrics (t = constant) are spaces of constant curvature he wrote down the metric as

$$ds^2 = g_{00}\,dt^2 - \frac{R^2}{(1+kr^2/4)^2}[dr^2 + r^2\,d\theta^2 + r^2\sin^2\theta\,d\phi^2]$$

and inadvertently introduced the assumption of k being independent of time which is however not demanded by the situation.

We have already noted that the essential generality obtained in our solution is that we can go over to both Maiti's metric and Bergmann's metric as special cases of Equation (13).

What, if any, is the significance of the functions a and b appearing in our solution (13) and (16)? Obviously they have resulted in non-geodesicity of the velocity field. Naturally they are also linked with the heat flux and spatial dependence of the expansion scalar.

Thermodynamically, one associates a temperature non-uniformity with heat flux; the heat flux in the rest frame (*i.e.* the comoving frame as used here) may occur either by radiation or by conduction. A radiation flux would complicate matters as then the Maxwell equations have to be considered. Here we therefore take the simple case of conduction and assume the phenomenological relation

$$q_\alpha = -\kappa(\delta^\beta_\alpha - v^\beta v_\alpha)(T_{,\beta} - T\dot{v}_\beta) \tag{20}$$

where K is the heat conductivity and T, the temperature. In the present case this yields

$$q_1 = -\kappa e^{-\nu/2}[Te^{\nu/2}]_{,1}. \tag{21}$$

A non-vanishing heat flux thus means a temperature gradient as may be expected to occur following the formation of a gravitational condensation. However, in the present paper, we do not make any attempt to develop a physically realistic picture of such a situation, but we estimate the temperature in a simple case.

With $e^\nu = 1$ as in Bergmann's solution, this would directly link up the heat flux with the temperature gradient. In our case the calculation of T in general leads to quite complicated expressions. A comparatively simple formula is obtained with the metric (16). From Equation (21) we get

$$\kappa T = -\frac{\dot{R}}{4\pi R}\frac{\ln(1+br^2)}{(1+br^2)} + \frac{f}{(1+br^2)}, \tag{22}$$

where we have assumed the conductivity K to be constant and f is an arbitrary function of t alone.

Acknowledgement

The author thanks Professor A. K. Raychaudhuri for valuable guidance, the referee for helpful suggestions which have lead to an improved presentation and CSIR for the award of a fellowship.

References

Bergmann, O. 1981, *Phys. Lett.*, **82A,** 384.
Glass, E. N. 1979, *J. Math. Phys.*, **20,** 1908.
Maiti, S. R. 1982, *Phys. Rev.*, **D25,** 2518.
Tolman, R. C. 1939, *Phys. Rev.*, **55,** 364.
Raychaudhuri, A. K. 1953, *Phys. Rev.*, **89,** 417.

J. Astrophys. Astr. (1984) **5**, 323–330

Eruptive Prominences of 1980 April 27 Observed during STIP Interval–X

Rajmal Jain, A. Bhatnagar & R. N. Shelke *Vedhshala, Udaipur Solar Observatory. 11, Vidya Marg, Udaipur 313001*

Received 1984 March 15; accepted 1984 June 4

Abstract. Observations and analyses of two similar eruptive prominences on the north-east limb observed on 1980 April 27 at 0231 and 0517 UT, which are associated with the Boulder active region No. 2416 are presented. Both the eruptive prominences gave rise to white-light coronal transients as observed by C/P experiment of High Altitude Observatory on the Solar Maximum Mission. Type II and moving type IV radio bursts are reported in association with the first Hα eruptive prominence at 0231 UT.

Both the Hα eruptive prominences showed pulse activity with a quasi-periodicity of about 2–4 min. We estimate a magnetic field in the eruptive prominence of about 100 G and a build-up rate $\sim 10^{26}$ ergs^{-1}. The high build-up rate indicates that the shearing of the photospheric magnetic field, which fed the energy into the filament, was rapid. It is proposed that fast-moving Hα features must have initiated the observed coronal transients. From Hα, type II and coronal-transient observations, we estimate a magnetic field of 2.8 G at 1.9 $R_\odot$ from the disc centre, which agrees well with the earlier results.

Key words: Sun—eruptive prominences—mass ejection—coronal transients—magnetic field

1. Introduction

Time-lapse Hα observations of two eruptive prominences at the northeast limb of the Sun were made from the Vedhshala, Udaipur Solar Observatory on 1980 April 27, at 0231 and 0517 UT. These two eruptive prominences were associated with the Boulder active region No. 2416. Both the eruptive prominences gave rise to white-light coronal transients observed at 0241 and 0538 UT by C/P experiment of High Altitude Observatory (HAO) on the Solar Maximum Mission (SMM) satellite. The Culgoora Radio Observatory has also reported type II and moving type IV radio bursts associated with the eruptive prominence of 0231 UT.

In this paper, we present a detailed study of the development of these two eruptive prominences with a view that a comprehensive study could be made in collaboration with radio, coronal-transient and interplanetary traveling-wave observations.

2. Observations

2.1 *Eruptive Prominence at 0231 UT*

Time-lapse observations of the eruptive prominence at the northeast solar limb were made at an interval of 6 and 10 seconds in the Hα line centre, through a Halle filter of 0.5 Å passband in conjunction with a 15-cm aperture solar spar telescope. A total of 172 frames were obtained for the event. This eruptive prominence was perhaps associated with the Boulder active region No. 2416 behind the eastern limb. The observations were started at 023141 UT, when the mass ejection had already started and thus the beginning of the event was missed. This event gave rise to a coronal transient observed with C/P experiment and also type II and type IV radio bursts were recorded at Culgoora. Development of the eruptive prominence in Hα is shown in Fig. 1 and line drawings of this activity are shown in Fig. 2. The prominence motion also has a component in line of sight, but as mentioned earlier, the observations were made using a 0.5 Å passband Hα filter, so measurement for height and velocity refer to the sky plane. In the first frame of our observations at 023141 UT (Fig. 2) a huge mass A rose to a height of about 26320 km above the solar limb. At 023159 UT the top half of the feature A fragmented into small pieces as shown in Fig. 2. Within 30 s, at 023228 UT, all the small features seem to have vanished. The top of the remaining part of feature A, indicated as 'a' in Fig. 2, also fragmented into small pieces at 023234 UT. After about 23 s, all the small fragmented features vanished.

This eruptive prominence activity continued until 024630 UT, a total duration of about 20 min. Our time-lapse observations show various features which shot out from the prominence. Among all these features, e and f attained the highest velocity of ejection of about 770 km s^{-1}, whereas feature h attained velocity of about 540 km s^{-1}. On the other hand, features i, n and m were observed as a coronal condensation; they were showing a downward motion towards the solar surface. During the 20 min period of observations at least 4 distinct 'pulses' of activity took place in the underlying active region, which gave rise to four kinks in the diagram of maximum height *vs* time (Fig. 3). The beginning of these pulses appeared at 023153, 023417, 023533 and 023821 UT, thus indicating that a recurring pulse activity with a quasi-periodicity of 2–4 min was responsible for the observed repeated eruptions of the prominence material.

2.2 *Eruptive Prominence Activity at 0517 UT*

On 1980 April 27, the same Boulder active region No. 2416 again gave rise to a violent eruptive prominence activity about three hours after the first manifestation, beginning around 0517UT. The Hα filtergrams and line drawings of the development of the prominence activity are shown in Figs 4 and 5 respectively. At 051728 UT the prominence material A rose to a height of about 54100 km. Over and above the mass A, features b and c were seen at 051754 UT. The feature b vanished from the frame at 051813 UT. The feature c was seen to have expanded in size before it fragmented into a number of small pieces at 051849 UT. One of the pieces of this fragmented material, designated as c rose to a maximum height of about 27800 km with a velocity of 270 km s^{-1}, whereas another piece e moved upwards with velocity of 510 km s^{-1} and vanished at 051923 UT. The maximum height and velocity attained by various features in this

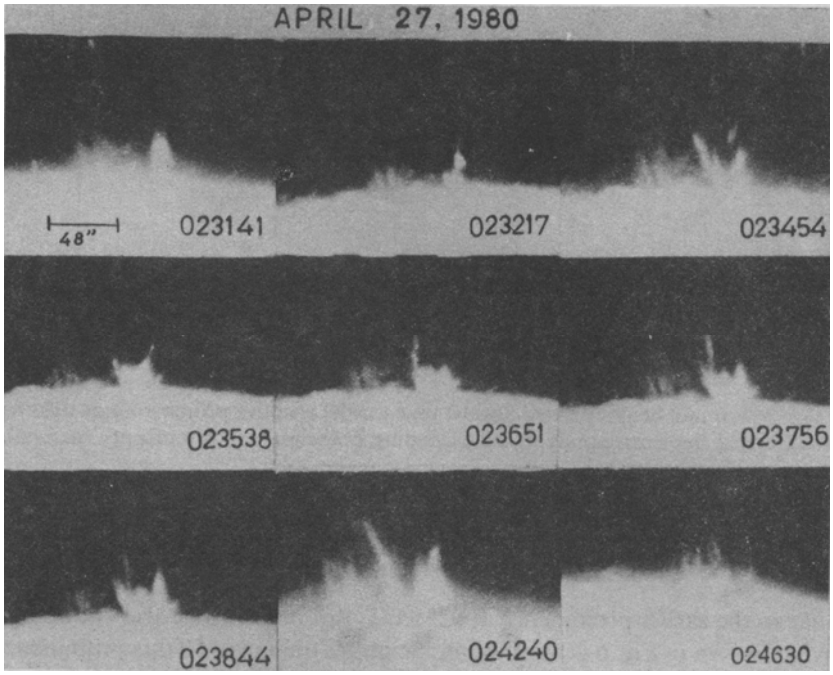


Figure 1. Sequence of Hα filtergrams, showing the development of the eruptive prominence of 1980 April 27 at 0231UT.

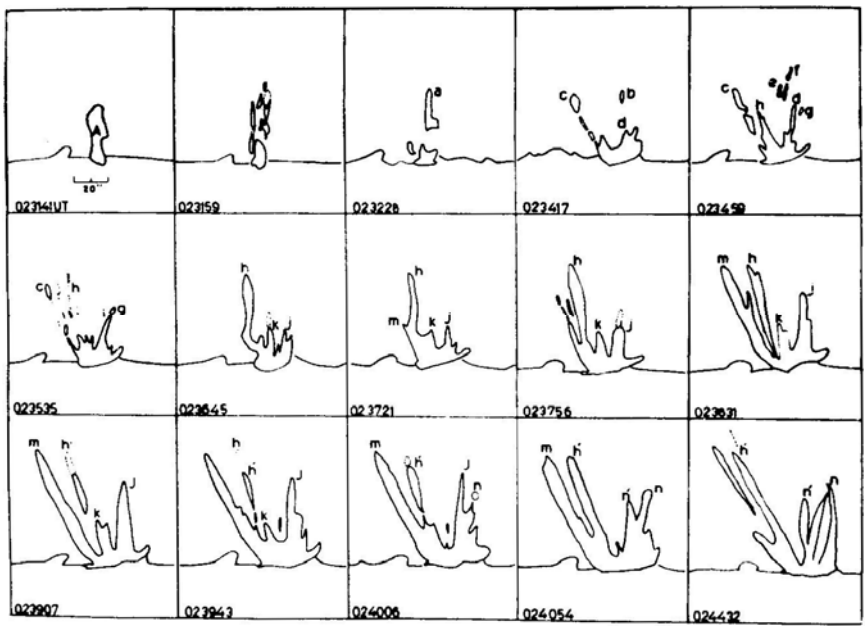


Figure 2. Line drawings of the sequence of the eruptive prominence activity of 0231 UT.

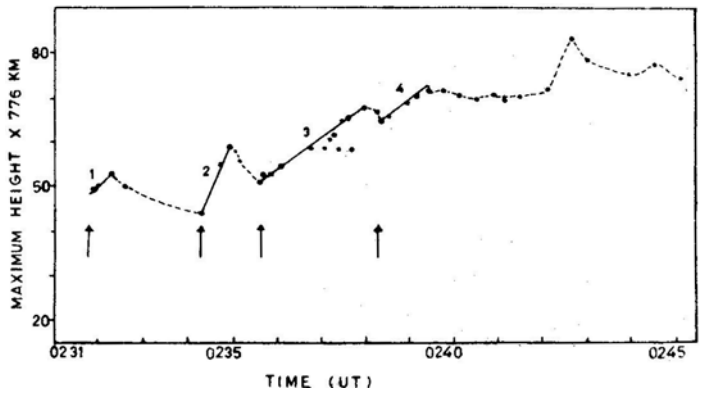


Figure 3. Maximum height plotted against time for the eruptive prominence of 1980 April 27 at 0231 UT. Solid line correponds to the ascending branch of the prominence material. Four distinct kinks are indicated.

eruptive prominence observed through the 0.5 Å passband Hα filter are given in Table 1.

Similar to the earlier prominence at 0231 UT, this prominence also showed 'pulses' of activity. Shown in Fig. 6 is a maximum height *vs* time plot. In this prominence, five distinct pulses of activity were observed with a quasi-periodicity of 2–4 min which is similar to the earlier prominence.

3. Estimation of magnetic field and energy

The strength of the magnetic field in the eruptive prominence may be estimated from the fact that its outward acceleration increased steadily through the period of observations (Engvold, Malville & Rustad 1976). Thus the eruptive prominence material could not have been on a ballistic trajectory but must have been driven and

Table 1. Maximum height and velocity attained by different features in the prominence at 0517 UT.

Features	Maximum height 10^3 km	Velocity km s^{-1}	Remarks
c	78	270	erupted from A
e	74	510	detached from C
h	84.5	582	erupting feature from a
g	101.0	770	
f	66	910	shot out from A
t	61.5	780	shot out from A
s	48.3	145	
i	108.7	360	the tip of B; detached from it
j	62.6	220	newly rising spike in the region
k	117.5	300	rising from A

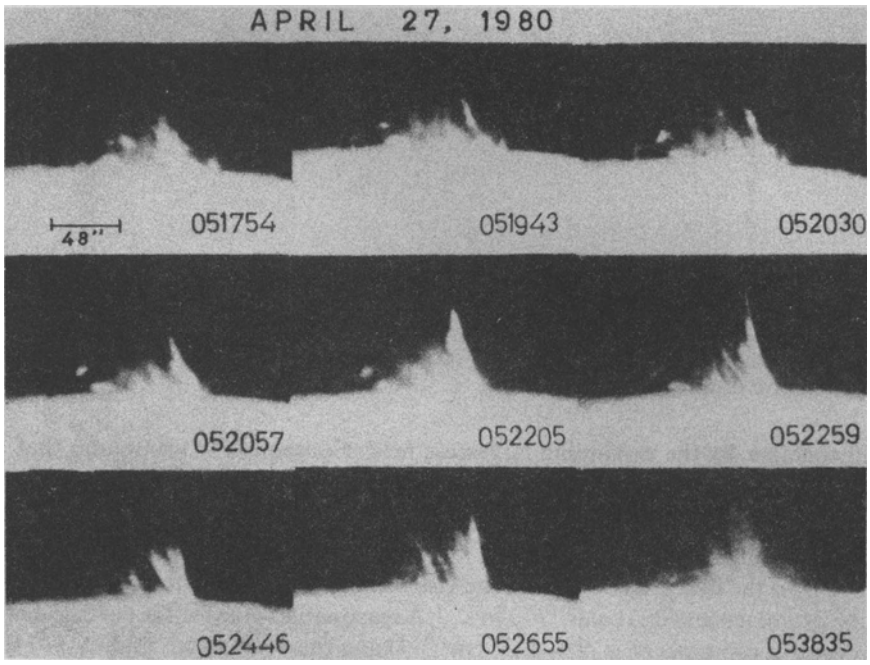


Figure 4. Sequence of Hα filtergrams of the eruptive prominence of 1980 April 27 at 0517 UT.

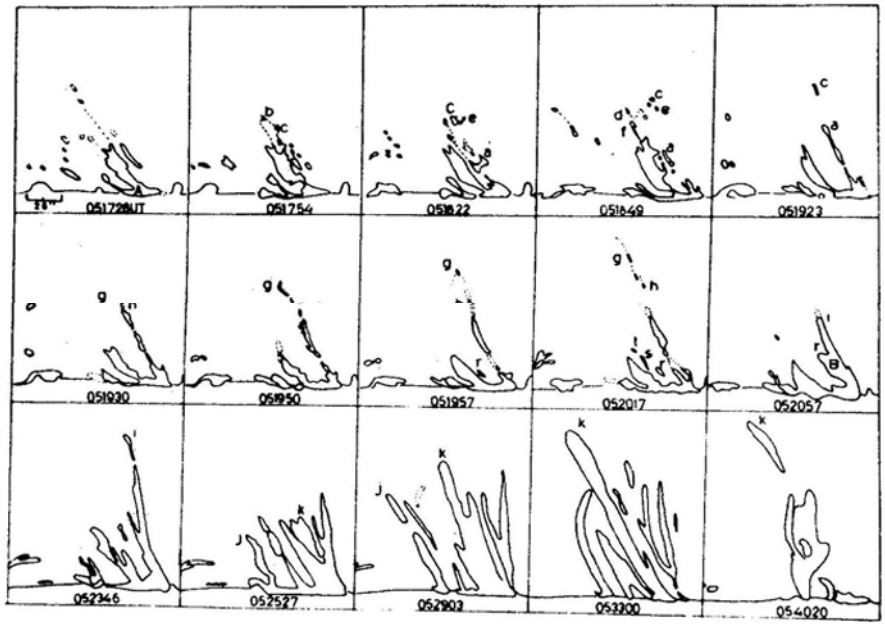


Figure 5. Line drawings of the sequence of the eruptive prominence activity of 0517 UT.

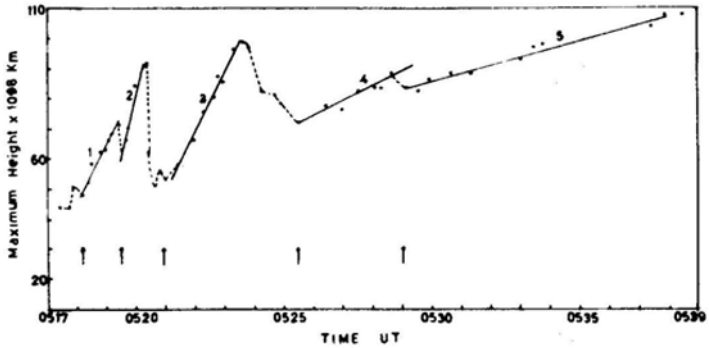


Figure 6. Maximum height plotted against time for the eruptive prominence event at 0517 UT indicating five distinct 'kinks'.

held together by the expanding largescale field. Consequently we require that

$$\frac{B_0^2}{4\pi\rho v^2} > 1 \tag{1}$$

where B_0 is the field associated with the visible material. The greatest projected velocity in the prominence at 0231 was 770 km s^{-1} . A reasonable value for the particle density in an eruptive prominence is $N = 10^{11}$ cm^{-3}. Using these values we find $B_0 > 113$ G in the prominence at 0231 UT and $B_0 > 133$ G in the prominence at 0517 UT.

The eruptive material with a velocity greater than V_{esc} cannot be contained by the magnetic field of strength ~ 100G because to control an object with velocity V_{esc} requires a magnetic field of at least 275 G (Tandberg-Hanssen 1974).

The two eruptive prominences of 1980 April 27 appear to have been similar events. We estimate the total mass (M_{prom}) of each prominence as seen in Hα from

$$M_{prom} = V_{prom}\rho \tag{2}$$

where V_{prom} is the volume of the prominence deduced from Hα line-centre data.

We determine the true height of the prominence above the limb to be 101750 km at 0533 UT. The width of the prominence is about 63000 km. Taking into account the fine structure of the prominence, we estimate its effective thickness to be about 1200 km. Thus we get $V_{prom} \simeq 7.3 \times 10^{27}$ cm^3. If the particle density $N = 10^{11}$ cm^{-3}, we get a total mass of 1.24×10^{15} gm as seen in Hα.

The initial velocity of mass ejection is 230 km s^{-1}. According to the mass estimate given above, the kinetic energy in the Hα prominence would be

$$E_{kin} = 4.5 \times 10^{29} \text{ erg}, \tag{3}$$

and the potential energy at a height 111750 km above the photosphere

$$E_{pot} = 3.8 \times 10^{29} \text{ erg}. \tag{4}$$

The fact that two similar prominences erupted in the same region, on 1980 April 27, would mean that a build up of energy of 10^{30} erg took place in a time interval of 2 h 46min ($\simeq 10^4$ s). This corresponds to an energy buildup rate

$$\frac{dE}{dt} \simeq \frac{10^{30}}{10^4} \simeq 10^{26} \text{ erg s}^{-1}. \tag{5}$$

This high build-up rate indicates that the shearing of the photospheric magnetic field which feeds the energy into the filament was fast.

4. Discussion

Munro *et al.* (1979) reported a good correlation between high-speed flare-associated Hα phenomena and mass ejections observed in white light during the Skylab mission. The flares, associated with coronal transients were accompanied by high-velocity ejections, sprays or eruptive prominences. Many coronal transients are associated with dynamic phenomena in the chromosphere which occur without flares. Over 70 per cent of all observed coronal transients are known to be associated with the eruption of solar prominences (Pneuman 1980). Thus in some Hα mass ejections, some parts can be thrown off with very high speed in various directions and can be seen in the centre of the Hα line as far as 3×10^5 km above the photosphere and much farther in the white light corona.

HAO's C/P experiment on the SMM satellite observed two white light coronal transients at 0241 and 0538 UT on 1980 April 27, which may be associated with observed Hα eruptive prominences at 0231 and 0517 UT, respectively. The coronal transients associated with the eruptive prominences at 0231 UT and at 0517 UT were at position angle 100° and 92° as reported by Research Observatory results for solar events selected for collaborative study by the Solar Maximum Year (SMY) study coordinators on 1981 August 31. The coronal transient associated with prominence at 0231 UT was a rising loop with a dark curved edge which moved with a velocity of ~ 610 km s^{-1}. In Hα, the observations of the eruptive prominence in the sky plane showed that several bits and pieces of material were moving outwards. However, many of them returned to the solar surface. On the other hand, a few pieces moved with a very high speed. In the case of prominence at 0231 UT, the feature b fragmented at 023447 UT into small bits e and f which shot out with a velocity of about 774 km s^{-1} . Similarly, in the eruptive prominence of 0517 UT many features (h, g, f & t) showed velocities ranging from 580 to 910 km s^{-1} . The velocities of these features are similar to the observed velocities of the coronal transients.

Our Hα time-lapse observations indicate that prominence material or fragments of material shot out recurrently to great heights. At least 4 and 5 distinctive 'thrusts' are clearly identified which might have taken place in the active region during the first and the second prominence respectively. We suggest that the recurrent rise of prominence material was caused by these thrusts or 'pulses' injected into the corona. These pulses showed a quasi-periodicity of 2 to 4 min.

Type II (0245–0253) and moving type IV radio bursts in metric band were observed at the Culgoora radio observatory. The optical disturbances seen in Hα moving with a very high speed and attaining greater heights (coronal transients) produce shocks which could be observed as type II radio bursts. Gosling *et al* (1975) have reported type II events in association with coronal transients and Hα mass ejections. The observed velocities of fragments of material (feature e and f) are sufficient to generate a shock which might have produced type II burst at 0245 UT. An eruptive prominence can give rise to a moving type IV burst (Robinson & MacQueen 1975). The moving type IV bursts are associated with sprays as seen in Hα and as expanding loops, arches, or

bottles in the white-light corona. The observed coronal transient, in association with the prominence at 0231 UT, was an expanding loop. It seems that both type II and moving type IV radio bursts were generated by the eruptive prominence at 0231 UT. Assuming that the observed type II burst at 0245 UT in the metric band was produced by the shock generated by the eruptive prominence material e and fat 023417 UT, we estimate the velocity of shock to be about 1100 km s^{-1}. The electron density at a height 0.9 $R_\odot$ above the photosphere where the type II burst was observed by Culgoora radio observatory is estimated to be 10^8 cm^{-3}. Type II bursts are usually interpreted as originating in weak shocks, of magnetic Mach number $M \leqslant 2$ (Smerd, Sheriden & Stewart 1975). If we assume M = 1.8, we obtain an Alfvén velocity V_A = 611 km s^{-1}, which corresponds closely to the speed (610 km s^{-1}) of the rising loop observed in the associated coronal transient. This supports the results of Smerd, Sheriden & Stewart that type II bursts originate in weak shocks. Further, taking V_A = 611 km s^{-1} and N_e = 10^8 cm^{-3}, we estimate the magnetic field to be B = 2.8 G at 1.9 $R_\odot$ from disc centre, a value which falls well within the range of magnetic field strengths (1.0–10.0 G) derived from radio observations at 2.0 $R_\odot$ from the disc centre by Dulk & McLean (1978).

Acknowledgements

It is a pleasure to thank Dr Ashok Ambastha for several discussions. Financial support for this work has come from the Department of Science &Technology, Government of India, under SERC scheme.

References

Dulk, G. A, McLean, D. J. 1978, *Solar Phys.,* **57**, 279.

Engvold, O, Malville, J. M, Rustad, B. M. 1976. *Solar Phys.,* **48**, 137.

Gosling, J. T., Hildner, E., MacQueen, R. M., Munro, R. H., Poland, A. I., Ross, C. L. 1975, *Solar Phys.,* **40**, 439.

Munro, R. H., Gosling, J. T., Hildner, E., MacQueen, R. M., Poland, A. I., Ross, C. L. 1979, *Solar Phys.,* **61**, 201.

Pneuman, G. W. 1980, *Solar Phys.,* **65**, 369.

Robinson, R., MacQueen, R. M. 1975, *Bull. Am. Astr. Soc.*, **7**, 348.

Smerd, S. F., Sheriden, K. V., Stewart, R. T. 1975, *Astrophys. Lett.,* **16**, 23.

Tandberg-Hanssen, E. 1974, *Solar Prominences,* D. Riedel, Dordrecht, p. 103.

J. Astrophys. Astr. (1984) **5**, 331–348

Is the Universe Flat?

Martin J. Rees *Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, England*
(*Invited article*)

Abstract. The evidence for unseen mass (which is briefly reviewed) suggests that the cosmological density parameter Ω is at least 0.1–0.2. An Einstein-de-Sitter 'flat' universe with Ω = 1—which is appealing for theoretical reasons—can only be reconciled with the data if the galaxies are more 'clumped' than the overall mass distribution, and are poor tracers of the unseen mass even on scales of several Mpc. Possible forms for the unseen mass are discussed; and feedback processes are outlined whereby galaxy formation can be suppressed in underdense regions.

Key words: universe—cosmology—galaxies, distribution—galaxies, formation—unseen mass

1. Evidence for unseen mass

1.1 *Evidence in Different Types of Systems*

Zwicky's (1933) application of the virial theorem to the Coma cluster yielded the first evidence for unseen mass. Modern data on Coma and other clusters bear out the same trend: the overall mass-to-blue-light ratio M/L within the virialized parts of clusters is ~ 300 h (where h denotes the Hubble constant in units of 100 km s^{-1} Mpc^{-1}); but for the stellar content of ordinary luminous galaxies it is 1–10. There is still, however, an uncertainty of order 2 owing to the limited number of galaxy redshifts known in typical clusters. The X-rays from clusters of galaxies reveal that hot gas is present, with temperature such that the sound speed is similar to the virial velocity (Forman & Jones 1982). The amount of gas in the cores is not enough to satisfy the virial theorem—it is comparable with the amount in 'luminous' galactic material. However, gas could provide a bigger fraction of the mass in the outlying parts—because the X-ray emission per unit mass goes as n, diffuse gas is less conspicuous. (The X-ray spectrum shows that this gas has a temperature consistent with being gravitationally bound in a cluster potential whose depth is inferred from the velocity dispersions of the galaxies. This fact argues against 'radical' ideals, such as that the clusters are not virialized, or that the galactic redshifts are not true velocity indicators.)

Historically, *groups of galaxies* were the second type of astronomical systems to reveal a 'mass paradox'. Kahn & Woltjer (1959), studying the motion of Andromeda (M 31) with respect to our Galaxy, concluded that the total mass of the Local Group must be $4 \times 10^{12}\ M_{\odot}$. Several later reanalyses (Einasto & Lynden-Bell 1982; Lynden-Bell 1982 and references cited therein) have essentially confirmed this result: the inferred M/L is 150. Results for other groups are vulnerable to observational errors,

misattributions of group membership, *etc.*, but nonetheless recent studies generally confirm a similar trend (Einasto *et al.* 1975; Huchra & Geller 1982; Geller & Huchra 1983). A statistical method due to Geller & Peebles (1973) does not depend on knowledge of what galaxies belong to what groups. A related 'cosmic virial' analysis (Bean *et al.* 1983) yields a density ~ 0.15 of the critical value (*i.e.* $M/L \simeq 300\,h$).

In principle one might have expected *binary galaxies* to be easier to analyse than the groups, but the results here have been controversial and reveal few clear correlations (White *et al.* 1983 and references cited therein).

There may be a hidden mass problem even closer at hand. The density of visible matter in the solar neighbourhood can be derived by counting all known stellar and gaseous components of the Galaxy (Oort 1932; Joeveer & Einasto 1976; Bahcall 1984). These determinations give, in good mutual agreement, a value $0.1\ M_{\odot}\mathrm{pc}^{-3}$. The total density of matter can be calculated from dynamical considerations, treating stars as test bodies and modelling their spatial distribution and velocity perpendicular to the galactic plane. This comparison has been made many times by various authors, with differing results. Some have found that the dynamical density coincides, within the errors, with the directly-observed density, whereas others (*e.g.* Bahcall 1984) have found a significant discrepancy by a factor of 2. If these latter determinations are correct, our Galaxy must contain an invisible component responsible for the difference in the mass density. Galactic mass models incorporating a dark component have been constructed by Bahcall & Soneira (1980), who find that the invisible component (or at least part of it) must form a disc, with velocity dispersion ~ 40 km s^{-1}.

The unseen mass in galactic discs, though significant for our understanding of galaxy formation, is less important in the cosmic 'mass budget' than the *extensive halo* component. An impressive body of work on the rotation of disc galaxies (see Rubin 1983 for a recent review) shows that rotation curves remain almost flat out to ~ 80 kpc in some cases. This implies local values of M/L exceeding ~ 300 at the periphery, and an extensive dark halo more massive than the disc component.

X-ray observations from the Einstein Observatory (Fabricant, Lecar & Gorenstein 1980; Fabian, Nulsen & Canizares 1984, and references cited therein) indicate that giant galaxies are surrounded by coronae of hot gas. This gas serves as a diagnostic for the gravitational potential, and offers independent support for the presence of unseen mass. The motions of hydrogen clouds and globular clusters offer probes for the gravitational potential in the outlying parts of massive galaxies; these data also suggest the presence of unseen mass. A comprehensive discussion of unseen mass is given by Pebles (1980) and by Dekel, Einasto & Rees (1985).

1.2 *Global Properties of the Unseen Matter*

Direct lower limits on M/L can be obtained for the matter in the outlying parts of galaxies with measured rotation curves, for the haloes of edge-on galaxies (Hegyi & Gerber 1977; Boughn, Saulson & Seldner 1981), and for the matter between the galaxies in (for instance) the Coma cluster (Melnick, White & Hoessel 1977).

There is of course no reason why dark matter should all be the same stuff—conceivably different forms dominate on the scales of individual galaxies, clusters, and superclusters. (Indeed, any component with a disc-like distribution would be likely to have undergone some dissipation; contrariwise, the distinctive feature of most forms of

unseen mass is that they are *less dissipative* than the 'luminous' material.) An important question is the dependence of *M/L* on scale. The available data are summarized in Fig. 1(a), adapted from Faber (1984) and Blumenthal *et al.* (1984), which displays the often-cited trend for *M/L* to increase with scale. However, if we interpret this evidence as implying that there is some strange constituent of the universe (massonium?) a more relevant quality than *M/L* is the ratio of 'ordinary' matter to massonium. The ordinary baryonic matter comprises the stars (which themselves have an *M/L* depending on galaxy type), and also the diffuse gas. The graph of $M_{\rm lum}/M_{\rm tot}$ does not display the same monotonic rise [Fig. 1(b)—see caption for further explanation]; on the contrary, it suggests that on scales $\gtrsim 50$ kpc the ratio $M_{\rm lum}/M_{\rm tot}$ remains steady at a value ~ 0.1. The fact that this quantity is no larger for a rich cluster than for a single galaxy-plus-halo, even though the *M/L* for the cluster is larger, is due to the different stellar populations in the ellipticals in rich clusters, plus the large contribution of X-ray emitting gas to $M_{\rm lum}$. On scales $\gtrsim 2$ Mpc evidence is less clear, because systems larger than this (*e.g.* the local supercluster) cannot be assumed to be virialized.

This evidence, taken at face value, suggests that Ω (defined as the ratio of the actual

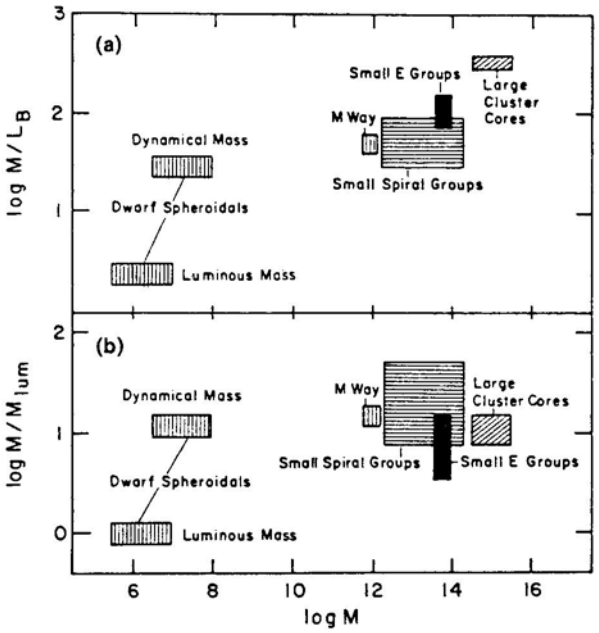


Figure 1. (a) The apparent increase with scale of the mass-to-light-ratio. This increase is due to *two* distinct trends: (i) ordinary stars and gas are a decreasing proportion of the mass of the larger systems; and (ii) in the larger systems, even the 'ordinary' (star + gas) component has a higher *M/L*, because they consist primarily of elliptical galaxies with few young stars, and contain much hot gas revealed only by its X-ray emission, (b) Effect (ii) subtracted out. One finds that the physically more fundamental ratio of 'ordinary' matter to unseen matter is *independent* of scale in all virialized systems larger than galaxies, and has a value consistent with $\Omega \simeq 0.2$.

This figure is adapted from Faber (1984) and Blumenthal *et al.* (1984); fuller details of the data on which it is based can be found in the latter paper. (The issue of unseen mass in dwarf spheroidal galaxies is uncertain and controversial; the diagrams show these systems plotted twice, depending on whether they do or do not contain unseen mass.)

mean density to the critical density $\rho_c = (\frac{8}{3}\pi G t_H^2)^{-1/2}$ is in the range 0.1–0.2, but that only 10 per cent of this ($\Omega = 0.01$–0.02) is definitely baryonic. It poses the insistent questions: what is the unseen mass? However, there is no dynamical evidence for $\Omega = 1$: there are no systems which have $M/L \simeq 2000\,h$, which would be the universal value if $\Omega = 1$. In Section 3 I shall motivate and discuss whether it is indeed possible for the universe to have the critical density—at this stage I merely emphasize that this poses a *second* hidden mass problem, over and above that forced to our attention by well-established virial discrepancies.

2. Possible kinds of unseen mass

2.1 *Baryonic Forms for the Unseen Mass*

Most of the initial baryons might have been incorporated in a population of stars that were either pregalactic, or else formed during the initial collapse phase of a protogalaxy: these stars, or their remnants, could perhaps now have a high M/L and contribute to the unseen mass. Ideally, one would like to be able to calculate what happens when a cloud of 10^5–$10^7\ M_\odot$ condenses out from primordial material: does it form one (or a few) supermassive objects, or does fragmentation proceed efficiently down to low-mass stars? Our poor understanding of what determines the mass spectrum of stars forming now (in, for instance, the Orion nebula) gives us little confidence that we can calculate the nature of stars born in an environment very different from our (present-day) Galaxy.

Although we cannot confidently predict what these so-called 'population III' stars would be like (Kashlinsky & Rees 1983), there are several constraints which, in combination, imply that *if* there are enough of them to provide the unseen mass, the individual masses must *either* be less than $0.1\ M_\odot$ *or else* in the range 10^3–$10^6\ M_\odot$. Masses above $\sim 0.1\ M_\odot$ would contribute too much background light unless they had all evolved and died, leaving dark remnants. But the remnants of ordinary massive stars of 10–100 $M_\odot$ would produce too much material in the form of heavy elements. Limits on the range 100–1000 $M_\odot$ are uncertain because only helium may be ejected, the 'heavies' in the core collapsing into a black hole remnant. An uncertainty in the evolution of massive or supermassive stars is the amount of loss during H-burning; however the hypothesis that most mass goes into very massive objects (VMOs) of greater than about $10^3\ M_\odot$ is compatible with the nucleosynthesis constraints. A further consideration favouring these *high* masses is that VMOs are likely to terminate their evolution by a collapse which swallows most of the mass: if most of the material were ejected, 'recycling' through several generations would be necessary in order to end up with most of the material in black holes rather than gas. Detailed discussions of pregalactic stars are given by Carr, Bond & Arnett (1984); see also Tarbet & Rowan-Robinson (1982).

Black holes of $\geqslant 10^6\ M_\odot$ are excluded, at least within the haloes of individual galaxies, because dynamical relaxation processes could increase the random motions of stars in disc galaxies, and make the discs thicker than they are observed to be.

We do not yet understand star formation well enough to decide theoretically between these possibilities, but there is real hope of observational progress by searching for manifestations of gravitational lensing. The angular separation θ_1 of lens images is a

diagnostic of the masses involved: for a path length of order the Hubble radius, $\theta_1 \sim 10^{-6}\ (M_1/M_\odot)^{½}$ arcsec, where M_1 is the lens mass. For $M_1, \gtrsim 10^6\ M_\odot$, very long baseline radio interferometers provide adequate resolution. For $M \lesssim 0.1\ M_\odot$ ('Jupiters') the scale is less than a micro-arcsec. This cannot be resolved by any techniques, until we have optical interferometers in space. However, there is a genuine prospect of detecting lensing of this kind because of the variability that would occur if the lens were to move transversely. (It takes only a few years for an object at the Hubble radius moving at $\sim 10^3$ km s^{-1} to trasverse an angle 10^{-6} arcsec.) The possibility of observing such an effect along a given line of sight is not small: it is of order Ω_1, the fraction of the critical density contributed by the lensing objects. (This probability is independent of M_1 because the cross section per object scales with M_1, whereas the number of objects scales as M_1^{-1} for a given Ω_1)

2.2 *Non-baryonic Unseen Mass*?

If neutrinos have negligible rest mass, the present density expected for relic neutrinos from the big bang is $n_\nu = 110\ (T_\gamma/2.7\ \mathrm{K})^3\ \mathrm{cm}^{-3}$ for each two-component species. This is of order the photon density n_γ, differing just by a factor 3/11 (*i.e.* a factor 3/4 because neutrinos are fermions rather than bosons, multiplied by 4/11, the factor by which the neutrinos are diluted when e^+-e^- annihilation boosts the photon density). This conclusion holds for *non-zero* masses, provided that $m_\nu c^2$ is far below the thermal energy (~ 5 MeV) at which neutrinos decoupled from other species and that the neutrinos are stable for the Hubble time. Comparison with the baryon density, related to Ω *via* $n_\mathrm{b} = 1.5 \times 10^{-5}\ \Omega_\mathrm{b}\ h^2\ \mathrm{cm}^{-3}$, shows that neutrinos outnumber baryons by such a big factor that they can be dynamically dominant over baryons even if their masses are only a few electron volts. In fact, a single species of neutrino would yield a contribution to Ω of $\Omega_\nu = 0.01\ h^{-2}\ (m_\nu)_\mathrm{eV}$, so if $h = 0.5$, only 25 eV is sufficient to provide the critical density.

The entire range 100 h^2 eV–3 GeV is incompatible with the hot big-bang model (Gunn *et al.* 1978). (For $m_\nu c^2 > 3$ GeV, the rest mass term in the Boltzmann factor would kill off most of the neutrinos before they decoupled; the number surviving would be less than $\sim n_\mathrm{b}$.) If any species of neutrino were discovered to have a mass in this excluded range, it would show that one cannot extrapolate the hot big bang back to $kT \gtrsim 5$ MeV, and that most of the photons must have been generated at later times.

(Such arguments are familiarly expressed by saying that the hypothetical particles of nonzero rest mass would 'close the universe by a large factor'. This loose phrase is in fact rather misleading. The geometry of the universe is likely to have been laid down at very early stages by mechanisms that do not 'know' what the dominant constituents will be after 10^{10} years. If the universe were indeed 'flat' (*e.g.* for 'inflationary' reasons as discussed in Section 3), it would expand with $\Omega = 1$ [*i.e.* with a density of $(\frac{8}{3}\pi Gt^2)^{-1}$] for all t. If the neutrinos were in the excluded mass range—or if there were, for instance, too many primordial monopoles—the observational incompatibility would be that Ω_b, the baryonic fraction of the total density would, for $t \simeq 10^{10}$ years, be much less than 10^{-2}.)

Neutrinos of nonzero mass would be dynamically important not only for the expanding universe as a whole but also for large bound systems such as clusters of galaxies. This is because they would now be moving slowly: if the universe had cooled

homogeneously, primordial neutrinos would now be moving at around 200 $(m_\nu)_{\rm eV}^{-1}$ km s^{-1}. They would be influenced even by the weak ($\sim 10^{-5}\ c^2$) gravitational potential fluctuations of galaxies and clusters. If the three (or more) types of neutrinos have different masses, then the heaviest will obviously be gravitationally dominant, since the numbers of each species should be the same.

It was conjectured more than a decade ago (Cowsik & McClelland 1973; Marx & Szalay 1972) that neutrinos could provide the 'unseen' mass in galactic haloes and clusters. In recent years, astrophysicists have explored this possibility in some detail, and considered scenarios for galaxy formation in which neutrino clustering and diffusion play a key role. Despite its attractiveness in many ways (Doroshkevich, Shandarin & Zel'dovich 1983), the neutrino-dominated picture of galaxy formation has serious weaknesses. The main constraint is that the smallest scale to survive damping due to free-streaming is very large—comparable with a supercluster (see Fig. 2). Studies of nonlinear clustering (on scales $\lesssim 10h^{-1}$ Mpc) show that supercluster collapse must have occurred quite recently, at redshift $z < 2$ (Frenk, White & Davis

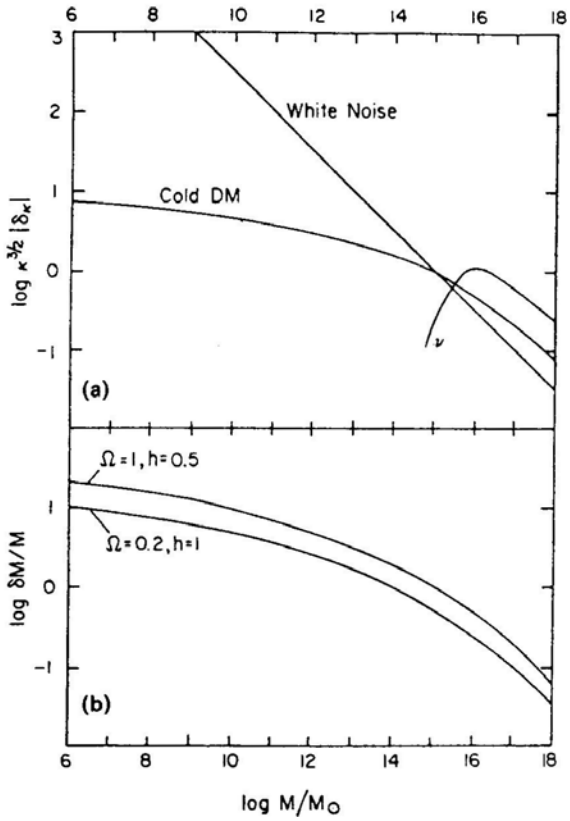


Figure 2. (a) The density fluctuation spectrum ($\kappa^{3/2}\ |\delta_x| = \delta\rho/\rho(M)$, where $M = 4\pi^4\rho_0/3\kappa^3$) for isothermal white-noise fluctuations, and for adiabatic 'constant-curvature' fluctuations in universes dominated by neutrinos and 'cold dark matter' respectively. Note that the neutrino fluctuations are damped out by free-streaming on scales below superclusters. (b) The root-mean-square fluctuations within a randomly placed sphere containing mass M for cold dark matter for two specific models calculated by Blumenthal & Primack: $\boldsymbol{\Omega} = 1$, $h = 0.5$ and $\boldsymbol{\Omega} = 0.2$, $h = 1$ (from Blumenthal *et al.* 1984).

1983). However, the best limits on galaxy ages coming from globular clusters and other stellar populations, plus the possible association of QSOs with galactic nuclei, indicate at least some galaxies had formed, and evolved to the extent that they have well-defined nuclei, before $z = 3$. This seems inconsistent with the neutrino 'top-down' theory, in which superclusters form before galaxies rather than after them.

Another problem with the neutrino picture is that large clusters of galaxies can accrete neutrinos more efficiently than ordinary galactic haloes, which have lower escape velocities. One-dimensional numerical simulations (Bond, Szalay & White 1983) predict that the ratio of total to baryonic mass M/M_b should be ~ 5 times larger for clusters ($\sim 10^{14} M_\odot$) than for ordinary galaxies ($\sim 10^{12} M_\odot$). (*cf.* Fig. 1b).

Particle physicists have other particles 'in reserve' which could make a substantial (non-baryonic) contribution to Ω, but which differ from neutrinos in that their free-streaming velocity is negligible, so that small-scale adiabatic perturbations are not phase-mixed away. Such particles can be described as 'cold dark matter', in contrast to neutrinos whose free streaming velocity renders them 'hot'. A currently popular candidate is the axion (Ipser & Sikivie 1983), a pseudoscalar field proposed originally to avoid large CP violation in the strong interactions (which would imply, for example, much too large a value for the neutron electric dipole moment). Instanton effects generate a nonzero axion mass at the quark deconfinement temperature ($kT \simeq 10^2$ MeV), below which the axions act as a nonrelativistic, massive, pressureless fluid. The requirement that the axion density be less than the critical density implies that the axion mass $m_a \gtrsim 10^{-5}$ eV, (Preskill; Wise & Wilczek 1983; Abbott & Sikivie 1983, Dine & Fischler 1983), while the longevity of helium-burning stars implies that $m_a < 10^{-1}$ eV (Fukugita, Watamura & Yoshimura 1982). Thus, if axions exist, they may be cosmologically important, and, for $m_a \simeq 10^{-5}$ eV, they would be gravitationally dominant.

Another candidate particle for cold dark matter is the photino, the spin-½ supersymmetric partner of the photon. Photinos are thought to be the lightest supersymmetric particles with $m \gtrsim 0.5$ GeV (the lower limit corresponding to cosmological critical density). Since photino annihilation at high temperatures is incomplete, the remnant photinos can, because of their large mass, contribute a critical density today.

Small primordial black holes offer another option. Still another idea, recently proposed by Witten (1984), involves 'nuggets' of quark matter which survive stably from a quark-hadron phase transition.

There is no shortage of 'cold dark matter' candidate particles—although each of them is highly speculative, to say the least. The motivation for nonetheless considering the hypothesis that the universe is dominated by cold dark matter is that it leads to a cosmogonic scheme that avoids the difficulties of the neutrino-dominated scheme and correctly predicts many of the observed properties of galaxies, including their range of masses, irrespective of the identity of the cold particle (Peebles 1984; Blumenthal *et al.* 1984).

A baryon-dominated universe ($\Omega = \Omega_b$) is consistent with the deuterium limit for Ω and h near their observational lower limits (see Section 2.3 below). There are, however, independent reasons for preferring the dark matter to be non-baryonic. In particular, the process of galaxy formation is then easier to understand. The existence of galaxies and clusters today requires that perturbations in the density must have become nonlinear before the present epoch. In a baryonic universe, for adiabatic perturbations at recombination, this implies present-day fluctuations in the microwave background

an order of magnitude larger than the present observational upper limits of $\sim 2 \times 10^{-5}$ on scales of 2 arcmin (Uson & Wilkinson 1984). For baryonic dark matter, this problem can be avoided only if there is significant reheating of the intergalactic medium after recombination or if the primordial fluctuation spectrum is isothermal rather than adiabatic. (However, grand unified models of baryosynthesis favour adiabatic fluctuations.) For nonbaryonic dark matter, on the other hand, the predicted fluctuations in the microwave background are consistent with the observations—though only if $\Omega h^{4/3} > 0.2$ (Bond & Efstathiou 1984)—since the baryonic fluctuations are small at recombination and only later grow to the same size as fluctuations in the dark matter.

2.3 *Primordial Nucleosynthesis: Need for some Baryonic Dark Matter if* $h = 0.5$

Primordial nucleosynthesis depends on two things: the expansion timescale at 0.1–1 MeV and the baryon density at that same epoch (which is proportional to $\Omega_b h^2$). The predicted ^{4}He abundance is rather insensitive to the matter density: for $\Omega_b h^2) \gtrsim 10^{-2}$ the density of baryons is high enough to ensure that most of the neutrons that survive when the neutron-proton ratio 'freezes out' at $kT \simeq 1$ MeV get incorporated in ^{4}He.

The cosmic helium abundance can however be measured with sufficient precision to suggest that the primordial ^{4}He is less than 26 per cent at the 3 σ level (Pagel 1982). This is compatible with $\Omega_b h^2) \lesssim 0.1$ but not with $\Omega_b h^2 = 1$ (for $\geqslant 3$ species of neutrinos). The strongest constraint on Ω_b from primordial nucleosynthesis comes, however, not from He but from deuterium. This is an intermediate product in helium formation, the amount emerging from the big bang being a steeply decreasing function of Ω_b. Only if $\Omega_b h^2) \lesssim 0.035$ can the observed deuterium abundance be produced in a standard hot big bang. The strength of this constraint stems from the failure of astrophysicists in the last decade to suggest any other plausible way of making deuterium.

The combined arguments from primordial nucleosynthesis suggest that $\Omega_b h^2$ is in the range 0.01–0.035, permitting a value of Ω_b no higher than 0.14, even for $h = 0.5$. See Yang *et al.* (1984) for a recent summary. If the lepton number for ν_1 and $\bar{\nu}_1$, were nonzero, then the neutron–proton equilibrium ratio would be shifted, affecting He production. It is thereby possible in principle to accommodate a higher Ω_b (David & Reeves 1980). (However, in order to make much difference, the neutrino lepton number must be of order the photon number; that is, $\gtrsim 10^8$ times larger than the baryon number.) Other possible 'escape clauses' can be invoked—for instance, there might be large-amplitude inhomogeneities in the initial baryon distribution, such that all the baryonic material we can now sample comes from *under*dense regions, the overdense regions having turned into dark population III objects (Rees 1983). Because the relevant parameter in primordial nucleosynthesis is, $n_b/n_\gamma \propto \Omega_b h^2$, more precise comparison of models with observation must await a firmer value of the Hubble constant. If $h = 1$ (corresponding to a Hubble time of 10^{10} years) then the simplest inference would be that most of the unseen mass—both in the haloes of individual galaxies and in clusters and groups—was nonbaryonic; but if $h = ½$ (corresponding to a Hubble time of 2×10^{10} years) the lower limit to h set by the requirement not to overproduce D + ^{3}He implies that some unseen matter—maybe that in haloes, if not in intergalactic space—is baryonic, though only enough to contribute $\Omega_b \simeq 0.1$. It is

remarkable that the simplest and least arbitrary form of hot big-bang model can, for a suitable choice of n_b/n_γ, account for D, ^{4}He and ^{7}Li (Yang *et al.* 1984).

3. Fine tuning, or Ω = 1?

If the universe had resembled a Friedmann model since the Planck time t_P, a comoving scale initially equal to the Planck length would by now (when $t \simeq t_H \simeq 10^{60}\ t_p$) have grown by ~ 10^{30}, because the scale factor varies as $t^{1/2}$ for a radiation-dominated expression. But the present Hubble radius is ~ 10^{60} Planck lengths. The Robertson-Walker curvature radius, which is certainly not much smaller than the present Hubble radius, is thus ~ 10^{30} larger than the 'natural' scale if the universe were 'set up' at a time of order t_P. This problem—the so-called 'flatness problem'—introduces a large dimensionless number,

$$\mathcal{R} = \left(\frac{\text{Robertson–Walker curvature radius}}{\text{Planck scale expanded to present epoch } t_H}\right) = \left(\frac{t_H}{t_P}\right)^{\frac{1}{2}} \left(\frac{t_{eq}}{t_H}\right)^{\frac{1}{6}} \frac{1}{|(\Omega-1)|^{\frac{1}{2}}}. \qquad (1)$$

The extra factor $(t_{eq}/t_H)^{1/6}$, where t_{eq} is the time at which the expansion switches from being radiation-dominated to being matter-dominated, arises because for $t > t_{eq}$ the scale factor grows as $t^{2/3}$ rather than as $t^{1/2}$.

The value of $\mathcal{R}$ is thus at least ~ 10^{30}; the vastness of this number seems to imply some kind of fine tuning in the initial conditions.

One reaction to this requirement might be an anthropic one: the universe would not have offered a hospitable environment for galaxies and stars if $\mathcal{R}$ were not so huge. But it would be more satisfactory if there were some fundamental reason. Such reasons are offered by new ideas in inflationary and quantum cosmology. Such ideas indeed suggest that the comoving scale of the Robertson-Walker curvature radius could readily have inflated by *far more than* 30 powers of ten, yielding a value of Ω indistinguishable from unity: to have anything else demands a coincidence in that it requires the curvature scale to be just of order the present Hubble distance. Inflation must increase the comoving scale by at least $\mathcal{R} \simeq e^{65}$ (see Equation 1) in order to explain the observed largescale homogeneity (the 'horizon problem'). If it involved an improbable effort for the universe to 'heave itself up' to this size, one could perhaps invoke anthropic arguments that the Robertson-Walker curvature scale would be no larger than our present Hubble radius. But the inflation is so readily achieved that it would probably overshoot by many powers of *e*, yielding a present Ω indistinguishable from unity. [In any case, if the universe had inflated by just the amount needed to yield Ω of (say) 0.1, it would be unnatural to find such small amplitude anisotropies in the microwave background—this would require the pre-inflation geometry of our observable universe to have been rather special, with a curvature which was essentially constant (to 1 part in ~ 10^4) over scales of order the curvature radius.] I shall not comment further on these ideas: they merely provide a motivation for seeing if Ω can indeed be unity—or if, contrariwise, there *has* to be some fine tuning that makes the present epoch special.

The evidence for hidden mass which I summarized in Section 1—the evidence that *M/L* increases as we consider scales from 10 kpc out to a few Mpc, is incontrovertible in general terms, even though particular details are all subject to debate. However, the amount of mass reliably inferred by dynamical arguments, though maybe 10 times

more than the amount of luminous material (in ordinary stars, or in gas), still falls short by a factor at least ~ 5 of yielding the critical density (see Fig. 1).

The problem of hidden mass, central to all studies of galaxies and clusters, acquires a new dimension: the possibility that the universe has exactly the critical density is now of renewed interest. Whereas, until recently, astronomers would have been disposed to accept an Ω of 0.1 or 0.2 as the 'best' (though still uncertain) estimate, the growing attitude that anything different from unity is in some sense unnatural provides an extra motivation for investigating whether there can be enough unseen matter to contribute all of this critical density—in other words, is the $\Omega = 1$ postulate compatible with what we observe?

I shall briefly mention three possibilities, but there is one feature common to them all: if $\Omega = 1$ the hidden mass must be *more smoothly distributed*, on large scales, than are the galaxies (or, at least, the conspicuous galaxies included in the surveys).

(i) The primordial (baryonic) material from which galaxies formed may have become more clumped than the (non-dissipative) unseen mass which is gravitationally dominant.

(ii) The process that converts some assemblages of baryons into (luminous) galaxies may proceed with an efficiency that is in some way sensitive to environmental or feedback effects, the consequence being that the galaxies are poor tracers even of the baryons. For instance, one might imagine a threshold density above which galaxy formation is efficient, and below which it is impossible. There could then be huge contrasts in the galaxy density, even if the overall density or gravitating stuff were only slightly non-uniform.

(iii) The overall dynamics are dominated by some kind of non-interacting relativistic material.

If $\Omega = 1$, the mean M/L for cosmic matter must be $\sim 2000\,h$, so the rising trend in Fig. 1(a) must continue out to length scales $\gtrsim 10\ h^{-1}$ Mpc. Direct evidence on these scales is limited, basically because the relevant systems do not have a large enough density contrast to be virialized. Models for our infall into the local supercluster, based on reasoning along the lines of Kahn & Woltjer's old (1959) local group argument, apparently suggest $\Omega < 1$ (see Davis & Peebles 1983 for a review), and are compatible with $\Omega = 1$ only if the mass is less concentrated towards the Virgo cluster than the distribution of galaxies indicates. But if one abandons the assumption that galaxies trace the mass distribution, then any value of Ω is permitted (Hoffman & Salpeter 1982). This is not surprising, because the peculiar velocity induced by the supercluster essentially depends on the *excess mass* in it, which is proportional to Ω times the fractional enhancement in the *total* density. Analogously, statistical arguments based on the cosmic virial theorem (Bean *et al.* 1983), which yield $\Omega \simeq 0.15$ [and are based on the relative velocities of galaxies separated by (1–2) h^{-1} Mpc], could be reconciled with $\Omega = 1$ if the dominant form of mass did not participate fully in the apparent clustering of galaxies.

3.1 *Could Baryonic Material Yield* $\Omega = 1$?

The largest amount of reliably inferred gas is that which produces the thermal X-ray emission from clusters of galaxies. The quantity in the core of (*e.g.*) the Coma cluster is not enough to bind it: it is comparable with the amount of luminous matter in the

galaxies themselves. However, it is harder (and more model-dependent) to infer how much still more diffuse gas could pervade the outlying parts of the cluster.

Gas that gets heated to $kT \geqslant 10$ keV would not be confined in cluster potential wells, but would constitute a roughly homogeneous intercluster medium. It has been suggested that such a gas is the prime contributor to the X-ray background at > 10 keV; even if it is not, the observed strength of this background constrains the density and thermal history of any such gas. More gas can be 'hidden' in intergalactic space if it is ultra-hot (~ 40 keV) rather than at a lower temperature. Indeed, there could be almost enough such gas to provide the critical density if it were heated up at a redshift $z < 3$. The difficulties with this idea stem from the very large energy input and special thermal history necessary to avoid conflict with various observational constraints (Fabian & Kembhavi 1982).

Although there is no natural reason why compact objects (*e.g.* black holes) should not participate in clustering, it is worth noting that the firmest constraint on intergalactic black holes comes from the absence of evident gravitational lensing (Press & Gunn 1973, Canizares 1982). Masses $\geqslant 10^6 M_\odot$ (which are of course not constrained by dynamical friction arguments if they lie outside galaxies) can be ruled out as important contributors to Ω, because the resultant lensing, on angular scales susceptible to VLBI measurements, would have been detected; the situation is less clear for compact bodies of smaller mass.

Note that if the universe were closed by hot gas, or by black holes which were not primordial, but formed by astrophysical processes from baryonic material, the cosmic abundances of the light elements would require some unorthodox explanations (see Section 2.3).

3.2 *Biased Galaxy Formation*

Let us consider how the distribution of galaxies could give an exaggerated impression of the overall density contrasts on large scales—how, in particular, large 'voids' might be manifest in the galaxy distribution, without necessarily implying that the regions are completely empty.

One possibility is that the baryons have been evacuated from large regions of space, without the depletion being equally severe for nonbaryonic material. If these effects are to be relevant to the largescale voids and clustering observed now, across which the differential Hubble speed is $\gtrsim 3000$ km s^{-1}, then the gas must be pushed with at least this speed. However, to push all the baryons with the requisite speed would require $\gtrsim 10^5$ eV per particle, and still more than this if radiative cooling behind the shock fronts dissipates much of the energy. Hogan (1984) points out that, if large volumes of gas were raised to such high temperatures that pressure gradients could drive the requisite motions, the Sunyaev-Zeldovich effect might lead to smallscale anisotropies in the microwave background exceeding the measured limits. But, despite this constraint, there may be no insurmountable objection to this general possibility; it is conceivable that the configurations we now see could be determined by gas-dynamical rather than primarily gravitational effects.

Maybe the voids are not deficient even in baryonic matter, but are domains where the formation of luminous galaxies from these baryons has been somehow inhibited. This is indeed a natural consequence of the 'neutrino pancake' model (Doroshkevich,

Shandarin & Zel'dovich 1983): no primordial perturbations survive on scales as small as individual galaxies (see Fig. 2), which form only from secondary perturbations produced in shocked and compressed regions: the underdense gas acquires no inhomogeneity on galactic scales; and some of it may, moreover, be heated by shocks so that its Jeans mass exceeds that of a galaxy.

In cosmogonic schemes involving cold dark matter, there is no equally obvious process that might inhibit galaxy formation in the incipient voids; but it is interesting to consider this possibility a bit further, because the effective thermal energy required to provide enough pressure to raise the Jeans mass above a galactic mass is 'only' a few hundred eV per baryon, compared with the $\gtrsim 10^5$ eV per baryon of kinetic energy needed to evacuate a void. If this kind of energy input were provided by some bound systems—galaxies or their precursors—the first such systems to form could exert an important feedback on what happens next. It may then be only the high amplitude (2σ or 3σ) peaks in the initial fluctuation spectrum which evolve straightforwardly into the non-linear stage: the collapse of more 'typical' fluctuations, which would not turn around as early as those on the tail of the gaussian distribution, may be pre-empted by feedback effects.

If such a 'feedback' process is to inhibit galaxy formation in a large void, its influence must propagate fast enough to traverse the void. It is therefore interesting to explore ways in which the effective sound speed could be raised to 100 km s^{-1} by a process whose influence propagates faster than this (*i.e.* at $\gtrsim$ 3000 km s^{-1}). Agents that can transmit influences at high speed include photons and relativistic particles.

The amount of energy needed is not exorbitant; however there are physical problems of coupling it to the gas in a suitable way. For instance, the most obvious possibility is photoionization. The temperature attained depends on the mean energy of an ejected photoelectron, and in a typical HII region is never more than a few times 10^4 K. At first sight, one might suppose that a source spectrum peaking sharply at photon energies 100 eV (*e.g.* a blackbody spectrum with $T \gtrsim 10^6$ K) would yield a correspondingly hot HII region. But this is not so (except in the implausible case of very sudden ionization) because a photoelectron would then collisionally ionize several neutral atoms before they had a chance to be photoionized, the energy of each photoelectron consequently being shared among several electrons and ions (Bardeen 1984, personal communication).

Violent activity and Supernovae generate cosmic ray (suprathermal) particles. The speeds of individual particles may be ~ c, and their energy density, if they diffused uniformly through the universe, could well exceed 100 eV per baryon. Subrelativistic particles would be slowed down, and would transmit their energy to the thermal component. However, the relativistic particles could themselves exert a pressure if they were coupled (*e.g. via* magnetic fields) so that they constituted, with the thermal gas, a composite fluid, to which they contributed most of the pressure. Although there is here even less problem in fulfilling the energy density requirement than there is for ultraviolet radiation, there is uncertainty about how uniformly it can spread. If the cosmic-ray energy remains concentrated around the sources, it is irrelevant in the present context; at the other extreme, if the particles diffuse too freely, they do not couple well enough to protogalactic gas for their pressure gradients to oppose gravitational collapse.

What about radiation pressure? The microwave background is of course the dominant energy density. However, it is too weakly coupled (*via* Thomson scattering)

to exert any effective damping on non-Hubble flows at epochs $z \lesssim 100$. A somewhat more hopeful possibility is the pressure of Lyman line radiation. If the protogalactic gas is not completely ionized, most of the cosmic background emitted in the ultraviolet would have been transformed into Lyman alpha, whose energy density could exceed 100 eV per ion. The cross-section is certainly large enough to ensure that radiation has a short mean free path. However, in a collapsing (or expanding) medium, the photons would be shifted out of the line wings after the density has changed by a fraction ($\sim 3\Delta v/v$). Even though the line width may be as much as 50 Å, this means that the effectively trapped energy density at any instant would be only a tenth of the total radiation background density in the ultraviolet.

Either of the above effects could inhibit collapse of a protogalaxy if it produced enough energy (per baryon) to balance the gravitational binding energy. We note that the same inhibition effect would result if a largescale gradient in the cosmic ray or radiation pressure pushed the baryonic fluid at ~ l00 km s^{-1} relative to the dark matter: it would not then get captured by the potential wells which develop from inhomogeneities in the dark matter distribution.

Any effect of the above kind would imply that there were many haloes (*i.e.* gravitationally bound galactic mass systems composed of dark matter) in which no baryonic matter had condensed and cooled. Such systems might be responsible for those instances of gravitational lensing where a 'luminous' candidate for the lens is puzzlingly absent.

3.3 *Environmental Effects at Pregalactic Epochs*

Even if gaseous bound systems of galactic mass can form, the types of galaxy they develop into may depend sensitively on the composition of the gas—in particular, whether it contains heavy elements, which permit more efficient cooling (and whether molecular hydrogen, an important coolant at $T < 10^4$K, can form). Such factors control the stellar initial mass function (IMF), and the efficiency of star formation at large radii; they thereby determine the brightness profile (and hence the observability) of a galaxy. If, for instance, star formation went to completion before the baryons had become concentrated (via dissipative infall) in the central parts of the halo potential, a galaxy might not show up in the usual surveys. It is therefore interesting to ask whether pregalactic processes could have led to inhomogeneities in environmental factors (*e.g.* the heavy element or H_2 abundances) on very large scales: if so, galaxy formation would proceed in a 'patchy' way.

At first sight, it might seem unlikely that any *large*scale correlations could be produced at the (earlier) time when subgalactic systems are condensing out, and when galactic and cluster scale perturbations are still essentially in the linear regime. However, there are two reasons why such effects cannot be dismissed.

(i) Even the rare high-amplitude peaks in the fluctuation spectrum (3σ or even 4σ above the mean) can cause significant feedback effects, and these high amplitude peaks are always more strongly clustered than the typical peaks.

(ii) The fluctuation spectrum appropriate to the currently popular 'cold dark matter' model is flat at small masses, so that the amplitude is not vastly larger on scales $10^6\ M_\odot$ than on galactic mass scales. This accentuates effect (i).

In the most straightforward version of the cold-dark matter picture (Peebles 1984;

Blumenthal *et al.* 1984) the first baryonic systems to go non-linear will have masses 10^5–$10^6\ M_\odot$. The 3σ peaks on this scale turn around at z in the range 50–100. The main uncertainty is in the types of stars that form in the first such systems that fragment. If they were all low-mass 'Jupiters', they would have no significant non-gravitational effects on the subsequent course of galaxy and cluster formation. On the other hand, pregalactic high-mass stars or VMOs could have important effects.

Even if only $\sim 10^{-4}$ of the primordial matter went into massive stars, enough processed material could be ejected to contaminate a large fraction of the universe with a Population II abundance of heavy elements. The extra cooling due to heavy elements would certainly affect the complex processes of agglomeration and fragmentation by which galaxies subsequently form.

Ultraviolet radiation from massive stars or VMOs would photoionize uncondensed gas (Hartquist & Cameron 1977). The details of this process depend on the redshift at which it occurs, and on how long-lived the UV sources are (see Carr, Bond & Arnett 1984 for a fuller discussion), but in general only 10^{-4}–10^{-5} of the mass, if converted into objects radiating at near the Eddington limit, would suffice to ionize the entire universe. The intergalactic medium would have cooled to $T \lesssim 100$ K by the time ($z \simeq 50$) when the first bound systems formed. If the gas were heated to $\sim 10^4$ K, the Jeans mass ($\propto T^{3/2}$) would rise by $\sim 10^3$. The actual factor may not be as drastic as this, because the photoionized gas may cool again due to Compton and molecular cooling (Couchman 1984). Nevertheless, the gas would certainly end up on a higher adiabat than before the heating occurred. Consequently the photoionization would terminate (or at least exert a negative feedback on) the formation of systems of the kind which are expected to form first in the 'cold dark matter' cosmogony: there would then be a pause in the cosmogonic process until larger systems (exceeding the new Jeans mass) went non-linear.

These considerations suggest that feedback processes may modify the primordial state of the gas (and change the scale or character of bound systems forming later) after a fraction $f \lesssim 10^{-4}$ condensed out—if there were a (gaussian?) spread in $(\delta M/M)$, these would be the rare ($> 3\sigma$) high-amplitude peaks on a mass scale $M_1 \simeq 10^6\ M_\odot$. Doroshkevich, Zel'dovich and Novikov (1967) were the first to discuss an effect of this kind; they argued that the feedback process would generate a natural 'secondary' scale $\sim f^{-1}\ M_1$, which they related to galaxies. They implicitly assumed that first systems to turn around would be randomly distributed. This is probably not a bad assumption for a spectrum where $\Delta(M) = \langle(\delta M/M)^2\rangle^{\frac{1}{2}}$ decreases steeply with M. However, it is not difficult to see that for a flat spectrum (as expected for cold dark matter) the first bound systems would tend to occur in clumps. This is because the long wave Fourier components are not negligible in amplitude, and there is a significantly higher chance of getting large over-densities (on mass-scales of, say $10^6\ M_\odot$) if these are riding on a peak (rather than being in a trough) of a wave mode with scale $\sim 10^{10}\ M_\odot$. This effect is still more marked if we are concerned with very rare fluctuations on the steeply falling high-amplitude tail of the gaussian distribution. ($f \lesssim 10^{-4}$ corresponds to $\gtrsim 3.72\sigma$).

Let me give a—purely illustrative—numerical example of the resultant effects. Suppose that the first bound systems, of mass $M_1 \simeq 10^6\ M_\odot$, fragment into massive stars, which produce UV radiation and heavy elements, and that all the remaining gas can be photoionized when a fraction $f = 10^{-4}$ has condensed into such systems (*i.e.* $f = 10^{-4}$ is enough to turn the universe into an H II region). Suppose further that the (more massive) systems which form from the reheated gas evolve differently, and in

a manner which depends on whether they have been contaminated with heavy elements.

We can then ask: what is the largest scale M_2 on which the subsequent evolution will be inhomogeneous, owing to clumping of the first $M \simeq M_1$ ($\sim 10^6\ M_\odot$) bound systems, and consequent inhomogeneous deposition of the first heavy elements? Let us calculate what difference it makes if these systems lie within a 2σ peak (protocluster) rather than a 2σ trough (protovoid) on a much larger mass-scale M_2. The density contrast between these regions is equivalent to a difference of $4[\Delta(M_2)/\Delta(M_1)]$ standard deviations in the amplitude on scale M_1. Thus a 3.72σ peak on scale M_1 (corresponding to $f = 10^{-4}$) in the protocluster has as great an overdensity as a $\{3.72 + 4[\Delta(M_2)/(\Delta(M_1)]\}\sigma$ peak in the protovoid: the first bound systems and the production of heavy elements, occur preferentially in the protoclusters. A value $f = 5\times10^{-5}$ corresponds to 3.89σ. So, even on mass-scales M_2 such that $\Delta(M_2)/\Delta(M_1)$ is as small as $0.25 \times 0.17 \simeq 0.042$, there would be a factor 2 difference between the heavy element abundance produced before photoionization choked off the process. For the 'cold dark matter' spectrum in Fig. 2, we find that the value of M_2 is $\sim10^{15}\,M_\odot$.

This example shows how pregalactic processes at $z \simeq 50$, involving gravitationally bound systems whose mass is only $M_1 \simeq 10^6\ M_\odot$, could be responsible for an important 'environmental' parameter, the pregalactic heavy element abundance; and that this abundance could vary by a factor 2 on scales $10^{15}\ M_\odot$—scales which would still not have turned around even at the present epoch. If the process of galaxy formation (or the stellar IMF in galaxies) were sensitive to this abundance, and consequently proceeded differently in protoclusters and in 'voids', then the contrasts in the density of bright galaxies could be greater than the overall density contrast in baryons. This possibility (or analogous ones based on different feedback processes) would be important if we were in a high density ($\Omega = 1$) universe, as it would help to reconcile the low non-Hubble velocities of galaxies with the largescale non-uniformities' in their distribution.

3.4 *Other Options*

A more radical and speculative viewpoint is that sufficient mass-energy to yield $\Omega = 1$ pervades the universe uniformly, and is qualitatively quite different from the well-established unseen mass discussed in Section 2. Three such ideas deserve a mention. All, however, are severely constrained by the fact that galaxies have indeed formed: gravitational instability of a baryonic or 'cold' component is suppressed if this material is embedded in hot material whose higher overall density determines the expansion timescale for the universe.

(a) The critical density could be contributed by non-interacting relativistic particles that resulted from decay of unstable heavy particles (see, for instance, Wilczek 1982). If the decays occurred relatively recently, *e.g.* at $t \simeq 10^9$ yr, galaxy and cluster formation could have taken place beforehand, when the universe was dominated by heavy cold particles. One problem with this idea is. that the universe would have expanded according to an $R \propto t^{1/2}$ law for most of its history; the time since the big bang is then little more than half the Hubble time, aggravating the 'age problem' unless $h < 0.5$.

(b) Among the exotic kinds of unseen mass that particle physicists now envisage are 'strings'—topologically stable structures which may appear in the vacuum as a result of

inhomogeneous breaking of a Yang-Mills symmetry in the ultra-early universe. Strings can have many interesting astrophysical effects—*e.g.* by triggering fluctuations, lens effects and a gravitational wave background—even if they contribute only $\lesssim 10^{-4}$ to Ω (Vilenkin 1984; Hogan & Rees 1984 and references cited therein). They could conceivably contribute $\Omega = 1$, but under very special (and maybe contrived) conditions: there would need to be a network of almost straight strings spanning the universe, and stretching as it expands (with $R \propto t$). Only under these conditions could strings contribute a high background density without producing unacceptably large Newtonian gravitational effects in galaxies and clusters.

(c) The universe could be flat, even though the. ordinary mass-energy density yielded $\Omega < 1$, if a non-zero cosmological constant Λ contributed to the curvature. Such a possibility could be checked by classical cosmology—measurements of the deceleration parameter, *etc.* In some respects this idea resembles alternative (a) above; however, the Λ-term is unimportant at early epochs, and so would not have had such a serious inhibiting effect on galaxy formation. This option requires that Λ should be comparable to ordinary matter in dynamical importance at the present epoch—an unappealing idea, since it reinstates the necessity for a degree of fine tuning which the $\Omega = 1$ hypothesis is intended to avoid.

It is fair to note, however, that almost all theories which invoke non-baryonic matter require some level of coincidence in order that the luminous and unseen mass contribute comparable densities (to within one or two powers often). For instance, in a neutrino-dominated universe, $(m_\nu/m_{\rm proton})$ must be within a factor ~ 10 of $n_{\rm b}/n_\gamma$. The only model that seems to evade this requirement is Witten's (1984) idea that the quark-hadron phase transition may leave comparable amounts of material in 'ordinary' baryons and in 'nuggets' of exotic matter.

4. Concluding comments

Recent ideas on inflationary and quantum cosmology provide the first non-philosophical reasons for favouring an essentially 'flat' universe with $\Omega = 1$. (These concepts also, incidentally broaden our cosmological horizons by opening the prospect that there may be interesting structures and inhomogeneities on scales far exceeding our present Hubble radius: the extrapolation from an apparent $\Omega \leqslant 1$ to an infinite homogeneous Friedmann model may be unwarranted.) Cosmologists who are seduced by these ideas must postulate that the bulk of the matter in the universe in non-baryonic (or else they must adopt a non-standard model for nucleosynthesis of the light elements).

5–0 times more unseen mass is required for $\Omega = 1$ than is reliably inferred from virial-type considerations. This excess mass can only be reconciled with dynamical arguments—and, more specifically, with the rather small observed deviations of galaxy motions from the Hubble flow—if the galaxy distribution gives an exaggerated impression of the overall inhomogeneity on scales of 10–20 Mpc. This could happen ether if the excess mass were 'hot', or if there were more complex reasons (of the kind outlined in Section 3.3) why galaxies do not trace the mass distribution. Until these latter possibilities can be understood on a quantitative basis, data on galaxy clustering cannot reliably tell us what Ω really is.

Acknowledgements

I am grateful for helpful discussions with many colleagues on topics summarized here, and especially to G. Blumenthal, A. Dekel, J. Einasto, S. Faber, C. Hogan and J. Primack.

References

Abbott, L., Sikivie, P. 1983, *Phys. Lett.*, **120B,** 133.
Bahcall, J. N. 1984, *Astrophys. J.*, **276,** 169.
Bahcall, J. N., Soneira, R. M. 1980, *Astrophys. J. Supp. Ser.*, **44,** 73.
Bean, A. J., Efstathiou, G., Ellis, R. S., Peterson, B. A., Shanks, T. 1983, *Mon. Not. R. astr. Soc.*, **205,** 605.
Blumenthal, G., Faber, S. M., Primack, J. R., Rees, M. J. 1984, *Nature*, **311,** 517.
Bond, J. R., Efstathiou, G. 1984, *Astrophys. J. Lett.*, in press.
Bond, J. R., Szalay, A. S., White, S. D. M. 1983, *Nature*, **301,** 584.
Boughn, S. P., Saulson, P. R., Seldner, M. 1981, *Astrophys. J.*, **250,** L15.
Canizares, C. R. 1982, *Astrophys. J.*, **263,** 508.
Carr, B. J., Bond, J. R., Arnett, W. D. 1984, *Astrophys. J.*, **277,** 445.
Couchman, H. P. 1984, *Mon. Not. R. astr. Soc.*, (submitted).
Cowsik, R., McClelland, J. 1973, *Astrophys. J.*, **180,** 7.
David, Y., Reeves, H. 1980, in *Physical Cosmology*, Eds R. Balian, J. Audouze & D. N. Schramm, North-Holland, Amsterdam, p. 443.
Davis, M., Peebles, P. J. E. 1983, *A. Rev. Astr. Astrophys.*, **21,** 109.
Dekel, A., Einasto, J., Rees, M. J. 1985, in preparation.
Dine, M., Fischler, W. 1983, *Phys. Lett.*, **120B,** 137.
Doroshkevich, A. G., Shandarin, S. F., Zel'dovich, Ya. B. 1983, in *IAU Symp.* 104: *Early Evolution of the Universe and its Present Structure*, Eds G. O. Abell & G. Chincarini, D. Reidel, Dordrecht, p. 387.
Doroshkevich, A. G., Zel'dovich, Y. B., Novikov, I. D. 1967, *Sov. Astr.*, **11,** 233.
Einasto, J., Kaasik, A., Kalamees, P., Vennik, J. 1975, *Astr. Astrophys.*, **40,** 161.
Einasto, J., Lynden-Bell, D. 1982, *Mon. Not. R. astr. Soc.*, **199,** 67.
Faber, S. M. 1984, in *Proc ESO/CERN Conference on Cosmology and Particle Physics*, CERN Publications.
Fabian, A. C., Kembhavi, A. 1982, in *Extragalactic Radio Sources*, Eds D. Heeschen & C. Wade, D. Reidel, Dordrecht, p. 453.
Fabian, A. C., Nulsen, P. E. J., Canizares, C. 1984, *Nature*, **310,** 733.
Fabricant, D., Lecar, J., Gorenstein, P. 1980, *Astrophys. J.*, **241,** 552.
Forman, W., Jones, C. 1982, *A. Rev. Astr. Astrophys.*, **20,** 547.
Frenk, C., White, S. D. M., Davis, M. 1983, *Astrophys. J.*, **271,** 417.
Fukugita, M., Watamura, S., Yoshimura, M. 1982, *Phys. Rev. Lett.*, **48,** 1522.
Geller, M. J., Huchra, J. P. 1983, *Astrophys. J. Supp. Ser.*, **52,** 61.
Geller, M. J., Peebles, P. J. E. 1973, *Astrophys. J.*, **184,** 329.
Gunn, J. E., Lee, B. W., Lerche, I., Schramm, D. N. Steigman, G. 1978, *Astrophys. J.*, **223,** 1015.
Hartquist, T. W., Cameron, A. G. W. 1977, *Astrophys. Space Sci.*, **48,** 145.
Hegyi, D. J., Gerber, G. L. 1977, *Astrophys. J.*, **218,** L7.
Hoffman, G. L., Salpeter, E. E. 1982, *Astrophys. J.*, **263,** 485.
Hogan, C. 1984, *Astrophys. J.*, in press.
Hogan, C., Rees, M. J. 1984, *Nature*, **311,** 109.
Huchra, J. P., Geller, M. J. 1982, *Astrophys. J.*, **257,** 423.
Ipser, J., Sikivie, P. 1933, *Phys. Rev. Lett.*, **50,** 925.
Joeveer, M., Einasto, J. 1976, *Tartu astr. Obs. Publ.*, **54,** 77.
Kahn, F. D., Woltjer, L. 1959, *Astrophys. J.*, **130,** 705.
Kashlinsky, A., Rees, M. J. 1983, *Mon. Not. R. astr. Soc.*, **205**, 955.

Lynden-Bell, D. 1982, in *Astrophysical Cosmology*, Eds H. A. Brück, G. V. Coyne & M. S. Longair, Vatican Publications, p. 85.
Marx, G., Szalay, A. 1972, in *Proc, Neutrino '72*, Vol. 1, Technoinform, Budapest p. 191.
Melnick, J., White, S. D. M., Hoessel, J. 1977, *Mon. Not. R. astr. Soc.*, **180**, 207.
Oort, J. H. 1932, *Bull. astr. Inst. Netherl.*, **6**, 249.
Pagel, B. E. J. 1982, *Phil Trans. R. Soc. Lond.*, **A307**, 19.
Peebles, P. J. E. 1980, in *Physical Cosmology*, Eds R. Balian, J. Audouzc & D. N. Schramm, North-Holland, Amsterdam, p. 213.
Peebles, P. J. E. 1984, *Astrophys. J.*, **277**, 470.
Preskill, J., Wise, M., Wilczek, F. 1983, *Phys. Lett.*, **120B**, 127.
Press, W. H., Gunn, J. E. 1973, *Astrophys. J.*, **185**, 397.
Rees, M. J. 1983, in *Formation and Evolution of Galaxies and Large Structures in the Universe*, Eds J. Audouze & J. T. T. Van, D. Reidel, Dordrecht, p. 237.
Rubin, V. C. 1983, *Science*, **220**, 1339.
Tarbet, P. W., Rowan-Robinson, M. 1982, *Nature*, **298**, 777.
Uson, J., Wilkinson, D. T. 1984, *Astrophys. J.*, **277**, L1.
Vilenkin, A. 1984, *Phys. Lett.*, in press.
White, S. D. M., Huchra, J., Latham, D., Davis, M. 1983, *Mon. Not. R. astr. Soc.*, **203**, 701.
Wilczek, F. 1982, *Phys. Rev. Lett.*, **49**, 1549.
Witten, E. 1984, *Phys. Rev.*, **D30**, 272.
Yang, J., Turner, M. S., Steigman, G., Schramm, D. N., Olive, K. A. 1984, *Astrophys. J.*, **281**,493.
Zwicky, F. 1933, *Helv. Phys. Acta*, **6**, 110.

J. Astrophys. Astr. (1984) **5**, 349–368

Identifications and Spectra of Extragalactic Radio Sources

M. S. Longair *Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*

S. J. Lilly *Department of Astrophysical Science, Princeton University, New Jersey 08544, USA*
(*Invited article*)

1. Introduction

The identification of radio sources has turned out to be one of the most productive techniques for discovering new and important classes of astrophysical objects over the last 35 years. It is remarkable that this should have been so but we can now recognise with hindsight the reasons for this. The key point is that the radiation mechanisms responsible for the radio emission from many classes of astrophysical objects are not 'thermal' emission but involve either the radiation of very high energy electrons through the synchrotron or related mechanisms or else involve coherent processes of emission. As a result, the objects associated with these sources must be somewhat remarkable astrophysically since they must contain sources of very high energy particles or else act as hosts for the physical conditions which favour the emission of coherent radiation.

The intention of this brief review is to concentrate upon the identification and redshifts of extragalactic radio sources but it is worthwhile beginning with a historical survey of some of the more remarkable products of the process of identification of radio sources in both its galactic and extragalactic contexts. We will then look in more detail at some of our recent work on the identification of extragalactic radio sources at optical and infrared wavelengths and at the problems of measuring their redshifts and spectra. We will make the case that this procedure offers us one of the most effective techniques for studying the origin and evolution not only of radio sources but also of galaxies in general.

2. The objects associated with radio sources—a brief history

The discovery of cosmic radio emission by Jansky in the early 1930s, almost contemporaneous with the founding of the Indian Academy of Sciences, was followed by the first surveys of the radio sky after the Second World War. These early surveys established the existence of discrete sources of radio emission of extra-solar-system origin. By a quirk of nature, it turned out that the two brightest discrete radio sources, known then by the constellations in which they lay as Cassiopaeia A and Cygnus A, were found to lie close to the Galactic plane and for a time it was thought that most of the 'radio stars', as they were known then, were objects within our own Galaxy. However, some of the discrete sources had to be extragalactic. In 1949, Bolton, Stanley & Slee showed that the source known as Virgo A was associated with the giant elliptical

galaxy M 87 in the Virgo cluster and Centaurus A was associated with the giant elliptical galaxy NGC 5128. Other extragalactic identifications followed soon after.

Perhaps the most important of the early identifications were those of the two brightest radio sources in the northern sky, Cassiopaeia A and Cygnus A. Interferometry by F. Graham Smith (1951, 1952) at Cambridge resulted in accurate radio positions for these sources and subsequent optical observations by Baade & Minkowski (1954) led to their identification with very faint optical objects. Cassiopaeia A turned out to be associated with the remnant of a supernova which must have exploded unseen by optical astronomers about 250 years ago. Cygnus A turned out to be associated with a faint distant galaxy at a redshift of 0.05, by far the most distant radio source known at that time.

These sources were of importance for a variety of different reasons. During the 1950s, it became apparent to a number of authors that synchrotron radiation was the most likely emission mechanism for these sources and this automatically implied that sources such as Supernovae and radio galaxies must be intense sources of very high energy particles. It was immediately recognised that Cygnus A was such a powerful radio emitter that similar radio objects should be identifiable at cosmological distances if the radio surveys extended to somewhat fainter flux densities. This cosmological goal was one of the main motives for the intense activity which proceeded through the 1950s and 1960s to produce surveys of radio sources which would be suitable for cosmological studies.

At the same time, these extragalactic radio sources produced a wide range of new physical problems concerning the origin of the huge fluxes of relativistic particles which were inferred to be present in sources such as Cygnus A. The puzzle deepened with the discovery by Jennison & Das Gupta (1953) that the radio emission of Cygnus A did not originate within the galaxy itself but in two radio lobes on either side of the parent galaxy located about 100 kpc on either side of the nucleus of the galaxy.

Thus, by the mid-1950s two themes which have dominated extragalactic radio astronomy ever since were exposed. On the one hand, there is the problem of understanding how giant elliptical galaxies can generate huge fluxes of relativistic particles and eject them into radio lobes which can in some cases extend more than 1 Mpc beyond the parent galaxy. On the other hand, there is the problem of using extragalactic radio sources as cosmological probes.

The identification of radio sources continued apace throughout the 1950s as more accurate positions became available for the fainter sources known at that time. On the Galactic front, many supernova remnants and HII regions were discovered which were highly obscured or unobservable in the optical waveband. On the extragalactic front, it was found that many of the early identifications were giant elliptical galaxies, many of them being the brightest galaxies in clusters. A significant discovery whose significance was not fully appreciated at the time was the discovery by Morgan of the class of N-galaxies associated with some of the bright radio sources (see Matthews, Morgan & Schmidt 1964). Morgan's interest was in the morphological classification of the optical counterparts of the radio sources and he recognised this hitherto unknown class of galaxy in which the light is dominated by the emission from the nucleus which appears to be star-like on photographic plates. Indeed, all the N-galaxies known were radio galaxies and their relation to other galaxies with active nuclei such as the Seyfert galaxies which at that stage had not become a heavy industry was not apparent.

The most remarkable result of this period of activity was the discovery in 1960 of the

class of quasi-stellar objects (quasars) by Matthews & Sandage (1963) from these optical identification surveys. The nature of these radio sources which appeared to be stars on photographic plates and yet had properties which were different from any other known class of stars was obscure. The great breakthrough was made in 1962 by Schmidt when he discovered that these quasi-stellar objects were not stars at all but distant extragalactic objects—the key object was the radio source 3C 273 which is still by far the brightest known quasar in the sky having apparent magnitude about 12 to 13 and yet has redshift 0.158 (Schmidt 1963). The two puzzling features of this new class of objects were their very great optical luminosities, amounting in some cases to more than 1000 times the luminosity of a galaxy such as our own and the fact that this intense source of radiation must lie within a very compact volume indeed because this huge luminosity was found to vary on the timescale of years.

The discovery of quasars sparked off one of the most important growth areas of modern astronomy. Since 1960, many quasars have been discovered and it is now appreciated that there is a continuity in nuclear activity in galaxies all the way from galaxies such as our own, through the Seyfert galaxies to the N-galaxies and quasars. This same optical identification procedure continued to make fundamental contributions to these studies, in particular, with the recognition in 1968 of the class of what are now regarded as probably the most extreme class of quasar-like objects, the BL Lac objects (MacLeod & Andrew 1968; Schmitt 1968). These are compact high-frequency radio sources which exhibit rapid variability on timescales of days and weeks. In the case of BL Lac and other members of this class, the optical spectra are almost featureless and it may be that in these sources the nucleus itself is being observed more or less unobscured by surrounding gas.

Another remarkable discovery which has come from the ability of the radio identification technique to find quasars in large numbers has been the discovery of gravitational lenses. By 1979, well over 1000 quasars had been discovered and at that time a very close binary quasar system was discovered associated with the radio source 0957 + 561, the two quasars being separated on the sky by about 6 arcsec. The striking feature of these quasars was not only their proximity to one another on the sky but also the fact that their magnitudes were similar and their optical spectra identical (Walsh, Carswell & Weymann 1979). The subsequent discovery of a faint galaxy lying between the two quasars was convincing evidence that this was the first example of a gravitational lens system (Young *et al.* 1980; Stockton 1980). The single image of a distant quasar was split into two bright components by the gravitational lens effect. Since that time more examples of this same phenomenon have been discovered.

The process of identifying extragalactic sources continues to produce surprises. One of the more important of these recent discoveries has been that of a class of quasar-like objects by Rieke, Lebofsky & Kinman (1979) which was discovered because its members were readily detectable in the infrared waveband but invisible or very faint indeed in the optical waveband. They have very steep optical spectra which flatten in the far-infrared waveband. The key importance of these objects is that they have spectra which cut off at high frequencies almost as rapidly as is theoretically possible for incoherent radiation by relativistic particles. As such, they provide some important tests of the theory of the processes by which these compact objects emit intense optical and infrared emission.

In the meantime, we should not forget the important role which has been played in the identification of Galactic objects from radio surveys. Probably the most spectacular

example is that of the discovery of pulsars in 1967 by Bell & Hewish (Hewish *et al.* 1968). In this case, the identification is with the class of magnetized rotating neutron stars which emit intense radio pulses by coherent emission mechanisms. Optical identifications have only been made for two pulsars, those associated with the young supernova remnants, the Crab nebula and the Vela supernova remnant.

Another startling discovery of a singular galactic object is that of SS 433. This object had been catalogued as a star with strong Hα emission by Stephenson & Sanduleak (1977). It was however found to be a strong variable radio source, apparently one of a number of such compact sources which lay in the Galactic plane. Subsequent spectroscopy showed it to be a quite remarkable Galactic object in which the radio emission originates from a core source and from two jets which move out from the centre at velocities which are a significant fraction of the velocity of light (in fact $v = 0.26\ c$). Optical spectroscopy as well as radio maps made at different epochs have shown that these relativistic jets precess about the mean radio axis of the system. SS 433 bears some resemblance to extragalactic sources but, being a nearby Galactic object, the jet phenomenon can be observed on much shorter timescales.

This summary indicates the diversity of high-energy astrophysical phenomena which have been discovered from the objects associated with discrete radio sources. In a word, wherever there is relativistic matter present in the universe, it may leave a definite signature through the strong radio emission (and possibly X-ray emission as well) associated with the relativistic electrons and magnetic fields in the source region. This rich harvest is by no means exhausted and it is reasonable to expect that this procedure will lead to the discovery of other important objects which will in turn lead to a deeper understanding of high-energy astrophysical processes in different astronomical environments.

We will now look in detail into one of the classical applications of this procedure—the identification of extragalactic radio sources and their role in understanding the evolution of the radio source population as a whole and of the processes involved in the origin and evolution of galaxies.

3. The present status of optical identification programmes of bright extragalactic radio sources

One of the main reasons for studying the optical identifications of extragalactic radio sources was to understand the counts of radio sources. A recent survey of the counts of extragalactic radio sources is shown in Fig. 1 in a representation which shows the counts at a wide range of frequencies. As usual, the counts are differential and in each flux-density interval have been arbitrarily normalized to the Euclidean prediction, $\mathrm{dN} = K\ S^{-5/2}\mathrm{dS}$. These counts are inconsistent with the predictions of uniform world models in which the counts expressed as $\Delta N/\Delta N_0$ should decrease monotonically with decreasing flux density. It is the fact that the counts increase with decreasing flux density at high flux densities and then converge at low flux densities which leads to the inferred strong evolution of the radio source population.(see *e.g.* Longair 1978). Expressed simply, the point is that even at the very highest flux densities the typical radio source is at a significant cosmological distance and therefore when one looks fainter, the sources should be even more distant. If one observes more faint sources than one expects, there

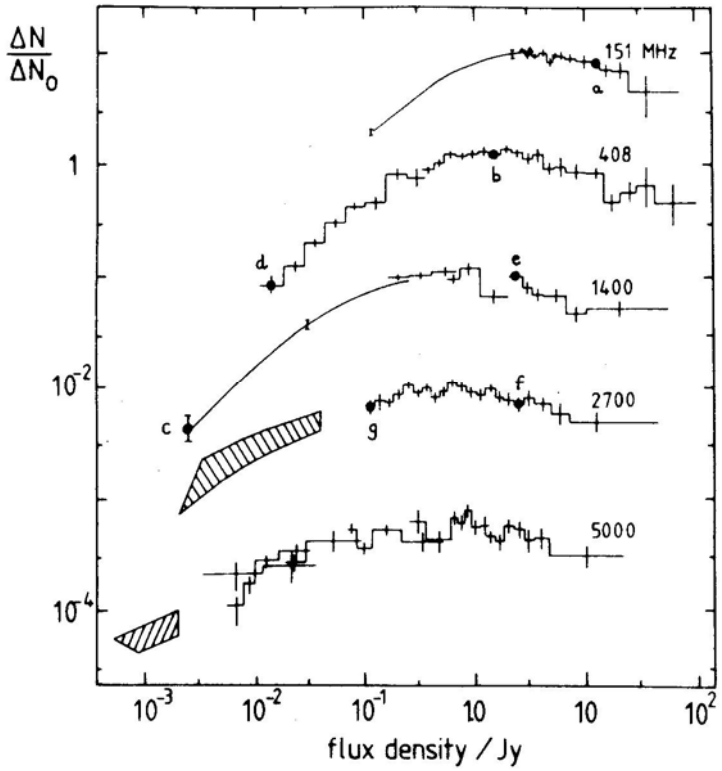


Figure 1. The counts of radio sources shown in differential form normalized to the expectation of a uniform Euclidean world model (from Wall & Benn 1982). Also indicated on the diagram are the flux densities at which major deep optical identification surveys have been carried out and which are described in the text:

(a) The 3CR sample
(b) The 1 Jy sample
(c) The Berkeley-Westerbork deep survey
(d) The 5C 12 survey
(e) The 1420 MHz survey of Bridle *et al.* (1972)
(f) The 2.7 GHz all-sky survey
(g) The 2.7 GHz deep survey

must be a significant increase in the number of sources observed at those earlier cosmic epochs.

Superposed on the diagram are certain flux density limits for which particularly intense optical identification programmes have been carried out. These surveys are listed in the caption to Fig. 1. At the highest flux densities, it is feasible to attempt to identify most of the sources in the sky but at lower flux densities, this becomes an excessively ambitious task and only small regions of sky can be systematically studied.

3.1 *The 3CR Complete Sample*

As an example of what can now be achieved if a great deal of effort is expended, it is instructive to look at the history of the optical identification content of the sample which has been studied most intensively of all samples to date—this is what is known as the 3CR sample of radio sources.

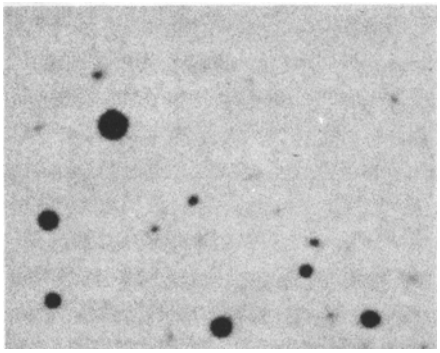
The 3C survey of radio sources (3C standing for Third Cambridge) was defined by Edge *et al.* (1959) and was intended to provide a reliable survey of all the bright radio sources in the sky selected at a frequency of 159 MHz. This survey was made with an interferometer and was completed at the time when the sensitivity of such catalogues to the effects of confusion and partial resolution were becoming fully appreciated. The survey was superseded by the Revised 3C Catalogue (3CR) which was prepared by Bennett (1962) using observations made with the fixed element of the Cambridge 4C interferometer at 178 MHz. These were total power measurements and the instrument had good angular resolution in right ascension but poor resolution in declination.

One of the most important purposes of these samples has been to define the statistical properties of the types of sources found and so particular care has to be taken to ensure completeness and reliability for the sources included in these samples. Two subsequent re-analyses of these samples have been undertaken in the light of higher and lower resolution observations available since the original survey and of deeper surveys of the northern hemisphere. These samples are listed by Jenkins, Pooley & Riley (1977) and the most recent and probably the most complete analysis—which is possible without starting absolutely from scratch again—by Laing, Riley & Longair (1983). This last paper in particular indicates the complexities involved in producing a reliable catalogue of sources which will be complete according to well-defined selection criteria. Such samples require observations with low angular resolution to ensure that the total flux density of the radio sources are measured and also high angular resolution to eliminate the effects of confusing nearby faint sources which systematically enhance the flux densities of the radio sources.

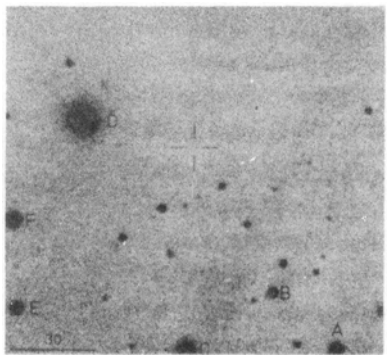
The high angular resolution is also needed in order to define more precisely the search area for the associated optical objects. The 3CR positions of Bennett (1962) were only adequate for obtaining optical identifications with objects which were relatively bright. The optical identification of sources with objects of 18th magnitude and sometimes fainter only became feasible when positions accurate to one arcmin and better became available about 1963 from the work of Clark (1964) at Cambridge and Fomalont *et al.* (1964) and Read (1963) at the California Institute of Technology. Optical identification surveys using these improved positions by Longair (1965) and Wyndham (1965) resulted in many new identifications of sources down to 19th magnitude and included for the first time significant numbers of quasars and many new radio galaxies including N-galaxies.

Progress was halted until higher resolution maps were made with Earth-rotation synthesis telescopes, in particular the Cambridge one-mile and 5-km telescopes. The latter provided maps with 2 arcsec resolution at 5 GHz and positional accuracy much better than 1 arcsec. The consequence was that optical identifications could be made well below the limit of about 20th magnitude of the Palomar–National-Geographic-Society Sky Survey. The principal surveys which were made in the early and mid-1970s were due to Kristian, Sandage & Katem (1974, 1978), Longair & Gunn (1975), Longair (1975), Laing *et al.* (1978) and Riley, Longair & Gunn (1980). The techniques used were either direct photography or photography using an image tube with the Palomar 5-metre telescope. These surveys extended the optical identifications to about 22 mag, at which level about 80 per cent completeness was obtained for the 3 CR sample.

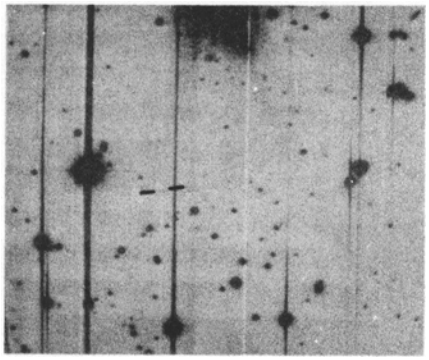
The penultimate step was taken in the late 1970s by Gunn *et al.* (1981) when a prototype Texas Instruments CCD detector became available for observations on the Palomar 5-metre telescope in a camera known affectionately as PFUEI (one



(a)



(b)



(c)

Figure 2. The field of the radio source 3C 65 observed on (a) the prints of the 48-inch Palomar Sky Survey (red plate); (b) deep plates taken with the 5-metre Palomar telescope with an image tube (IIIaJ plates), and (c) CCD exposures taken with the 5-metre Palomar telescope and the PFUEI camera (30 minute exposure in the *i* waveband).

interpretation of this acronym was Prime-Focus Universal Extragalactic Instrument). Because of the very high quantum efficiency of these detectors, a limiting magnitude of 23.5 to 24 could be obtained and at this level of sensitivity, about 95 per cent of all the 3CR sources in the complete sample of Laing, Riley & Longair (1983) were identified. Fig. 2 shows an example of the appearance of the optical field of the radio source 3C 65 as observed at these three sensitivity levels. Identifications for the sources added to the 3CR sample by Laing, Riley & Longair (1983) were made by Riley, Eales & Baldwin (1984).

The final step was taken not in the optical but in the infrared waveband at 2.2 μm by Lilly & Longair (1985). The three remaining unidentified sources in a complete subsample of the 3CR sample which had been searched to the faintest optical limits were detected using the UK Infrared Telescope in Hawaii using a high sensitivity InSb detector.

This history is summarized in Fig. 3 which shows a histogram of the apparent magnitude distribution of the radio galaxies and quasars in a complete 3CR sample with an indication of the relative sensitivities of the different types of detectors. Fig. 4 shows the corresponding apparent K magnitude distribution at 2.2 μm for the radio galaxies in the sample including the three objects which we detected as infrared sources but which were invisible in the optical CCD observations. The reason for this apparent increased sensitivity in the infrared waveband is the shape of the typical spectral energy distribution of a giant elliptical galaxy (Fig. 5). Because of the rapid cut-off in the spectral energy distribution of galaxies towards the ultraviolet region of the spectrum, they become fainter in the optical waveband much more rapidly than would be predicted by the inverse-square law when they are observed at large redshift. On the other hand, the galaxies still remain relatively bright objects in the infrared waveband. We will return to this point later in this article.

What should strike the reader is the fact that all this trouble should be necessary to identify the brightest 180 extragalactic radio sources in the northern sky. The contrast with the optical sky could not be more striking. The brightest 180 objects in the optical sky are nearby stars whereas the 180 brightest radio sources observed away from the Galactic plane include the most distant galaxies known. As was predicted by the radio astronomers of the early 1950s, the study of the brightest radio objects in the sky takes one very rapidly to cosmological distances and to detect the most distant of these, let alone measure their redshifts, requires the use of the most powerful facilities we possess in the optical and infrared wavebands.

Essentially all the sources in the 3CR sample have steep radio spectra and have extended radio structures, generally of the classical double source type. About 20 per cent of the sources are quasars and the remainder are radio galaxies, about 5–10 per cent of them having N-type morphology. We will look in more detail into some of their properties in Section 5.

3.2 *The 2.7 GHz All-Sky Survey*

It was realised in the 1960s that one could not obtain a complete picture of the extragalactic radio source population without undertaking surveys at high as well as low radio frequencies. By then it was realised that there exists a population of compact flat-spectrum radio sources which appear in considerable numbers in a high-frequency survey but are poorly represented in the low-frequency samples. The Parkes

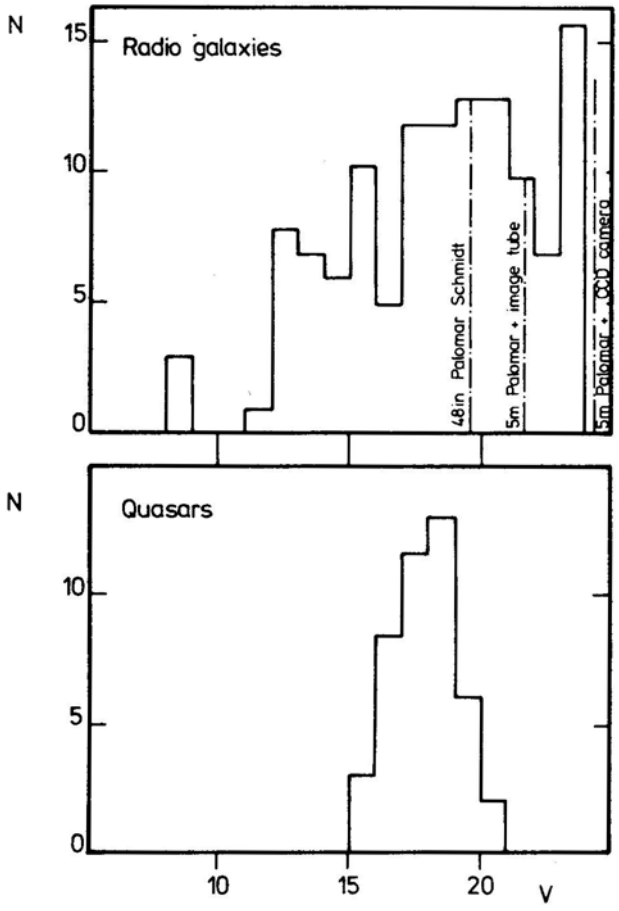


Figure 3. The apparent magnitude distribution of the optical identifications in the 3CR catalogue for radio galaxies and quasars. The practical limits of different procedures are shown on the diagram.

2.7 and 5 GHz surveys were the first to cover large regions of the sky and indeed large numbers of the fiat-spectrum radio sources were discovered. It turned out that a large fraction of the flat-spectrum radio sources were quasars and many of them had very large redshifts with values up to about 3. In the period 1969 to 1980, the northern hemisphere was surveyed at 5 GHz in a joint programme involving the National Radio Astronomy Observatory in the USA and the Max Planck Institute for Radio Astronomy in Bonn.

Using these data, Wall & Peacock (1985) began the systematic compilation of complete samples of sources at 2.7 GHz which fulfilled the analogous function to the 3CR sample but now for sources selected at high frequencies and for the whole of the extragalactic sky away from the Galactic plane. This survey has gone through the same trials as the 3CR sample. In some ways, it has been easier because, with a larger fraction of the sources being quasars, these have been easier to identify but, nonetheless, the faintest of the sources in the sample have required the use of CCD observations (Peacock *et al.* 1981). In the final sample of 233 sources selected to a limiting flux density of 2 Jy at 2.7 GHz, 93 per cent of the sources have optical identifications.

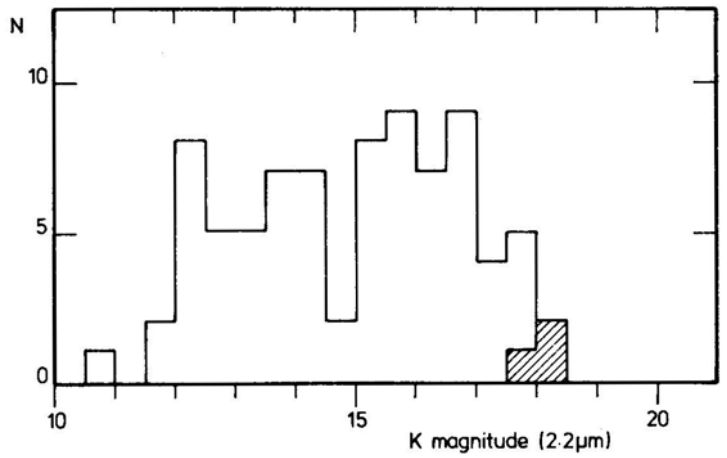


Figure 4. The apparent K (2.2 μm) magnitude distribution of radio galaxy identifications in a complete sub-sample of the 3CR sample (Lilly & Longair 1985). The shaded boxes are identifications which were made in the infrared waveband alone.

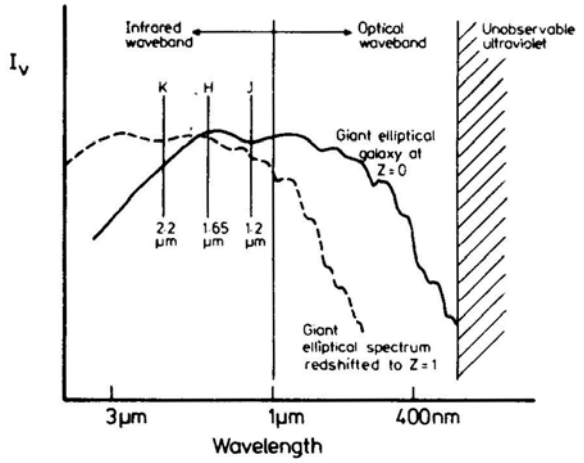


Figure 5. Illustrating the effect of redshifting the spectral energy distribution of a giant elliptical galaxy. Galaxies at redshifts of 1 are strong emitters in the infrared waveband but are relatively much fainter in the optical waveband.

The discovery of significant numbers of flat-spectrum sources associated with quasars comes as no surprise but there have been other surprises associated with these samples. Perhaps the most important is the discovery of a class of compact steep-spectrum radio sources which appear in only small numbers in the low-frequency samples. The reason for this is that they have spectra which are steep at frequencies greater than 1 GHz but which display spectral curvature at low frequencies so that they are not prominent members of the low-frequency samples. These turn out to be associated with quasar-like objects and radio galaxies. At least some of them turn out to be very intense infrared emitters and in fact it is a misnomer to call them optical objects at all (Ricke, Lebofsky & Kinman 1979). They have such steep optical–infrared spectra

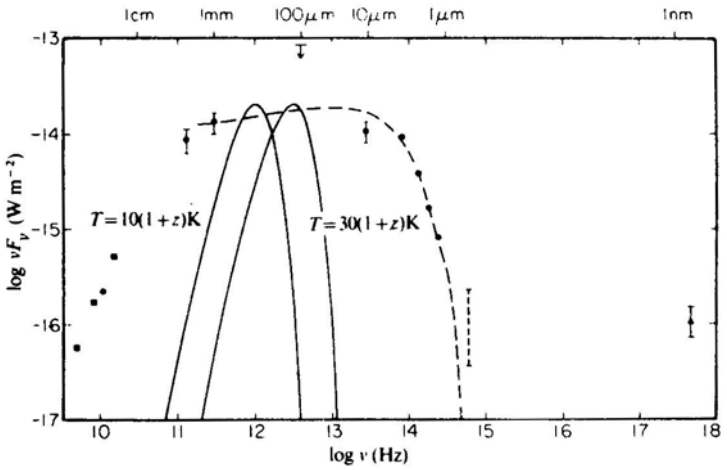


Figure 6. The infrared-optical spectrum of the radio source 1413 + 135 which is a very intense emitter in the infrared waveband but is very faint optically (Beichman *et al.* 1981).

that even the high-frequency exponential cut-off of synchrotron radiation is barely steep enough to explain them (Fig.6). It is also important that these are among the most intense radio emitters known, their intrinsic radio luminosities being as great as those of the most powerful known quasars.

3.3 *Other High Flux Density Radio Surveys*

Besides these surveys, other important surveys have been carried out, notably the NRAO-Bonn 5 GHz surveys, the Molongo survey at 408 MHz, the Jodrell Bank survey at 966 MHz (Porcas *et al.* 1980) and the 1420 MHz survey of Bridle *et al* (1972). These are all important in different ways and have contributed to the definition of the radio source populations found at high flux densities. In particular, the 1420 MHz survey—because of its sensitivity to sources of low surface brightness—included many sources of very large angular extent which had not been included in the other surveys. The 5 GHz survey has produced large numbers of quasar identifications (Kühr *et al.* 1981). It was as part of the Jodrell Bank identification survey that the double quasar 0957 + 561 A and B, which is now confirmed as a gravitational lens, was discovered.

4. Deep radio surveys

Granted the major effort required to identify even the brightest samples of radio sources, it is apparent that to extend these studies to deep radio surveys requires a very large effort indeed. The conclusion of the work on the bright-source surveys is that it is essential to work with well-defined samples of sources for which high-resolution radio maps are available. Radio structures with angular resolution of about 1 arcsec are generally necessary for the very faint sources but this may not be the whole story. A number of cases are known in which the extended lobes have very low radio surface brightness and this is only detected by radio telescopes with angular resolution of 10 to

20 arcsec (*e.g.* 3C16—see Riley, Longair & Gunn 1980). It is fortunate that the Very Large Array (VLA) has exactly the attributes necessary for carrying out these types of studies. Its combination of high angular resolution and sensitivity make it a unique tool for providing the radio maps necessary to undertake the identification of extragalactic radio sources. Besides the high-resolution radio structures of the lobes of the sources, the discovery of compact central radio components in many of the radio sources has been of particular value in confirming the identifications of many of the diffuse sources (Laing, Riley & Longair 1983).

A standard procedure has now been developed of first searching for identifications on the sky surveys made with the large Schmidt telescopes, the Palomar-National-Geographic-Society Sky Survey in the north and the ESO/SER C Schmidt Survey in the southern hemisphere. To proceed further, it is either necessary to use stacked plates, as has been undertaken in the optical identification survey of the 5C 12 region by Grueff *et al.* (1984) or to use 4-metre reflectors with sensitive detecting elements, the best of these being the CCD detectors which are now available on most large telescopes. An example of this is the optical identification searches of the 5C 6 and 7 fields by Perryman *et al.* (1982) which succeeded in finding identifications for about 35 per cent of the sources in these surveys.

In evaluating the significance of these deeper surveys, it is useful to look again at Fig. 1 which shows the source counts at a wide range of radio frequencies. The most detailed surveys have been carried out at the highest flux densities, in the region of the source counts in which the numbers increase more rapidly with decreasing flux density than would be expected. Then at lower flux densities, the numbers converge more or less as expected in the standard world models. If we consider the low flux densities first, the significance of the decrease in the counts is that if we consider a single luminosity class of sources in a uniform world model, the numbers of sources we expect to see decreases very rapidly beyond a redshift of about 1. This occurs for two reasons—first of all, the sources become fainter more rapidly than would be predicted by the inverse-square law because of the additional effects of redshift and, secondly, because the element of comoving volume decreases with increasing redshift beyond redshifts of about $1/\Omega$ where Ω is the density parameter of the universe. This means that, as we proceed to fainter and fainter flux densities, rather than see further and further away, we tend to see lower luminosity objects at redshifts of about 1 and the much more distant powerful sources are rare objects. The reason why there are so many objects in the bright source samples with redshifts greater than 1 is entirely due to the anomalies in the source counts. The excess of faint radio sources is the reason why we observe large redshift objects with much greater ease than we have any right to expect. The consequence of this is that, if we make identification surveys at very faint radio flux densities, we are not necessarily seeing any further away than we are at flux densities at which the counts display a maximum excess over the predictions of uniform world models.

These expectations are by and large fulfilled in deep optical identification surveys. Perhaps the largest of these is the survey carried out by a Westerbork-Berkeley collaboration in which systematic deep surveys were carried out by Windhorst at Westerbork and the complementary optical data provided by Koo and Kron in their deep optical survey of selected regions (see Windhorst 1984). The striking feature of the radio survey is the fact that at the faint radio flux densities at which they were working, not many of the radio sources exhibited the type of extended double structure which is typical of the sources found in the high flux-density surveys. Rather, the sources appear

to be much more similar to the intrinsically weak radio sources. This is confirmed by the optical identifications of these sources in which faint galaxies are observed. One of the very nice discoveries of this survey has been that of a number of blue galaxies which they identify with 'starburst' galaxies. It is likely that these galaxies make a major contribution to the flattening of the radio source counts found in the very deepest radio surveys which have been carried out at the VLA by Condon and Mitchell (see *e.g.* Mitchell 1983).

5. The 1 Jy survey—optical and infared observations

If one's main interest in the optical identification of extragalactic radio sources is the study of the universe in the distant past, the claim can now be substantiated by direct observation that the way to do this is to concentrate upon that region of the radio-source counts where there is the maximum discrepancy between the observed counts of sources and the predictions of uniform world models. There are two major surveys currently underway which address this problem directly. One of these was carried out by Allington-Smith and consisted of a complete sample of 59 radio sources selected at a flux density of 1 Jy at a frequency of 408 MHz. This corresponds to a sample selected at flux densities about 5 times fainter than those of similar sources in the 3CR catalogue and, as can be seen from Fig. 1, this corresponds to the flux density at which the differential source counts at 408 MHz reach a maximum before decreasing at flux densities below about 0.1 Jy. The second consists of a major survey of a sample of about 180 sources selected at 2.7 GHz at a flux density of about 0.1 Jy by Downes, Peacock & Savage. The first of these surveys has now produced some intriguing results.

The 1 Jy sample was selected by Riley & Longair in 1978 with the intention of extending the type of analysis already underway at that time for the 3CR sample. The 59 sources were mapped with the Cambridge radio telescopes by Allington-Smith (1982). Optical identifications of these sources were first sought on the prints of the Palomar Sky Survey and then CCD observations were obtained for all those which were identified close to the limits of the Sky Survey and for those which were unidentified (Allington-Smith *et al.* 1982). The net result was that to a limiting apparent magnitude of $r \simeq 23$, about 75 per cent of the sources in unobscured regions could be identified. To the same magnitude limit, well over 90 per cent of the sources were identified in the 3CR sample.

All the accessible sources in the sample were observed with the UK Infrared Telescope including those which remained optically unidentified. Lilly, Longair & Allington-Smith (1985) have found that virtually all the sources could be detected at 2.2 μm. The apparent magnitude distribution in the K waveband for the sources in the 1 Jy sample is shown in Fig. 7. It can be seen by comparison with Fig. 4 that the whole distribution is fainter by about 1.5 to 2 magnitudes as compared with the 3CR sample. This result is entirely consistent with the hypothesis that the members of the 1 Jy sample are the more distant counterparts of the same types of objects seen in the 3CR sample. Our interpretation of these results for the empty fields is that we have detected the radio galaxies associated with these radio sources (see Section 7).

This interpretation has been confirmed by subsequent very deep exposures taken with the Palomar 5-metre telescope by Gunn & Lilly (1985). They made exposures of up to 1 hour of the optically unidentifed source fields in this sample using the four-array

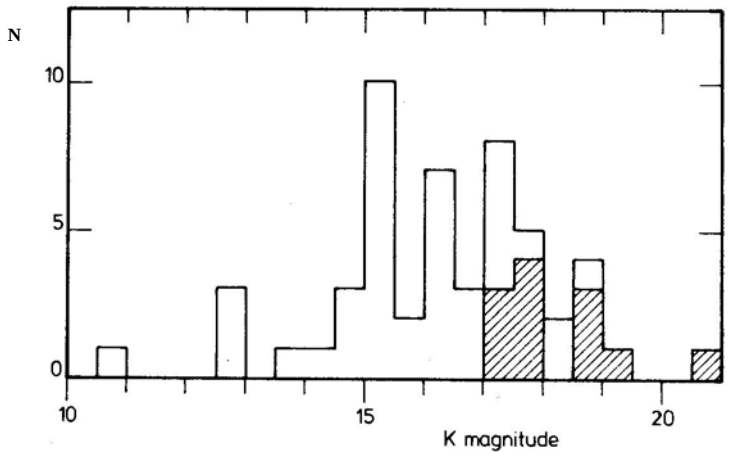


Figure 7. The *K* magnitude distribution of the radio sources in the complete sample of Allington-Smith selected at 1 Jy at 408 MHz. The detections fainter than $K = 19$ are detected at less than 2 standard deviations above the noise and are at best marginal detections (Lilly, Longair & Allington-Smith 1985).

CCD camera. An example of the enormous increase in sensitivity obtained in this work is shown by a comparison of the fields of the radio source 1056 + 39 observed with the Palomar 48-inch Schmidt telescope, with the Palomar 5-metre telescope with a 5 minute exposure with PFUEI and with a 1 hour exposure with the new four-array camera (Fig. 8). It is evident from the last image that there is a system of faint galaxies in exactly the position of the radio source. It is virtually certain that this is the most distant system of galaxies yet identified by any means. From the considerations of Section 7, it is likely that this system is at a redshift of about 2.

Our conclusion is that the standard interpretation of the content of these samples of radio sources is correct and that by this means we can select radio galaxies and quasars with very large redshifts. It is important that it is not only quasars which can now be discovered at these very large redshifts but galaxies as well in which the light is dominated by ordinary starlight.

6. The redshifts of the radio source identifications

The next crucial step is the measurement of the optical spectra of the identifications and the determination of their redshifts. The measurements of the redshifts of the nearby giant elliptical galaxies were straightforward, the characteristic stellar absorption features of calcium, magnesium and sodium being observed. Frequently the galaxies also exhibited emission lines of [OII], [OIII] and strong Hα and *H*β. In the case of the N-galaxies, broad emission-line features were observed, particulary lines of the Balmer series and other permitted lines, similar in a number of ways to the Seyfert I galaxies.

The discovery of the quasars provided a rich spectroscopic harvest. Their spectra exhibit many strong, broad emission lines and, because of their large redshifts, ultraviolet lines such as CIII], [CIV], Lα and many other high-excitation lines were observed. The quasar industry was particularly productive because, although they were

almost all objects with apparent magnitudes in the range 16 to 19 magnitude, the strong emission lines made them easy and exciting targets spectroscopically.

For objects brighter than about 19th magnitude, the measurement of redshifts and spectra is a relatively straightforward procedure if efficient spectrographs on large optical reflectors are used. By the late 1970s, however, many faint radio galaxy identifications had been made with objects fainter than 19 mag and because they mostly seemed to be associated with apparently normal giant elliptical galaxies, they were unattractive and difficult objects for spectroscopic study. One fact, however, was established towards the end of the 1970s which made this enterprise not entirely without hope. Hine & Longair (1979) showed that, with increasing radio luminosity, there is a much greater probability that the radio galaxies will exhibit strong emission-line spectra. Since samples such as the 3CR sample are flux-density-limited, the most distant and hence the faintest galaxies are also the most luminous radio sources.

By a stroke of good fortune, this expectation has been more than fulfilled, largely through the efforts of Spinrad and his collaborators. From the late 1970s to date, they have measured the spectra of the very faint 3CR radio galaxies, first with the Lick 3-metre telescope and then with the Kitt-Peak 4-metre telescope (see Spinrad & Djorgovski 1984 and references therein). Although there are some important exceptions, the very faint galaxy identifications have very strong, narrow emission-line spectra, the lines of [OII] and [OIII] being particularly prominent. Very often the lines of[NeIII]and [Nev] are also observed as is shown in the example of 3C22 which has a redshift of 0.96 (Fig. 9—Perryman *et al*. 1984). The result has been that for the 3CR sample of radio sources, the redshift measurements are more than 90 per cent complete for more than half of the northern sky, the redshifts ranging up to 1.6 .For comparison, it should be recalled that optical studies of the brightest galaxies in clusters extend only to redshifts of about 1. Therefore this sample of radio galaxies is the most distant sample of objects studied so far in which the light is dominated by starlight.

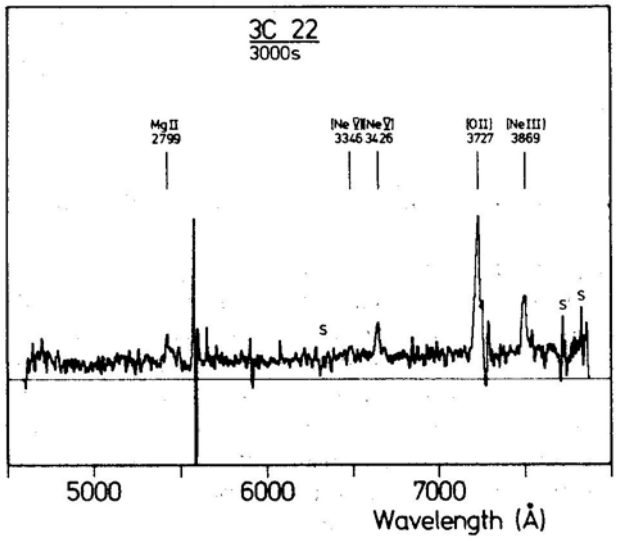


Figure 9. The spectrum of the radio galaxy 3C 22 (Perryman *et al.* 1984).

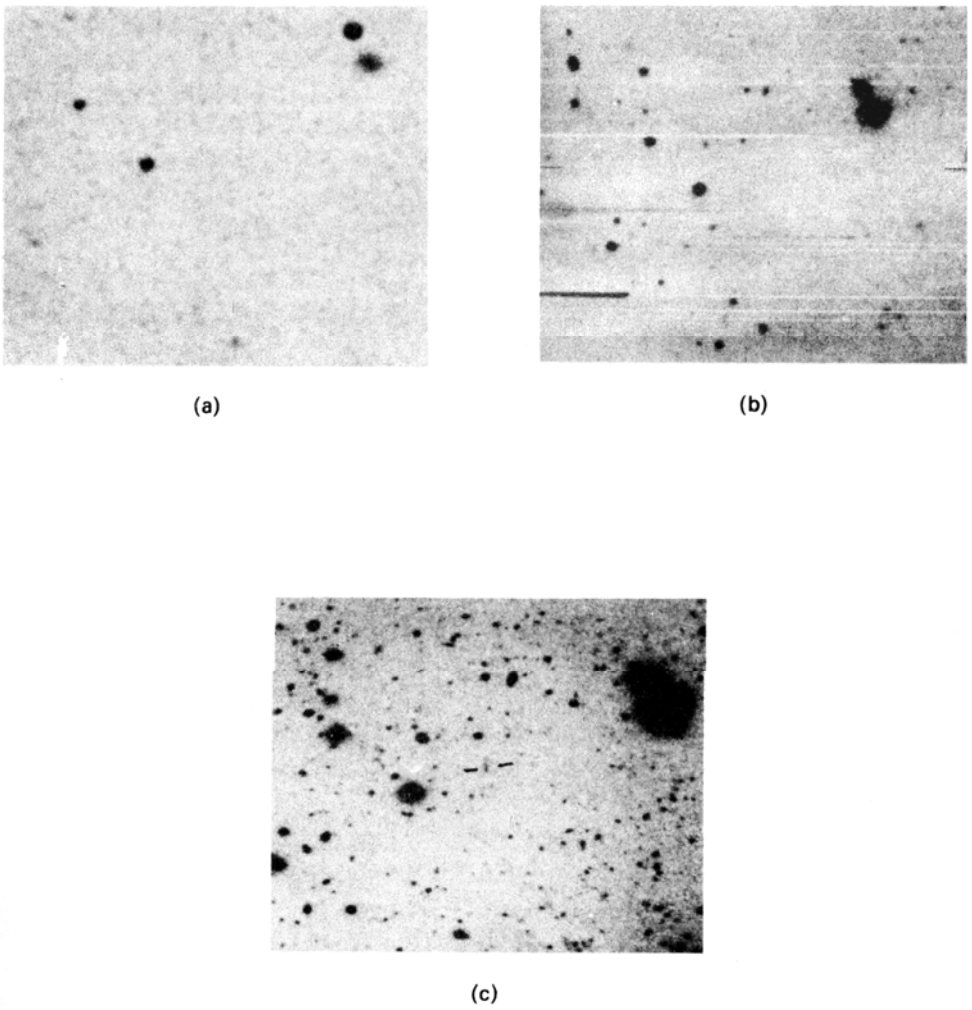


Figure 8. The optical field of the radio source 1056 + 39 observed with: (a) the 48-inch Sky Survey plates, (b) the 5-metre telescope with the PFUEI CCD camera in a 5 minute exposure, (c) the 5-metre telescope with the new 4-array CCD camera in a 1-hour exposure.

The redshift distribution for a complete sample of 3CR radio galaxies and quasars is shown in Fig. 10. This figure illustrates another remarkable feature of the redshift distribution of these radio sources—the redshift distributions of the radio galaxies and quasars in the 3CR sample are similar at redshifts $z \gtrsim 0.4$. Thus, as was inferred indirectly from studies of the counts alone, the quasars and radio galaxies span more or less the same redshift range and there must be a close astrophysical relation between their properties.

One other point is worth noting. In many ways, the radio galaxies which have these very large redshifts form a distinct class of astrophysical objects. The emission-line spectra are very strong but the lines are narrow. In a number of cases, the spectrograms show that the [OII] emission originates in an extended region which can be significantly larger than the image of the galaxy. These properties are quite different from those of the emission-line regions in quasars and N-galaxies in which the emission originates very close to the galactic nucleus itself. The excitation process for the strong, narrow emission-line regions is not established, possibilities including the photoionization by a dilute ultraviolet continuum and collisional excitation by shock waves. Whichever is correct, there is a strong suggestion that the emission results from gas liberated in the interaction between galaxies which may be the gas which ultimately fuels the inner regions of the active nucleus. This in turn may be responsible for generation of the fluxes of relativistic electrons which power the strong radio source phenomenon.

It is apparent that a great deal can be learned from the relatively straightforward processes of identification and spectroscopy of these galaxies.

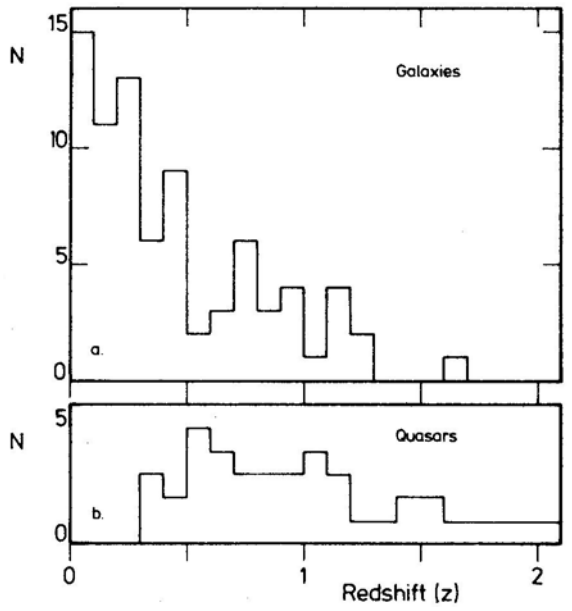


Figure 10. The redshift distribution for (a) radio galaxies and (b) quasars in the 3CR samples studied Lilly & Longair (1985). All the quasars have known redshifts. In the case of the radio galaxies, only 4 in unobscured regions lack redshifts.

7. The astrophysical significance of surveys of the optical and infrared properties of radio sources

We can identify at least three separate themes which run through this story. The first is the fact that the process of identifying extragalactic radio sources has resulted in the discovery of new classes of objects which are of great importance in understanding high-energy processes in astrophysical objects. Perhaps the most striking example is that of quasars and BL-Lac objects. The most recent is that of the strong, narrow-line emission galaxies which seem to be the dominant radio galaxy population at large redshifts.

The second aspect concerns the cosmological evolution of the radio source population as a whole. It must be recalled that the samples of sources which have now been almost completely identified are quite small and refer to high flux densities. To determine the evolutionary history of the whole radio source population requires much larger identified samples at a wider range of flux densities. This is a major undertaking as can be appreciated from the amount of effort which has been necessary to complete the identifications of the samples described above. Some of these programmes are now well underway but it is a time-consuming task to obtain identifications and optical spectra for samples of objects which are fainter than 20 mag.

Despite the incompleteness of the data, it has been possible to make considerable headway in delineating the evolution of the overall properties of the radio source population. The most recent models of this evolution have been described by Peacock (1985) who has elaborated the procedures outlined by Peacock & Gull (1981). The new generation of models can account for all the identification and redshift data which are now available. A sample of his results is shown in Fig. 11. These show how the

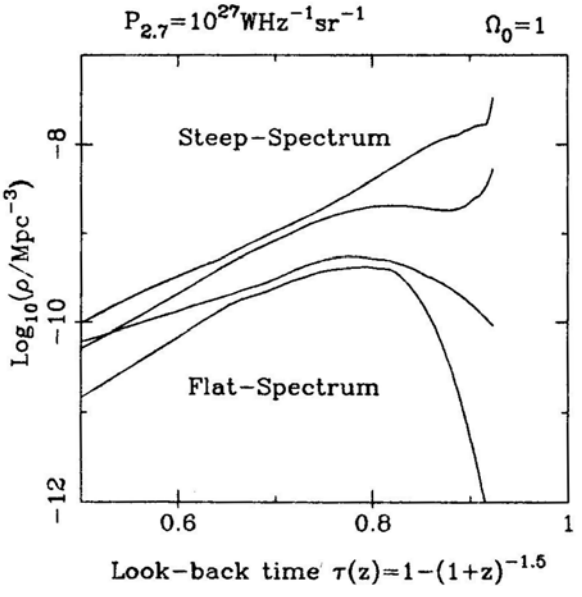


Figure 11. Sample results of the modelling of the evolution of the radio source population by Peacock (1985). The diagrams show separately the evolution functions for steep and flat spectrum radio sources.

comoving space density of radio sources with steep and flat radio spectra have evolved with cosmic epoch (or redshift). They show the very rapid changes in the probability of radio source events as a function of cosmic epoch. A second important point is the fact that the models show evidence for the convergence of the evolution of the source population at large redshifts. This is particularly striking for the flat-spectrum sample selected at 2.7 GHz where all the models show a decrease in the comoving space density of objects at redshifts greater than about 2.5. This result has important implications for the evolution of high-energy activity as the universe evolves. Apparently, the epoch of maximum high-energy astrophysical activity occurred in the relatively recent past and it is intriguing to speculate how this is related to the processes of the origin and evolution of galaxies in general.

The third aspect of these studies is the information they provide about the properties of galaxies in general at epochs earlier than the present. The galaxies associated with extragalactic radio sources are amongst the most luminous in the universe, being similar in absolute magnitude to the brightest galaxies in clusters. This continues to be the case even for the faintest radio galaxies known. The range of absolute magnitude for the 3CR radio galaxies is remarkably narrow, particularly when infrared *K* absolute magnitudes are considered. The most recent determination of the redshift-magnitude relation in the *K* waveband (2.2 μm by Lilly & Longair (1985) is shown in Fig. 12. The dispersion about the mean relation for the radio galaxies in which the light is dominated by star-light remains unchanged with redshift and has standard deviation about 0.5 mag. In the infrared, the dominant stellar population is old stars and therefore the integrated infrared luminosity averages over the long timescale behaviour of the galaxy. A similar diagram using optical magnitudes shows a much greater scatter at large redshifts. We have shown that this occurs because the optical waveband is much more sensitive to the presence of even small numbers of very young blue stars and we have direct evidence that this phenomenon strongly influences the optical-infrared colours

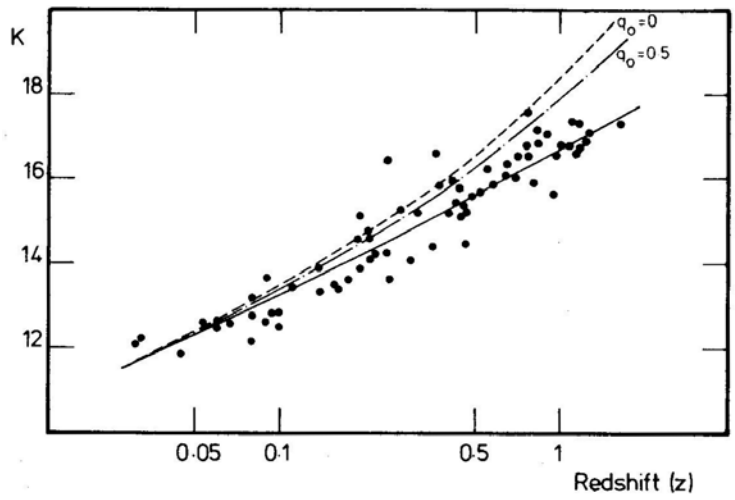


Figure 12. The *K* magnitude-redshift relation of Lilly & Longair (1985). The dashed lines show the expectations of the standard world models and the solid line a best fit to the data of a uniform world model which incorporates evolution of the stellar population of the radio galaxies with cosmic epoch.

of the faint galaxies. Our study of the complete 3CR sample of radio galaxies has demonstrated that a number of those at large redshifts possess young stellar populations in the sense that they show an extended ultraviolet excess in the rest frame of the galaxy. This excess is correlated with the strength of the strong, narrow emission lines suggesting that these phenomena have a common origin. We believe that both are evidences of bursts of star formation which may be associated with the interaction with nearby companions. This may in turn be responsible for fuelling the energy source in the nucleus of the radio galaxy.

Underlying this behaviour, the infrared redshift-magnitude relation describes the properties of the bulk of the stellar population of the galaxies. The most important result is that this relation maintains a small dispersion out to the largest redshifts and that the mean absolute magnitude of the galaxies is about a magnitude brighter at a redshift of 1 than would be expected in any of the standard cosmological models which have values of the deceleration parameter in the range 0 to 1. This phenomenon is, however, entirely consistent with the passive evolution of giant elliptical galaxies. At a redshift of 1, we are looking back more than half the present age of the universe and therefore the galaxies are expected to be brighter as stars more massive than the Sun evolve off the main sequence and onto the giant branch. It turns out that for reasonable assumptions the amount of evolution expected is about 1 mag at a redshift of 1. We believe that we have now seen this effect in these radio galaxies. This is the first time that it has been possible to study directly the stellar populations of galaxies at redshifts of 1 and greater.

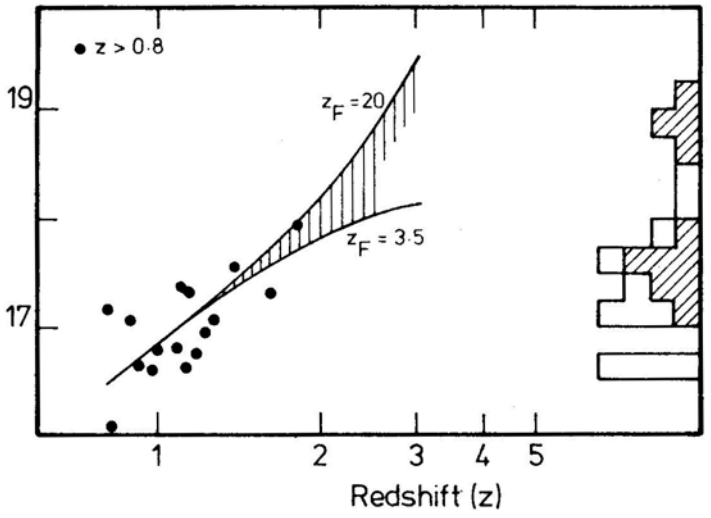


Figure 13. The *K* magnitude-redshift relation for large redshift radio galaxies. Also shown are a range of reasonable extrapolations of the observed relation depending upon the epoch at which the galaxies first formed (labelled by the formation redshift z_F). The solid dots show the *K* magnitudes of radio galaxies with measured redshifts. On the vertical axis is shown the apparent magnitude distribution for the faint infrared detections in the 1 Jy radio survey. The shaded boxes are the optically unidentified sources which we detected in our infrared survey. It is inferred that a number of these radio galaxies are likely to have redshifts greater than 2 (Lilly, Longair & Allington-Smith 1985).

Perhaps even more intriguing is the nature of some of the faintest galaxies detected in our infrared survey of the radio galaxies in the 1 Jy sample. Fig. 13 shows the redshift-apparent magnitude relation for radio galaxies with the apparent magnitude distribution of the faint infrared detections indicated on the vertical axis. It can be seen that the faintest of these radio galaxies must have redshifts of 2 to 3 or possibly even greater on the basis of reasonable extrapolations of the *K* redshift-magnitude relation. This is the same range of redshifts which is currently spanned by the quasars.

Equally important is the fact that we have measured lower limits to the optical –infrared colours for all these distant radio galaxies and some of them at least must have integrated stellar spectra which are consistent with the passive evolution of the stellar population of these galaxies. These galaxies must have redshifts of 2 to 3, and must have formed much earlier than the epoch at which they are now observed. Our calculations suggest that the epoch of formation of these galaxies must have taken place earlier than the epoch corresponding to redshifts of about 3 or 4. The importance of extending these studies to much larger samples of sources is obvious.

References

Allington-Smith, J. R. 1982, *Mon. Not. R. astr. Soc.*, **199**, 611.

Allington-Smith, J. R., Perryman, M. A. C., Longair, M. S., Gunn, J. E., Westphal, J. A. 1982, *Mon. Not. R. astr. Soc.*, **201**, 331.

Baade, W., Minkowski, R. 1954, *Astrophys. J.*, **119**, 206.

Beichman, C. A., Neugebauer, G., Soifer, B. T., Wootten, H. A., Roellig, T., Harvey, P. M. 1981, *Nature*, **293**, 711.

Bennett, A. S. 1962, *Mem. R. astr. Soc*., **68**, 163.

Bolton, J. G., Stanley, G. J., Slee, O. B. 1949, *Nature*, **164**, 101.

Bridle, A. H., Davis, M. M., Fomalont, E. B., Lequeux, L. 1972, *Astr. J.*, **77**,405.

Clark, M. E. 1964, *Mon. Not. R. astr. Soc.*, **127**, 405.

Edge, D. O., Shakeshaft, J. R., McAdam, W. B., Baldwin, J. E., Archer, S. 1959, *Mem. R. astr. Soc.*, **68**, 37.

Fomalont, E. B., Matthews, T. A., Morris, D., Wyndham, J. D. 1964, *Astr. J.*, **69**, 772.

Grueff, G., Vigotti, M., Wall, J. V., Benn, C. R. 1984, *Mon. Not. R. astr. Soc.*, **206**, 475.

Gunn, J. E., Hoessel, J. G., Westphal, J. A., Perryman, M. A. C., Longair, M. S. 1981, *Mon. Not. R. astr. Soc.*, **194**, 111.

Gunn, J. E., Lilly, S. J. 1985, in preparation.

Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F., Collins, R. A. 1968, *Nature*, **217**, 709.

Hine, R. G., Longair, M. S. 1979, *Mon. Not. R. astr. Soc.*, **188**, 111.

Jenkins, C. J., Pooley, G. G., Riley, J. M. 1977, *Mem. R. astr. Soc.*, **84**, 61.

Jennison, R. C., Das Gupta, M. K. 1953, *Nature*, **172**, 996.

Kristian, J., Sandage, A. R., Katem, B. 1974, *Astrophys. J.*, **191**, 43.

Kristian, J., Sandage, A. R., Katem, B. 1978, *Astrophys. J.*, **219**, 803.

Kühr, H., Witzel, A., Pauliny-Toth, I. I. K., Nauber, V. 1981, *Astr. Astrophys. Suppl. Ser.*, **45**, 367.

Laing, R. A., Longair, M. S., Riley, J. M., Kibblewhite, E. J., Gunn, J. E. 1978, *Mon. Not. R. astr. Soc.*, **183**, 547.

Laing, R. A., Riley, J. M., Longair, M. S. 1983, *Mon. Not. R. astr. Soc.*, **204**, 151.

Lilly, S. J., Longair, M. S. 1982, *Mon. Not. R. astr. Soc.*, **199**, 1053.

Lilly, S. J., Longair, M. S. 1985, *Mon. Not. R. astr. Soc.* (in press).

Lilly, S. J., Longair, M. S., Allington-Smith, J. R. 1985, *Mon. Not. R. astr. Soc.* (submitted).

Longair, M. S. 1965, *Mon. Not. R. astr. Soc.*, **129**, 419.

Longair, M. S. 1975, *Mon. Not. R. astr. Soc.*, **173**, 309.

Longair, M. S. 1978, in '*Observational Cosmology*', 8th Saas-Fee School of Astronomy and Astrophysics, Eds A. Maeder, L. Martinet & G. Tammann, Geneva Obs., p. 127.

Longair, M. S., Gunn, J. E. 1975, *Mon Not. R. astr. Soc.*, **170**, 121.

MacLeod, J. M., Andrew, B. H. 1968, *Astrophys. Lett.*, **1**, 243.
Matthews, T. A., Morgan, W. W., Schmidt, M. 1964, *Astrophys. J.*, **140**, 35.
Matthews, T. A., Sandage, A. R. 1963, *Astrophys. J.*, **138,** 30.
Mitchell, K. J. 1983, *Dissertation*, Pennsylvania State Univ.
Peacock, J. A. 1985, in preparation.
Peacock, J. A., Gull, S. F. 1981, *Mon. Not. R. astr. Soc.*, **196**, 611.
Peacock, J. A., Perryman, M. A. C., Longair, M. S., Gunn, J. E., Westphal, J. A. 1981, *Mon. Not. R. astr. Soc.*, **194**, 601.
Perryman, M. A. C., Lilly, S. J., Longair, M. S., Downes, A. J. B. 1984, *Mon. Not. R. astr. Soc.*, **209,** 159.
Perryman, M. A. C., Longair, M. S., Allington-Smith, J. R., Fielden, J. 1982, *Mon. Not. R. astr. Soc.*, **201**, 957.
Porcas, R. W., Urry, C. M., Browne, I. W. A., Cohen, A. M., Daintree, E. J., Walsh, D. 1980, *Mon. Not. R. astr. Soc.*, **191**, 607.
Read, R. B. 1963, *Astrophys. J.*, **138,** 1.
Rieke, G. H., Lebofsky, M. J., Kinman, T. D. 1979, *Astrophys. J.*, **232,** L151.
Riley, J. M., Eales, S. A., Baldwin, J. E. 1984, *Mon. Not. astr. Soc.*, **209,** 641.
Riley, J. M., Longair, M. S., Gunn, J. E. 1980, *Mon. Not. R. astr. Soc.*, **192,** 233.
Schmidt, M. 1963, *Nature*, **197**, 1040.
Schmitt, J. L. 1968, *Nature*, **218**, 663.
Smith, F. G. 1951, *Nature*, **168**, 555.
Smith, F. G. 1952, *Mon. Not. R. astr. Soc.*, **112,** 497.
Spinrad, H., Djorgovsky, S. 1984, Preprint.
Stephenson, C. B., Sanduleak, N. 1977, *Astrophys. J. Suppl. Ser.*, **33,** 459.
Stockton, A. 1980, *Astrophys. J.*, **242,** L141.
Wall, J. V., Benn, C. R. 1982, in *IAU Symp.* 97: *Extragalactic Radio Sources*, Eds D. S. Heeschen & C. M. Wade, D. Reidel, Dordrecht, p. 441.
Wall, J. V., Peacock, J. A. 1985, in preparation.
Walsh, D., Carswell, R. F., Weymann, R. J. 1979, *Nature*, **279,** 381.
Windhorst, R. 1984, *Faint Radio Galaxy Populations*, PhD Dissertation, Beugelsdijk Leiden B. V.
Wyndham, J. D. 1965, *Astr. J.*, **70**, 384.
Young, P., Gunn, J. E., Kristian, J., Oke, J. B., Westphal, J. A. 1980, *Astrophys. J.*, **241**, 507.

J. Astrophys. Astr. (1984) **5,** 369–388

Arrival-Time Analysis for a Millisecond Pulsar

Roger Blandford, Ramesh Narayan* & Roger W. Romani
Theoretical Astrophysics, California Institute of Technology, Pasadena CA 91125 USA

(*Invited article*)

Abstract. Arrival times from a fast, quiet pulsar can be used to obtain accurate determinations of pulsar parameters. In the case of the millisecond pulsar, PSR 1937 + 214, the remarkably small rms residual to the timing fit indicates that precise measurements of position, proper motion and perhaps even trigonometric parallax will be possible (Backer 1984). The variances in these parameters, however, will depend strongly on the nature of the underlying noise spectrum. We demonstrate that for very red spectra *i.e.* those dominated by low-frequency noise, the uncertainties can be larger than the present estimates (based on a white-noise model) and can even grow with the observation period. The possibility of improved parameter estimation through pre-whitening' the data and the application of these results to other pulsar observations are briefly discussed. The post-fit rms residual of PSR 1937 + 214 may be used to limit the energy density of a gravitational radiation background at periods of a few months to years. However, fitting the pulsar position and pulse-emission times filters out significant amounts of residual power, especially for observation periods of less than three years. Consequently the present upper bound on the energy density of gravitational waves $\Omega_g \lesssim 3 \times 10^{-4} R_{\mu s}^2$, though already more stringent than any other available, is not as restrictive as had been previously estimated. The present limit is insufficient to exclude scenarios which use primordial cosmic strings for galaxy formation, but should improve rapidly with time.

Key words: millisecond pulsar—arrival times—gravitational background radiation

1. Introduction

The discovery of the millisecond pulsar, PSR 1937 + 214 (Backer *et al.* 1982), has opened up several new possibilities in the study of pulsar timing. The high-spin frequency (642 Hz) and the apparently small intrinsic timing noise combine to make this object an excellent clock. Arrival times have been monitored with an accuracy exceeding 1 μs over periods of two years (Backer, Kulkarni & Taylor 1983; Backer 1984; Davis *et al.* 1984) and it appears that we are already limited by the accuracy of planetary ephemerides and the stability of atomic clocks. As has been pointed out by several authors, PSR 1937 + 214 can be used as a sensitive detector of low-frequency gravitational radiation (*e.g.* Hogan & Rees 1984), as a probe of electron-density

* On leave from: Raman Research Institute, Bangalore 560080, India.

fluctuations in the interstellar medium (Armstrong 1984, Cordes & Stinebring 1984, Blandford & Narayan 1984a, b) and perhaps for the study of neutron-star seismology (*e.g.* Cordes & Greenstein 1981). Our purpose in the present paper is twofold. Firstly, we wish to develop the analysis of pulsar arrival times so as to estimate the sensitivity of fast pulsars as detectors of gravitational radiation and dispersion-measure fluctuations under the assumption that they remain as good clocks as is indicated by present observations. Secondly, we explore the limits to the use of accurate arrival times to measure pulsar spin-down, position, proper motion and parallax distance, in the presence of a particular noise spectrum.

In Section 2, we give a general analysis of the fitting of residuals in the measured pulse arrival times with an assumed timing model that includes the pulsar phase, period and period derivative, together with its position, proper motion and parallax. We specialize to the case of a stationary noise source and consider in Section 3 the particular case of a power-law power spectrum. We give estimates of the accuracy with which the pulsar parameters and the noise strength can be determined with standard least squares and suggest that 'pre-whitening' could lead to improvement if the noise spectrum is very 'red' (*i.e.* noise-power increasing strongly towards low frequencies). In Section 4, we apply our results to PSR 1937 + 214 and give quantitative estimates of its sensitivity to three potential sources of noise—gravitational waves, interstellar electron-density fluctuations and intrinsic pulsar noise. Applications to other pulsars are discussed in Section 5.

2. Analysis of timing residuals

Measured sequences of pulsar arrival times are conventionally fitted to a linear expression, whose parameters (essentially the corrections to various unknown quantities) are determined by the method of least squares. Unfortunately, contributions to the residuals that have quite different physical origins—for example the response to a gravitational wave of period exceeding several years and the slowing down of the pulsar's spin—can have very large covariances and are therefore not easily separated. In this section we describe a method for estimating the true sensitivity of a rapid pulsar to gravitational radiation and interstellar effects. We do this by analysing a simple timing model that includes all of the essential sources of covariance, omitting some inessential terms that would otherwise lengthen the analysis. We emphasize that the timing model has been chosen purely for analytical convenience and is not to be used in fitting real data, which should be fitted to a model based on a complete ephemeris, including general-relativistic corrections (*e.g.* Romani & Taylor 1983; Backer 1984).

In our model we assume that a point earth describes a circular orbit of known radius α about the solar system barycentre and so the transverse Doppler shift and gravitational redshift terms represent constant offsets (*e.g.* Manchester & Taylor 1977). This is equivalent to assuming that we possess a sufficiently accurate planetary ephemeris determined by independent means so that errors in the telescope position relative to the barycentre do not contribute to the timing noise. We discuss this approximation further in Section 4. We also assume that the pulsar position on the sky is known well enough that a linear fit to its true position, proper motion and distance is adequate.

We restrict our attention to stationary sources of noise that can be completely described by a power spectrum. In order to keep the algebra manageable, we further

idealize the observations by assuming that they are uniformly spaced and extend over an integral number of years starting at a particular epoch which we shall specify. This restriction greatly simplifies the theory and will slightly overestimate the sensitivity of the timing data if our results are applied to non-uniform observations taken over a non-integral number of years.

For a pulsar with parallax $p = a/d$ (with d the pulsar's distance from the barycentre), whose heliocentric latitude and longitude measured from the vernal equinox are respectively β and λ ,the distance a pulse travels to earth is given by

$$\begin{aligned} D &= [d^2 \sin^2\beta + \{d\cos\beta + a\cos(\phi-\lambda)\}^2 + a^2\sin^2(\phi-\lambda)]^{1/2} \\ &= a[\cos\beta\cos(\phi-\lambda) - \frac{1}{4}p\cos^2\beta\cos 2(\phi-\lambda)], \end{aligned} \tag{2.1}$$

where ϕ is the earth's mean anomaly, and we have dropped some constant terms. Let the small errors in the pulsar latitude and longitude be

$$\delta\beta = \delta\beta_0 + \delta\mu_\beta t \tag{2.2}$$

$$\delta\lambda = \delta\lambda_0 + \delta\mu_\lambda t \tag{2.3}$$

where $\delta\mu_{\beta,\lambda}$ are the two components of the proper motion and t is the time of observation, which we measure in years from the midpoint of the observation, fixed to occur at an anomaly $\phi = \lambda + \pi/4$.

As usual, we fit the time of emission of the pulses to a quadratic function parametrized by the unknown phase, frequency and frequency derivative. Ignoring constant additive and multiplicative factors, the pulse arrival time is given by the emission time plus the variable part of the propagation time to earth, D/c. We define the timing residual $R(t)$ to be the difference between the observed arrival time of a pulse and the arrival time predicted on the basis of our best guesses to the unknown parameters. These residuals are fitted to an expression that is linear in the corrections to the unknown parameters, *i.e.*,

$$R(t) = \sum_{a=1}^{8} \alpha_a \psi_a(t) \tag{2.4}$$

where

$$\begin{aligned}
\alpha_1 &= -K, & \psi_1 &= 1, \\
\alpha_2 &= -\frac{\delta\nu}{\nu}, & \psi_2 &= t, \\
\alpha_3 &= -\frac{\delta\dot{\nu}}{2\nu}, & \psi_3 &= t^2, \\
\alpha_4 &= \frac{a}{c\sqrt{2}}(\delta\beta_0\sin\beta + \delta\lambda_0\cos\beta), & \psi_4 &= \sin 2\pi t, \\
\alpha_5 &= \frac{a}{c\sqrt{2}}(-\delta\beta_0\sin\beta + \delta\lambda_0\cos\beta), & \psi_5 &= \cos 2\pi t, \\
\alpha_6 &= \frac{a}{c\sqrt{2}}(-\delta\mu_\beta\sin\beta + \delta\mu_\lambda\cos\beta), & \psi_6 &= t\cos 2\pi t, \\
\alpha_7 &= \frac{a}{c\sqrt{2}}(\delta\mu_\beta\sin\beta + \delta\mu_\lambda\cos\beta), & \psi_7 &= t\sin 2\pi t, \\
\alpha_8 &= \frac{a}{4c}p\cos^2\beta, & \psi_8 &= \sin 4\pi t.
\end{aligned} \tag{2.5}$$

Timing parallax has not so far been measured in any pulsar. Therefore, we have repeated our calculations for a linear combination of 7 parameters, leaving out ψ_8.

Now suppose that we measure n equally spaced and comparably accurate arrival times each year for a total of N years, *i.e.*, we have Nn residuals $R_i = R(t_i)$, $i = 1, Nn$. We wish to obtain least squares estimates of the parameters α_a. As there are 8 independent parameters to fit, it turns out to be algebraically easier to diagonalize the normal equations by introducing a set of orthonormal fitting functions, $\psi'_{ai} = \psi'_a(t_i)$, which are linear combinations of the original ψ_i, *i.e.*,

$$R(t) = \sum_{a=1}^{8} \alpha'_a \psi'_a(t), \qquad \psi'_{ai} = L_{ab}\psi_{bi}, \tag{2.6}$$

where

$$\sum_{i=1}^{Nn} \psi'_{ai}\psi'_{bi} = \delta_{ab}. \tag{2.7}$$

In fact, the number of observations is usually so large that the sum in Equation (2.7) can be approximated by an integral over the observing period; *i.e.*, $\sum_{i=1}^{Nn} \sim \eta \int_{-N/2}^{N/2} \mathrm{d}t$. A convenient choice of orthonormal functions for the case in point is defined uniquely through the Gram-Schmidt orthogonalization procedure:

$$\begin{aligned}
\psi'_1 &= L_{11}, \\
\psi'_2 &= L_{22}t, \\
\psi'_3 &= L_{31} + L_{33}t^2, \\
\psi'_4 &= L_{42}t + L_{44}\sin 2\pi t, \\
\psi'_5 &= L_{51} + L_{53}t^2 + L_{55}\cos 2\pi t, \\
\psi'_6 &= L_{62}t + L_{64}\sin 2\pi t + L_{66}t\cos 2\pi t, \\
\psi'_7 &= L_{71} + L_{73}t^2 + L_{75}\cos 2\pi t + L_{77}t\sin 2\pi t, \\
\psi'_8 &= L_{82}t + L_{84}\sin 2\pi t + L_{86}\cos 2\pi t + L_{88}\sin 4\pi t,
\end{aligned} \tag{2.8}$$

where the L_{ab} *are* constants that depend upon N. The best-fitting primed parameters, α'_a, are given by the solution of the normal equations

$$\alpha'_a = \sum_{i=1}^{Nn} R_i\psi'_{ai}. \tag{2.9}$$

Now suppose that the residuals are entirely due to timing noise generated by a stationary power spectrum $P(f)$ so that

$$\langle R_iR_j \rangle = \int_0^\infty \mathrm{d}f\, P(f)\cos 2\pi f(t_i - t_j) \tag{2.10}$$

where $\langle\ \rangle$ signifies an ensemble average over many realizations of the fitting procedure. We obtain an expression for the mean-square residual after subtracting the best-fitting solution to Equation (2.6)

$$\begin{aligned}
\overline{R^2(t)} &= \frac{1}{Nn}\sum_i\sum_j \langle R_iR_j\rangle\left[\delta_{ij} - \sum_a \psi'_{ai}\psi'_{aj}\right] \\
&= \int_0^\infty \mathrm{d}f\, P(f)T(f)
\end{aligned} \tag{2.11}$$

were the transmission or filter function, $T(f)$, is given by

$$T(f) = 1 - \frac{1}{N}\sum_a \tilde{\psi}'_a(f)\tilde{\psi}'^*_a(f). \tag{2.12}$$

and

$$\tilde{\psi}'_a(f) = \int_{-N/2}^{N/2} \mathrm{d}t\, \psi'_a(t) \exp(2\pi i f t) \tag{2.13}$$

are the Fourier transforms of the orthonormal fitting functions.

Equation (2.11) is an expression of the fact that when we try to detect background timing noise, much of this noise will be filtered out by the fit for the pulsar period, position and other parameters. We can think of the factor $T(f)$ as being a transmission coefficient for the noise and the individual factors $|\tilde{\tilde{\Psi}}'_a(f)|^2$ as being absorption coefficients associated with the individual fitting functions. The latter are presented for $N = 3$ in Figs 1 and 2 and the transmission function $T(f)$ is presented for $N = 1$, 3, 10 in Fig. 3. The pulsar will thus be a less sensitive detector of the noise than if we had prior knowledge of the exact phase, period, position, *etc.* (in which case the filter function is $T(f) = 1$).

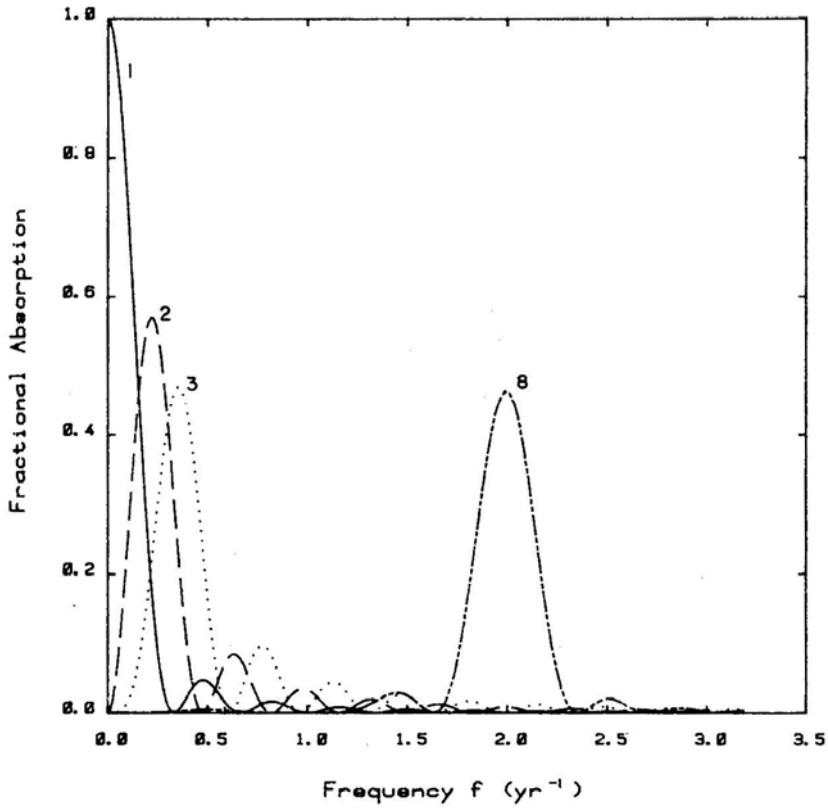


Figure 1. Absorption coefficients $|\tilde{\tilde{\Psi}}'_a|^2$ for $a = 1$, 2, 3 and 8 at $N = 3$ years. The first three functions generate the dip near the origin in Fig. 3, and the last function generates the feature at $f = 2$ yr^{-1}.

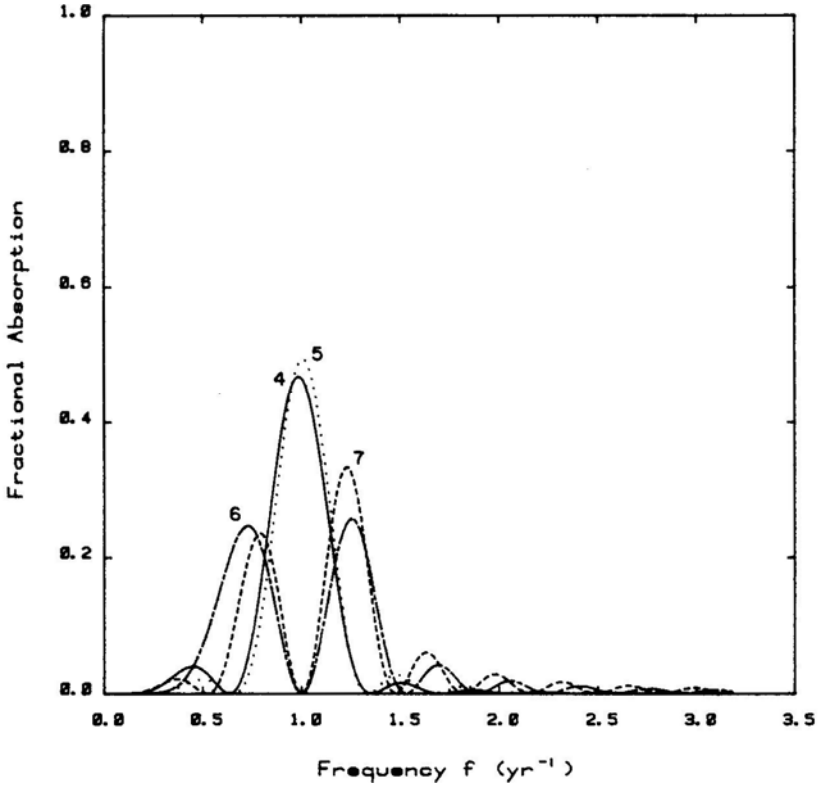


Figure 2. Absorption coefficients $|\tilde{\Psi}'_a|^2$ for a = 4–7 at N = 3 years. The functions 4 and 5 are largely due to position errors while 6 and 7 are dominated by the proper motion terms. These generate the minimum at $f = 1\ \mathrm{yr}^{-1}$ in Fig. 3.

We can also use Equation (2.6) to estimate the covariance matrix of the parameters α'_a after performing a least-squares fit to the measured arrival times

$$\langle \delta\alpha'_a \delta\alpha'_b \rangle = \int_0^\infty \mathrm{d}f\, P(f) \tilde{\psi}'_a(f) \tilde{\psi}'^*_b(f) \tag{2.14a}$$

or

$$\langle \delta\alpha'_a \delta\alpha'_b \rangle = \left[\frac{\int_0^\infty \mathrm{d}f\, P(f) \tilde{\psi}'_a(f) \tilde{\psi}'^*_b(f)}{\int_0^\infty \mathrm{d}f\, P(f) T(f)} \right] \bar{R}^2. \tag{2.14b}$$

Note that the quantity in square brackets is independent of the strength of the noise and depends only on the shape of its spectrum. Finally, the covariance matrix of the original fitting parameters is given by

$$\langle \delta\alpha_a \delta\alpha_b \rangle = \sum_a \sum_b L_{ca} L_{db} \langle \delta\alpha'_c \delta\alpha'_d \rangle. \tag{2.15}$$

Equations (2.14) and (2.15) allow us to make an unbiased estimate of the expected error in the various fitting parameters in terms of either the noise strength or the residual. However, as we discuss further in Section 3.4 below, we may be able to filter out much of the noise so as to obtain a much smaller variance for the unknown parameters. The

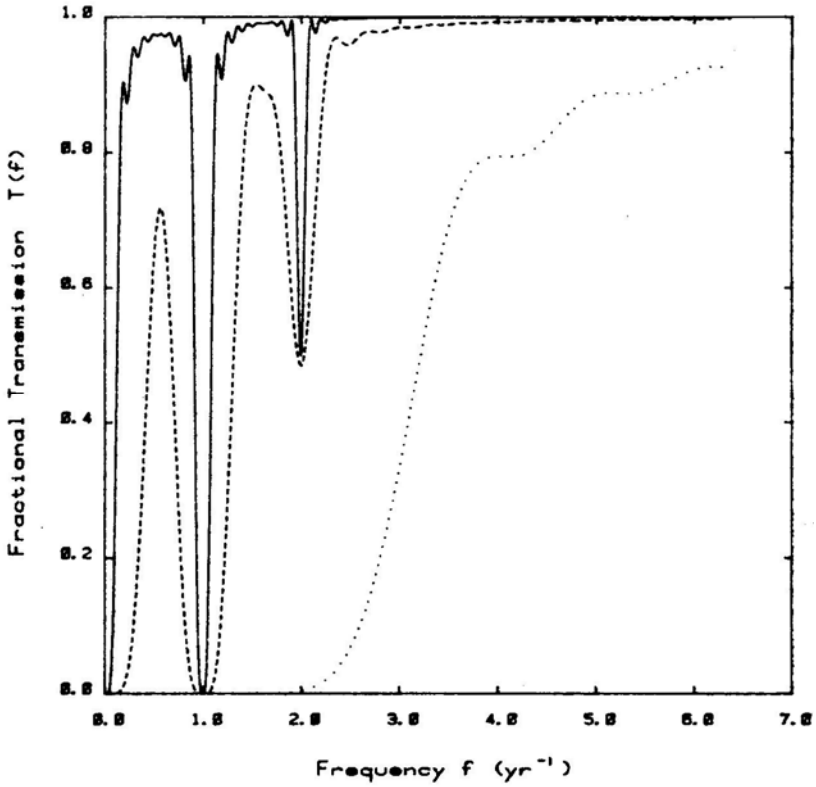


Figure 3. Transmission coefficient $T(f)$ defined in Equation (2.12) for an 8-parameter fit for N = 1 year (dotted line), 3 years (dashed line), 10 years (solid line). The dip near the origin corresponds to power removed by the polynomial fit, the dip at 1 yr^{-1} is from fitting position and proper motion and that at 2 yr^{-1} is due to parallax. As N increases, the three features become narrower (width $\propto 1/N$) showing that the corresponding sets of functions become more nearly orthogonal to one another.

usual variance estimated by standard least squares corresponds to the case of white noise, *i.e.*, $P(f)$ = constant.

3. Power-law noise spectra

3.1 *General Considerations*

We now assume that the noise spectrum has a power law form

$$P(f) = P_0 f^{-s}, \qquad f \geqslant 0. \tag{3.1}$$

P_0 is the noise power in waves with a period of around one year. We confine our attention to the exponents s = 0, 2, 3, 4, 5, 6 and data spans of N = 1, 2, 3, 5, 10 yr. The exponent s = 0 corresponds to white noise, which is the spectrum usually assumed (at least implicitly) when analysing the arrival times by least-squares fitting. It is appropriate when individual independent measurement errors dominate other sources

of noise. 'Red' spectra with slopes s = 2, 4, 6 correspond to random walks in phase, frequency and torque respectively. Spectra with slopes s = 3, 5 may be produced, respectively, by interstellar density fluctuations and a hypothetical background of primordial gravitational waves.

Our procedure is to compute the elements L_{ab} of the transformation matrix defined by (2.6) for each value of N and then to calculate the Fourier transforms of the orthonormal functions, $\tilde{\psi}'_a(f)$, by taking suitable linear combinations of the analytical Fourier transforms of the $\psi_a(t)$. Next, we evaluate the filter function $T(f)$ (Equation 2.12), and then compute the mean expected residual through Equation (2.11). In order to make contact with earlier work we express our results in terms of an equivalent filter which is 0 for $f < \beta/N$ and 1 for $f > \beta/N$. In other words, we determine β so that the calculated mean square residual $\bar{R}_2$ satisfies the relation

$$\bar{R}^2 = \int_0^\infty \mathrm{d}f P(f) T(f) = \int_{\beta/N}^{n/2} \mathrm{d}f P(f) \tag{3.2}$$

The upper cut-off in the frequency arises from the sampling theorem and is not important for red noise. The lower cut-off takes account of the fact that lower frequencies are fitted away by the polynomial fit and periods around 1 yr and 6 months are fitted by position/proper motion and parallax respectively. In the past β has been assumed to be ~ 1 (Detweiler 1979: Bertotti, Carr & Rees 1983; Romani & Taylor 1983), but no quantitative estimates have been reported to date.

We also compute the uncertainties in the various parameters α_1–α_8 and present each as the ratio, (variance)$^{1/2}$ per μs of post-fit rms residual. These can be converted to variance per unit power at 1 yr period, P_0, through Equations (3.1) and (3.2).

3.2 *White Noise*

To bring out the salient features of our formalism we first consider white noise, corresponding to s = 0. Calculations show that, for white noise with $n \gg 1$, β = 4 when all 8 parameters are fitted and β = 3.5 when parallax is not refined.

Consider next the variance in the position estimate of the pulsar. We can make the following approximate estimate. If N is sufficiently large, $\Psi_4(t)$ and $\Psi_5(t)$ are almost orthogonal to the other $\Psi_i(t)$. Then, the variance v_4 in the estimate of α_4 is approximately given by simplifying Equation (2.14a) to

$$v_4 = \langle \delta\alpha_4 \delta\alpha_4 \rangle = \frac{\int_0^{n/2} P(f) |\tilde{\psi}_4(f)|^2 \mathrm{d}f}{\left[2 \int_0^{n/2} |\tilde{\psi}_4(f)|^2 \mathrm{d}f \right]^2} . \tag{3.3}$$

The denominator is necessary because $\psi_4(t)$ is not normalized and the factor of 2 is because the integral has been restricted to positive f. There is a similar expression for v_5. Taking $P(f) = P_0$ for white noise and substituting

$$\tilde{\psi}_4(f) = \frac{iN}{2} \left[\frac{\sin \pi N(f-1)}{\pi N(f-1)} - \frac{\sin \pi N(f+1)}{\pi N(f+1)} \right], \tag{3.4}$$

$$\tilde{\psi}_5(f) = \frac{N}{2} \left[\frac{\sin \pi N(f-1)}{\pi N(f-1)} + \frac{\sin \pi N(f+1)}{\pi N(f+1)} \right], \tag{3.5}$$

we obtain

$$v_4 \sim v_5 \sim \frac{2\bar{R}^2}{nN}. \tag{3.6}$$

We thus recover the well-known result that the variance decreases inversely with the number of independent measurements. Substituting $a = 1.5 \times 10^{13}$ cm in (2.5) we thus have

$$\sin\beta\,(\delta\beta_0)_{\rm rms} \sim \cos\beta\,(\delta\lambda_0)_{\rm rms} \sim \frac{0.6\,(\bar{R}^2)^{1/2}_{\mu s}}{(nN)^{1/2}}\ \text{mas}. \tag{3.7}$$

More detailed calculations through the Gram-Schmidt orthogonalization procedure described in Section 2 confirm the coefficent as well as the scaling with n and N. The rms error in the proper motion is given by

$$\sin\beta\,(\delta\mu_\beta)_{\rm rms} \sim \cos\beta\,(\delta\mu_\lambda)_{\rm rms} \sim \frac{2\,(\bar{R}^2)^{1/2}_{\mu s}}{(nN^3)^{1/2}}\ \text{mas yr}^{-1}. \tag{3.8}$$

3.3 *Red Noise*

Red noise spectra have $s > 0$, *i.e.* the residuals are dominated by low-frequency noise. In the cases of interest, all the integrals converge rapidly at high/and so none of the results are sensitive to n so long as $n \gtrsim 10$. This is an important qualitative feature of red noise, showing that one cannot improve the precision of the refined parameters by increasing the number of measurements. As we demonstrate below, one does not gain by increasing the number of years of data either since the variances often increase as N increases.

Red noise has a divergent spectrum at low f. However, since the filter function $T(f) \propto f^6$ at. low f (for the present problem), the post-fit mean-square residual $\bar{\bar{R}}^2$ converges so long as $s < 7$. Equation (3.2) can now be written in the form

$$\bar{R}^2 = \frac{P_0}{(s-1)}\left(\frac{N}{\beta}\right)^{(s-1)} \tag{3.9}$$

where the upper limit in the integral should ideally be $n/2$ but has been set to ∞ (continuous sampling) because the integral converges rapidly, β has been evaluated for various values of N and s; the results are presented in Table 1. We give β for a 7-parameter fit (without parallax) for $N = 1, 2, 3, 5$ and also for an 8-parameter fit for $N = 5, 10$. Note that β is large, $\gtrsim 2$ for $N < 3$, showing that the parameter fit removes a substantial part of the noise. Our values of β are somewhat larger than those assumed by Bertotti, Carr & Rees (1983) and Hogan & Rees (1984).

Press (1975) and Lamb & Lamb (1976) have developed a least-squares analysis of pulsar timing noise in terms of a complete set of orthogonal polynomials, but considered only a white-noise spectrum. Our approach, which involves an orthogonalization of the functions relevant to physical parameters, can be extended to accommodate red-noise processes. Groth (1975a) and Cordes (1980) have analysed red-noise spectra as well, but employ a model in the time domain. This time series approach, in principle, has more information than is contained in the power spectrum alone; we

Table 1. Values of the effective spectral cut-off β (*cf.* Equation 3.2) corresponding to a 7-parameter fit (no parallax) for $N = 1, 2, 3, 5$ and an 8-parameter fit (including parallax) for $N = 5, 10$.

N/s	2	3	4	5	6
7-parameter fit:					
1	2.91	2.80	2.73	2.67	2.59
2	3.12	2.95	2.61	2.05	1.53
3	1.75	1.34	1.15	1.01	0.88
5	1.35	1.15	1.05	0.95	0.84
8-parameter fit:					
5	1.36	1.16	1.05	0.96	0.84
10	1.22	1.11	1.03	0.94	0.83

make a comparison between the time domain and power-spectrum methods in the Appendix.

Equation (3.9) shows that the post-fit residuals grow rapidly as data are collected over longer spans of time. Physically, large-amplitude low-frequency noise becomes increasingly important over longer data spans. The rate of growth of $\overline{R}^2$ with N can be used to estimate the spectral index s, as Groth (1975b) and Cordes (1980) have emphasized. Deeter & Boynton (1982) and Deeter (1984) describe another interesting technique (based on a formalism that has some similarity to our methods) for estimating the shape of the noise spectrum. Their analysis treated finite' samples of unevenly spaced data, but considered only even integral values of s, and did not include the refinement of intrinsic pulsar parameters. Odd s can, however, be of physical interest. In principle, since $T(f)$ is known, it should always be possible to recover $P(f)$ from the power spectrum of the residuals. With the complexities of a finite time series of data, however, a discrete method such as that developed by Deeter (1984) may be more accurate.

As can be seen from Equation (2.14), the variances of the parameters involve integrals over the power spectrum $P(f)$ weighted by the appropriate absorption coefficient. All the integrals converge in the limit as $f \to \infty$ but their properties vary as $f \to 0$. It can be shown that the weighting functions vary as f^0, f^2 and f^4 for α_1 α_2 and α_3 and as f^6 for the rest of the parameters. Consequently, depending on the value of s, one or more of the parameters could have a divergent variance. Physically, this means that the error in the estimated parameter is dominated by noise of very long period and so the variance is essentially determined by the lower cut-off in the spectrum. Uncertainties in α_1 and α_2 are of no consequence. The variance in α_3, however, is of interest. Results are given in Table 2 for various values of N and s. For $s = 5, 6$, the answer depends on $N_{\max}$, the longest-period wave present in the spectrum. If the source of noise is gravitational radiation, $N_{\max}$ is the light travel time to the pulsar (since beyond $N_{\max}$ the effective spectral slope reduces by 2 and so the integral converges), while if it is intrinsic pulsar noise (say a random walk in the rate of spin-down), $N_{\max}$ will probably be of the order of the characteristic spin-down age of the pulsar, $P/2\dot{P}$.

The uncertainty in $\dot{P}$ also affects the accuracy with which $\ddot{P}$ can be measured. The error in $\ddot{P}$ is approximately the difference between the errors in $\dot{P}$ at the beginning and end of the observations, divided by N years. Clearly, errors in $\dot{P}$ caused by very-long-

Table 2. Root-mean-square error $\delta\dot{P}/P$ per μs post-fit residual R_{rms} in units of 10^{-20} s^{-1}. For $s = 5$ and 6 the results depend on the cut-off frequency $f_{min} = 1/N_{max}$ and hence two numbers, A and N^*, are given. For $s = 5$, $\delta\dot{P}/P = A\,[\ln(N_{max}/N^*)]^{1/2}$ and for $s = 6$, $\delta\dot{P}/P = A(N_{max} - N^*)^{1/2}$.

N / s	2	3	4	5	6
7-parameter fit:					
1	46.6	67.6	86.7	43.9, 0.010	74.4, −0.231
2	0.792	2.20	5.78	7.79, 2.83	4.14, 3.07
3	0.293	0.492	0.793	0.841, 3.94	0.389, 4.28
5	0.088	0.148	0.247	0.269, 6.67	0.096, 7.25
8-parameter fit:					
5	0.089	0.148	0.248	0.269, 6.67	0.096, 7.25
10	0.021	0.035	0.059	0.065, 13.6	0.016, 14.7

period waves are not relevant since they coherently affect $\dot{P}$ over the whole range of observations. Therefore, for this calculation, we have used the rms error in $\dot{P}$ contributed by waves with periods less than πN. We then find that the rms error in the braking index, $n_b = P\ddot{P}/\dot{P}^2$, contributed by a red noise process is

$$\delta n_b \sim 8\left(\frac{\tau}{10^7\,\mathrm{yr}}\right)^2\left(\frac{N}{3}\right)^{(s-7)/2}\left(\frac{R_{rms}}{1\,\mu s}\right) \tag{3.11}$$

where τ is the pulsar timing age $P/2\dot{P}$ and s is the index of the noise spectrum. This error is to be compared with $n_b = 3$ predicted by magnetic dipole braking.

Fig. 4 shows the rms uncertainties per μs post-fit residual of pulsar position, proper motion and parallax for $s = 4$ and various values of N. The results are relatively insensitive to s, particularly at large N. This can be understood on the basis of approximate analytical estimates of the variances similar to those made in Section 3.1. Noting that for large N and sufficiently steep spectra the respective variances are dominated by the integrals near $f \sim 1/N$ (below which the integrands fall off as f^{6-s}), it can be shown that the position and parallax variances $\propto \bar{R}^2 / N^2$ and the proper motion variances $\propto \bar{R}_2 / N_4$, with no dependence on s. These scalings are consistent with the more accurate calculations of Fig. 2. Combining with Equation (3.9), the surprising result is that for a given power spectrum, the position and parallax variances $\propto N^{s-3}$ and the proper motion variances $\propto N^{s-5}$, *i.e.*, for a sufficiently steep spectrum *the variance increases with increasing N*. This is quite contrary to the normal wisdom on parameter uncertainties in least squares, which is based on white noise. A comparison of the above scaling laws with those presented in Equations (3.7) and (3.8) shows that the true variance in the presence of red noise can be significantly greater than that estimated on the basis of standard least squares whenever $n \gg N$.

3.4 *Variance Reduction*

We now discuss how prior knowledge of the spectrum can, in principle, be used to reduce the variances in the estimated parameters. For simplicity consider a model consisting of only one parameter, *i.e.*

$$R(t) = \alpha\psi(t) \tag{3.12}$$

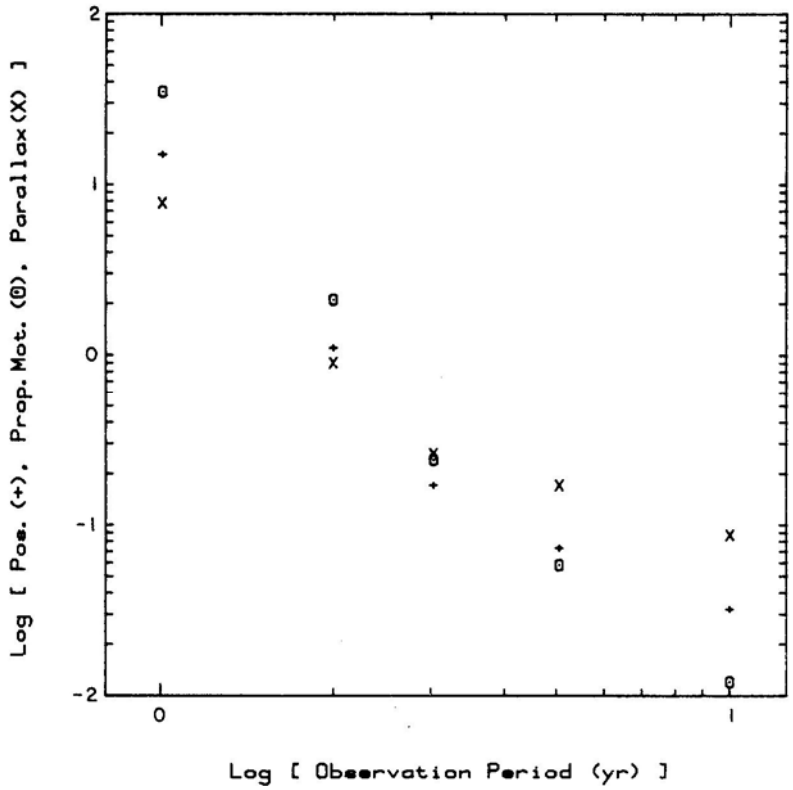


Figure 4. Root-mean-square error in pulsar parameters per μs post-fit residual $R_{\rm rms}$ as a function of the number of years of observation. The results are for $s = 4$, but do not vary a great deal for other values of s. The symbol + shows position errors, $\sin\beta(\delta\beta_0)_{\rm rms}$ and $\cos\beta(\delta\lambda_0)_{\rm rms}$, in as (milli-arcsec). For large N the error scales as $1/N$. The symbol ⊙ shows proper-motion errors, $\sin\beta(\delta\mu_\beta)_{\rm rms}$ and $\cos\beta(\delta\mu_\lambda)_{\rm rms}$, in mas yr^{-1}; scaling as $1/N^2$. The symbol × shows $(\sin^2\beta/d_{\rm kpc})$ × parallax error $\delta d/d$; scaling as $1/N$.

As before we take $\tilde{\psi}(f)$ to be the Fourier transform of $\Psi(t)$. Now let us suppose that we convolve the measured residuals $R(t_i)$ with an arbitrary function $K(t)$, which is equivalent to multiplying $P(f)$ by $|\tilde{K}(f)|^2$. Correspondingly, the new model that is to be fitted is $\tilde{\Psi}(f)\,\tilde{K}(f)$. Proceeding as in Section 2, the variance of α is given by

$$v = \frac{\int_0^\infty P(f)|\tilde{\psi}(f)|^2\tilde{K}(f)|^4\,\mathrm{d}f}{\left[2\int_0^\infty|\tilde{\psi}(f)|^2|\tilde{K}(f)|^2\,\mathrm{d}f\right]^2}. \tag{3.13}$$

We now optimize v with respect to the function $|\tilde{K}(f)|^2$. This gives

$$|\tilde{K}(f)|^2 = K_0/P(f) \tag{3.14}$$

where K_0 is an arbitrary constant. Thus the uncertainty in the parameter is minimum when the noise is 'pre-whitened' before the least squares is performed, with the fitting model being suitably modified.

When there are several parameters the analysis becomes a little more complicated because the variances in (2.14) depend on the orthogonal functions $\tilde{\psi}'_a(f)$ which change

as $\tilde{K}(f)$ is varied. However, a proof can be devised, based on a variational technique where one constantly rotates into a local orthogonal set of functions, to show that (3.14) continues to be optimal even for this case.

Simple estimates indicate that the 'pre-whitened' variances in pulsar position and parallax will be $\propto \bar{R}^2 / n^{s-1} N^{s-1}$ while the variances in proper motion will be $\propto \bar{R}^2 / n^{s-1} N^{s+1}$. The coefficients in these relations, however, are quite large and therefore significant gains are probably possible only for large *s*, *n* and *N*. A practical matter is that at high frequencies measurement errors, which behave like white noise, will dominate. Hence the appropriate *n* to use in the above estimates is not the actual sampling rate but some $n' < n$ where the spectrum changes from red to white noise. We are currently exploring the practicality of implementing this pre-whitening procedure.

4. Application to PSR 1937 + 214

4.1 *General Considerations*

For the particular case of PSR 1937 + 214, $\nu = 642$ Hz and $\dot{\nu} = -4.3 \times 10^{-14}$ Hz s^{-1} (Backer 1984). If we assume that the braking index is 3, then $\ddot{\nu} = 8.6 \times 10^{-30}$ Hz s^{-2}. If we were to include a cubic term in the fitting formula, then the contribution to the residual would be $7 \times 10^{-5} N^3$ μs. This may possibly be detectable after ~ 10 yr but will be significantly harder to measure than the parallax term. We have therefore omitted it from the fitting formula.

The heliocentric latitude and longitude of the pulsar are respectively $\beta = 42.3°$ and $\lambda = 301.3°$. The distance, determined from hydrogen absorption measurements (Heiles *et al.* 1983) is $d \sim 5$ kpc which is consistent with the dispersion measure of $DM = 71$ cm^{-3} pc. Scintillation studies suggest that the speed of the pulsar transverse to the line of sight is ~ 80 km s^{-1} (J. M. Cordes, personal communication) which translates into a proper motion of ~ 3.4 mas yr^{-1}. However, the pulsar is unusually close to the galactic plane for its apparent age and so we expect that the velocity lies within the plane. The parameters α_4–α_8 are expected to have the following magnitudes

$$\alpha_4 \sim \alpha_5 \sim 1.7 \left[\frac{\delta\beta_0, \delta\lambda_0}{1 \text{ mas}} \right] \mu\text{s}, \tag{4.1}$$

$$\alpha_6 \sim \alpha_7 \sim 4.2 \left[\frac{\mu_\beta, \mu_\lambda}{3.4 \text{ mas yr}^{-1}} \right] \mu\text{s}, \tag{4.2}$$

$$\alpha_8 \sim 0.066 \ \mu\text{s}. \tag{4.3}$$

It is clear that the signal given by Equation (4.3) will be very hard to measure; for this reason we have not included parallax within the fitting formula for observing periods $N < 5$. In fact, from the results of Fig. 2, we see that a ~ 30 percent measurement of the parallax will require that the rms. residual over 5 years from red noise should be less than 0.2 μs. Unfortunately, however, dispersion measure fluctuations alone introduce a residual of $\sim 2\,(N/10)^{1/2} \mu$s (*c.f.* Section 4.3).

We should also consider the accuracy of solar system ephemerides over ~ 10 yr timescales. The internal agreement over periods of ~ 10 yr for the best ephemerides is about 3000 metres, *i.e.* 10 μs in arrival time. There is some prospect that improvements

in our knowledge of the position of the telescope relative to the solar system barycentre, which must be known to better than 10 m to exploit the timing fully, will occur over the same period, especially if plans to land a ranger on Phobos in the early 1990's are realized (R. Hellings, personal communication). A related requirement is that local time as measured by atomic clocks be able to avoid drifts in excess of a few μs over ten year periods. Trapped ion clocks may achieve the necessary stability. Of course, the discovery of another quiet millisecond pulsar (or preferably several others) would allow the separation of intrinsic pulsar noise and ephemeris errors to a large extent.

4.2 *Gravitational Radiation*

Several authors (*e.g.* Detweiler 1979; Mashhoon 1982; Bertotti, Carr & Rees 1983) have suggested that an upper bound can be placed on the energy density of primordial gravitational radiation with periods ~ 1 yr using the pulsar timing residuals. In particular, a substantial energy density in gravitational radiation may be produced by primordial cosmic strings and indeed pulsar timing is probably the best way to set limits on the density of these strings (*e.g.* Hogan & Rees 1984). If the energy density in the gravitational radiation between frequency f and $f + \mathrm{d}f$ is $\rho_g(f)$ then the expected power spectrum for the timing noise is

$$P(f) = \frac{G\rho_g(f)}{3\pi^3 f^4}, \tag{4.4}$$

i.e. $P_0 = 1.3 \times 10^4\ \Omega_g(f)\ \mu s^2$ where $\Omega_g(f) = [8\pi\ G\rho_g(f)f]/(3\ H_0^2)$ is the ratio of the wave energy density per natural-logarithm frequency interval at frequency f to the critical cosmological density (setting the Hubble constant. $H_0 = 100$ km s^{-1} Mpc^{-1}). If a fixed fraction of the energy within a horizon during the radiation-dominated era is channelled in some self-similar way into gravitational radiation of comoving wavelength equal to a fraction of the horizon size, then we expect Ω_g to be constant, *i.e.* $P(f) \propto f^{-5}$. Under other circumstances, as discussed by Vilenkin (1981) and Hogan & Rees (1984), structure may be imprinted on the spectrum at the epoch when the universe becomes matter-dominated. Spectral slopes of 5.5 and 7 in the frequency range $0.1 \gtrsim f \gtrsim 10^{-4}$ have also been proposed. Existing observations of the millisecond pulsar can only place a rather modest limit on the energy density of gravitational radiation at frequencies on the order of a few cycles per year. Setting $N = 2$, we see that

$$\Omega_g < 3.4 \times 10^{-4}\ R_{\mu s}^2. \tag{4.5}$$

The difference between this estimate and that given by Hogan & Rees (1984) is due mainly to their assumed value of β. After observations have been carried out for more than 5 years, however, a limit

$$\Omega_g < 4.0 \times 10^{-7}\ R_{\mu s}^2 \tag{4.6}$$

may be set, which would certainly be more interesting. For instance, cosmological models in which primordial strings are created during the earliest epochs of the expanding universe and re-enter the horizon during the radiation era require the string parameter ε to be $\gtrsim 10^{-6}$ if the strings are to have a significant effect on formation. Since $(\varepsilon/10^{-6}) \sim (\Omega_g/2 \times 10^{-7})^2$ (Hogan & Rees 1984), 5 years of sub-μs residuals on PSR 1937 + 214 would be sufficient to exclude such scenarios.

4.3 *Interstellar Density Fluctuations*

Arrival-time fluctuations can also be caused by a variable dispersion measure along the line of sight to the pulsar (Armstrong 1984; Blandford & Narayan 1984a,b). Essentially what happens is that as the observations proceed, larger and larger interstellar clouds can cross the line of sight, causing progressively greater changes in the dispersion measure. The importance of this effect depends upon the spectrum of interstellar density fluctuations in the length-scale range 10^4–10^{16} cm. It has been argued that the spectrum of density fluctuations has a power law form,

$$\Phi_k = C_N^2 k^{-\eta} \tag{4.7}$$

where Φ_k is the three-dimensional power spectrum of the density fluctuations at spatial frequency k. The exponent η has been estimated to be close to the 'Kolmogorov' value of 11/3 (*e.g.* Armstrong, Cordes & Rickett 1981) although there are some indications that it may be somewhat larger (Blandford & Narayan 1984b). Here we adopt a value $\eta = 4$, *i.e.*, $s = 3$. For PSR 1937 + 214 we take C_N^2 to be 10^{-4}, compatible with the measured decorrelation bandwidth (Cordes & Stinebring 1984), together with a measured speed of the scintillation pattern relative to earth of 80 km s^{-1} (J. M. Cordes, personal communication) At an observing wavelength of 1400 MHz. we then find that

$$P(f) = 0.3 f^{-3}\ \mu s^2 \tag{4.8}$$

(*cf.* Armstrong 1984). If most of the measurement error is removed, leaving (4.8) as the dominant noise component in the spectrum, then after three years the timing position can be determined with an uncertainty of ~ 0.23 mas, and the proper motion can be measured to an accuracy of ~ 0.33 mas yr^{-1}. The scaling laws of Section 3.3 indicate that these uncertainties will remain constant for the first parameter and scale as $1/N$ for the second. The uncertainty in the braking index, $\delta\eta_b$, induced by DM fluctuations will be $\sim 2 \times 10^4/N$ (for $N \gtrsim 3$). After three years, the fractional uncertainty in the parallax distance, $\delta d/d$, will be ~ 2.6, and will not improve with time. Therefore, unless dispersion measure fluctuations are monitored and corrected for, parallax distance cannot be determined.

4.4 *Intrinsic Noise*

It has long been known that many pulsars exhibit intrinsic timing noise. The best-analysed case is the Crab pulsar for which successive studies have found that the noise is principally describable as a random walk in frequency (called frequency noise, FN) with $s = 4$ (*e.g.* Groth, 1975b; Cordes 1980).This also appears to be true for a variety of other pulsars, although there are indications that admixtures of random walks in phase and torque must also be included (*e.g.* Cordes & Helfand 1980). We can relate the expected mean squared residual to the diffusion coefficient expressed as the strength of the random walk in frequency P_0/P^2, through

$$\frac{P_0}{P^2} = 1.5 \times 10^{-25} \left[\frac{\beta}{N}\right]^3 \left[\frac{\bar{R}^2}{1\ \mathrm{ms}^2}\right] \left[\frac{1\mathrm{s}}{P}\right]^2 \mathrm{Hz}^2\,\mathrm{s}^{-1}. \tag{4.9}$$

If we assume that FN contributes the bulk of the residual (currently ~ 0.7 μs) in PSR 1937 + 214, then the present data imply that $P_0/P^2 \lesssim 1.4 \times 10^{-25}$ Hz^2 s^{-1}. For

comparison the measured strength of FN in the Crab pulsar is 5.3×10^{-23} Hz^2 s^{-1} (Groth 1975b) and the upper limit on FN for a quiet pulsar, PSR 1237 + 25, is $P_0/P^2 \lesssim 7 \times 10^{-30}$ Hz^2 s^{-1}. To measure P in the millisecond pulsar the rms residual must be less than 10^{-3} μs over a period of 10 years. This limits the strength of any FN random walk to $P_0/P^2 \lesssim 6 \times 10^{-32}$ Hz^2 s^{-1} We thus require the millisecond pulsar to be less restless (by this measure) than any other pulsar we know if the timing is to be exploited fully.

5. Application to other pulsars

Although other pulsars do not have the remarkably small timing residuals of PSR 1937 + 214, the time baselines of the observations are considerably longer ($\gtrsim$10 yr) and so the results of Section 3 for low-frequency noise can still be of interest. Following Bertotti, Carr & Rees (1983), we consider the orbit decay of the binary pulsar, PSR 1913 + 16. The secular decrease in the binary period has been measured to an accuracy of 4 per cent (Weisberg & Taylor 1984) and agrees to this accuracy with the result $\dot{P}/P = 3 \times 10^8$ yr predicted by general relativity. We can therefore take the error in $\dot{P}/P$ to be $< 0.04/3 \times 10^8$ yr $= 4.2 \times 10^{-18}$ s^{-1}. As we have demonstrated, gravitational waves with periods longer than the duration of the observations (but shorter than the light travel time to the pulsar) can cause unusually large variances in period derivatives. PSR 1913 + 16 can be used to set a limit on the energy density in such waves. A background with equal energy density in logarithmic intervals has a spectrum $\propto f^5$ with $P_0 = 1.3 \times 10^4\ \Omega_g^2\ \mu s^2$. The resulting rms timing residual is given by Equation (3.9) with $s = 5$. Therefore, taking $N = 10$ yr, $\beta = 0.94$, and $N_{max} = 10^4$ yr and using Table 2 for $s = 5$, we see that the variance in the measured orbit decay time is

$$\frac{\delta\dot{P}}{P} = 0.17 \times 10^{-20} R_{\mu s} = 1.1 \times 10^{-17} \Omega_g^{1/2} s^{-1}. \tag{5.1}$$

Thus, the measured limit $\delta\dot{P}/P < 4.2 \times 10^{-18}$ s^{-1} yields the upper bound $\Omega_g < 0.15$. The limit on the integrated Ω between $N = 10$ and $N_{max} = 10^4$ is $\Omega_{tot} < 1.0$.

A similar bound can be obtained from PSR 1952 + 29, which has the largest known timing age. We can consider its observed $\dot{P}/P = 4.7 \times 10^{-18}$ to be a statistical upper bound on the rms error in the estimate of its age. Using $N_{max} = 10^3$ yr and $N = 10$ yr, one obtains, as above, the limits $\Omega_g < 0.26$ and $\Omega_{tot} < 1.2$. Other noise spectra are also strongly limited. The expected variance for spindown noise (SN, $s = 6$) is

$$\frac{\delta\dot{P}}{P} = 9 \times 10^{-15} \left[\frac{R_{rms}}{1\ ms}\right] s^{-1} \tag{5.2}$$

so that SN processes are unlikely to contribute more than $\sim 10^{-4}$ of the measured timing residual.

Cordes & Helfand (1980) have determined the dominant noise process for a number of pulsars; the timing noise of PSR 0823 + 26, for example, is apparently described by SN. If the observed 12.6 ms residual is in fact SN dominated, then for ~ 10 years of observation, our model predicts the rms error in $\delta\dot{P}/P$ to be 4.4×10^{-15} s^{-1}. The measured timing age, $\tau = 4.9 \times 10^6$ yr could then be in error by as much as a factor of two or three. This suggests the interesting possibility that such pulsars with a

sufficiently small spindown rate could actually have an observed *spinup* because of strong noise with a steep red spectrum.

As has been previously noted, timing noise makes $\ddot{P}$ measurements and braking index determinations very uncertain. The nominal braking indices reported by Gullahorn & Rankin (1982), ranging up to 10^5 and of both signs, are evidently spurious and can be largely accounted for by the variance expressed by Equation (3.11). Both SN and FN processes as well as a gravitational radiation background can produce δn_b's of the appropriate magnitude.

There are three independent methods for estimating the proper motions of pulsars. Direct interferometry appears to be the most accurate and gives reproducible results (Lyne, Anderson & Salter 1982). Measuring the speeds of scintillation diffraction patterns at the Earth is less accurate and does not provide a direction for the motion but the results here appear to be in agreement with the interferometric determination. The third method, however, which relies on fitting arrival times has only produced a credible result in the case of PSR 1133 + 16 (Manchester, Taylor & Van 1974). Furthermore, the timing *positions* do not agree with those determined interferometrically (Fomalont *et al.* 1984). The discussion of Section 3.3 shows that, in the presence of red noise, uncertainties in the pulsar parameters are often much larger than the reported experimental errors which are calculated assuming white noise alone. The variances in position and proper motion determinations can, in fact, grow with increased observation time. It seems worthwhile to try to pre-whiten the timing noise in these pulsars to see if their timing positions and proper motions can be brought into agreement with the interferometrically determined values.

Acknowledgements

We thank Ron Hellings and Craig Hogan for several discussions and Rajaram Nityananda for comments on the manuscript. Support by the National Science Foundation under grant AST 82-13001 and the Alfred P. Sloan Foundation is gratefully acknowledged. RWR is grateful to the Fannie and John Hertz Foundation for fellowship support.

Appendix

In this paper we have described the timing noise exclusively in terms of power spectra in the arrival residuals. This approach differs from that followed by earlier authors and we now relate the two methods.

Following Boynton *et al.* (1972), Groth (1975) and Cordes (1980), consider three distinct forms of noise, which they describe as random walks in phase (PN), in frequency (FN) and in the time derivative of the frequency (SN). We have corresponding noise spectra with associated exponents s = 2, 4 and 6. However, we make an essential simplification in that we assume the noise to be completely described by its power spectrum. This restricts us to random walk steps that are sufficiently small and frequent to be unresolved by the observations. The formalisms of Groth and Cordes are developed to enable them to detect finite step sizes as well. In practice this has not yet been possible as these effects appear to be masked by measurement errors. (In fact, it

should also be possible to develop the power spectrum approach along these lines by considering bispectra and three-point correlation functions. We shall not pursue this.)

A second important difference is in the treatment of transients associated with the start of the observations. Cordes artificially assumes that the noise commences at the same instant as the observations. The influence of ail prior noise can then be absorbed in the fitted values for the phase, the period and its derivative. A Monte Carlo method is used to relate the ensemble-averaged rms phase residual after a least-squares polynomial fit to the rms phase residual that would have resulted from the same noise adopting the phase, the period and its derivative at the start of the observations. The ratio of these two rms residuals is the quantity $C_{\rm R}$ (m, $T_{\rm obs}$) where m denotes the order of the polynomial and $T_{\rm obs}$ the duration of the observations. $C_{\rm R}$ (m, $T_{\rm obs}$) is independent of $T_{\rm obs}$ provided the rate of occurrence r of random walk steps satisfies $rT_{\rm obs} \gg 1$. Groth deals with the transients in a related manner but instead makes an orthogonal polynomial fit to the observations and compares the coefficients of these polynomials with their expectation values. Both approaches accommodate the non-stationary nature of the phase residuals through a memory of the start of the observations, although the underlying noise process is white in the relevant parameter (*e.g.* frequency), is stationary and possesses a well-defined correlation function.

In our approach, we deal with the transients by assuming that the noise process has been switched on adiabatically in the distant past and "that the phase noise (or equivalently arrival time noise) has a power spectrum which is simply related to the frequency noise spectrum. If the Wiener-Khintchine theorem for the frequency is written

$$\frac{\langle \delta\nu(t)\delta\nu(t+\tau)\rangle}{\nu^2} = \int_0^\infty {\rm d}f Q(f) e^{2\pi i f\tau}; \tag{A1}$$

then the true underlying arrival time power spectrum is simply

$$P(f) = Q(f)/4\pi^2 f^2 \tag{A2}$$

and so on for other types of power spectra. These power spectra as defined here are all stationary.

In fact, we can compute the correction factors, $C_{\rm R}$ (m, $T_{\rm obs}$), introduced by Cordes and evaluated by him through a Monte Carlo procedure directly from these power spectra. Consider phase noise first. The quantity that Cordes considers is

$$\overline{R^2_{\rm PN}(T_{\rm obs})} = \left\langle \frac{1}{T_{\rm obs}} \int_0^{T_{\rm obs}} [R(t) - R(0)]^2 \, {\rm d}t \right\rangle. \tag{A3}$$

Taking Fourier transforms and expressing the result in terms of the power spectrum of the residuals yields

$$\overline{R^2_{\rm PN}} = \int_0^\infty P(f) T_{\rm PN}(f) \, {\rm d}f \tag{A4}$$

where the filter function, $T_{\rm PN}$ (f), is

$$T_{\rm PN}(f) = 2\left[1 - \frac{\sin 2\pi f T_{\rm obs}}{2\pi f T_{\rm obs}}\right] \tag{A5}$$

and T_{obs} is in years. In comparison, the formalism of Section 2 gives the filter function for a Quadratic fit (3 parameters. α_1, α_2, α_3 only) to be

$$T_3(f) = \left[1 - \frac{9}{2x^2} - \frac{9}{x^4} - \frac{45}{2x^6}\right] + \cos 2x\left[\frac{3}{2x^2} - \frac{36}{x^4} + \frac{45}{2x^6}\right] + \sin 2x\left[-\frac{12}{x^3} + \frac{45}{x^5}\right] \tag{A6}$$

where $x = \pi f\, T_{obs}$. Substituting $P(f) = P_0 f^{-2}$ for phase noise one can calculate $\overline{R^2_{PN}}$ from (A4), and $\overline{R_3^2}$ by substituting $T_3(f)$ instead of $T_{PN}(f)$. Their ratio is the correction factor C_R (2, T_{obs}) of Cordes; we obtain the same numerical value. In the case of frequency noise, $s = 4$, and Cordes considers

$$\overline{R^2_{FN}(T_{obs})} = \left\langle \frac{1}{T_{obs}} \int_0^{T_{obs}} \left[R(t) - R(0) - t\,(dR/dt)_{t=0}\right]^2 dt \right\rangle. \tag{A7}$$

The appropriate filter function in this case is

$$T_{FN}(f) = 2\left[1 + \frac{y^2}{6} - \frac{\sin y}{y} + \frac{y\cos y - \sin y}{y}\right] \tag{A8}$$

where $y = 2\pi f\, T_{obs}$. Finally, for spindown noise we have

$$T_{SN}(f) = 2\left[1 + \frac{y^4}{40} + 2\frac{y\cos y - \sin y}{y} + \frac{1}{2}(y^2 - 2)\frac{\sin y}{y}\right]. \tag{A9}$$

We verify the numerical results of Cordes in each case.

References

Armstrong, J. W. 1984 , *Nature*, **307**, 527.
Armstrong, J. W., Cordes, J. M., Rickett, B. J. 1981, *Nature*, **291**, 561.
Backer, D. C. 1984, *J. Astrophys. Astr.*, **5**, 187.
Backer, D. C., Kulkarni, S. R., Heiles, C., Davis, M. M., Goss, W. M. 1982, *Nature*, **300**, 615.
Backer, D. C., Kulkarni, S. R., Taylor, J. H. 1983, *Nature*, **301**, 314.
Bertotti, B., Carr, B. J., Rees, M. J. 1983, *Mon. Not. R. astr. Soc.*, **203**, 945.
Blandford, R., Narayan, R. 1984a, *Proc. Workshop Millisecond Pulsars*, (in press).
Blandford, R., Narayan, R. 1984b, *Mon. Not. R. astr. Soc.*, (in press).
Boynton, P. E., Groth, E. J., Hutchinson, D. P., Nanos, G. P., Jr., Patridge, R. B., Wilkinson, D. T. 1972, *Astrophys. J.*, **175**, 217.
Cordes, J. M. 1980, *Astrophys. J.*, **237**, 216.
Cordes, J. M., Greenstein, G. 1981, *Astrophys. J.*, **245**, 1060.
Cordes, J. M, Helfand, D. J. 1980, *Astrophys. J.*, **239**, 640.
Cordes, J. M., Stinebring, D. R. 1984, *Astrophys. J.*, **277**, L53.
Davis, M. M., Taylor, J. H., Weisberg, J., Backer, D. C. 1984, in preparation.
Deeter, J. E. 1984, *Astrophys. J.*, **281**, 482.
Deeter, J. E., Boynton, P. E. 1982, *Astrophys, J.*, **261**, 337.
Detweiler, S. 1979, *Astrophys. J.*, **234**, 1100.
Fomalont, E. B., Goss, W. M., Lyne, A. G., Manchester, R. N. 1984, *Mon. Not. R. astr. Soc.*, (in press).
Groth, E. J. 1975a, *Astrophys. J. Suppl. Ser.*, **29**, 443.
Groth, E. J. 1975b, *Astrophys. J., Suppl. Ser.*, **29**, 453.

Gullahorn, G. E., Rankin, J. M. 1982, *Astrophys. J.*, **260**, 520.
Heiles, C., Kulkarni, S. R., Stevens, M. A., Backer, D. C., Davis, M. M., Goss, W. M. 1983, A*strophys. J.*, **273**, L75.
Hogan, C. J., Rees, M. J. 1984, *Nature*, **311**, 109.
Lamb, D. Q., Lamb, F. K. 1976, *Astrophys. J.*, **204**, 168.
Lyne, A. G., Anderson, B., Salter, M. J. 1982, *Mon. Not. R. astr. Soc*., **201**, 503.
Manchester, R. N., Taylor, J. H. 1977, *Pulsars.* Freeman, San Francisco.
Manchester, R. N., Taylor, J. H., Van, Y. Y. 1974, *Astrophys. J.*, **189**, L119.
Mashhoon, B. 1982, *Mon. Not. R. astr. Soc.*, **199**, 659.
Press, W. H. 1975, *Astrophys. J.*, **200**, 182.
Romani, R. W., Taylor, J. H. 1983, *Astrophys. J.*, **265**, L35.
Vilenkin, A. 1982, *Phys. Lett.*, **107B**, 47.
Weisberg, J. M., Taylor, J. H. 1984, *Phys. Rev. Lett.*, **52**, 1348.

J. Astrophys. Astr. (1984) **5**, 389–402

Binary Progenitors of Supernovae

Virginia Trimble *Department of Physics, University of California, Irvine CA 92717, USA & Astronomy Program, University of Maryland, College Park MD 20742, USA*

(*Invited article*)

Abstract. Supernovae of both Type I (hydrogen-poor) and Type II (hydrogen-rich) can be expected to occur among binary stars. Among massive stars ($\gtrsim 10\ M_\odot$), the companion makes it more difficult for the primary to develop an unstable core of $\gtrsim 1.4\ .M_\odot$ while still retaining the extended, hydrogen-rich envelope needed to make a typical Type II light curve. Among 1–10 $M_\odot$ stars, on the other hand, a companion plays a vital role in currently popular models for Type I events, by transferring material to the primary after it has become a stable white dwarf, and so driving it to conditions where either core collapse or explosive nuclear burning will occur. Several difficulties (involving nucleosynthesis, numbers and lifetimes of progenitors, the mass-transfer mechanism, *etc.*) still exist in these models. Some of them are overcome by a recent, promising scenario in which the secondary also evolves to a degenerate configuration, and the two white dwarfs spiral together to produce a hydrogen-free explosion, long after single stars of the same initial masses have ceased to be capable of fireworks.

Key words: Supernovae, progenitors—binary stars, evolution

At least half the stars in the sky are double or multiple (Abt & Levy 1976, 1978); and one star in every few hundred must become a supernova in order to produce the event rate observed in massive spiral galaxies like the Milky Way (Tammann 1982); thus, one star out of 300 to 1000 ought to be a binary supernova progenitor. In some cases, the companion will be an innocent bystander to the fireworks, or even inhibit them somewhat. In other cases, it will be a vital participant.

1. Supernovae among single stars

Supernova (SN) events are characterized by the rapid release of large amounts of energy (Baade & Zwicky 1934). Thus the essential requirement for their production is that a star gets itself into a condition where its structure changes suddenly and exoergically. Fig. 1 (which was assembled from the references on single star evolution cited by Trimble, 1982a) presents the possible states of stellar interiors in the central temperature *vs.* central density plane. It reveals two regimes of rapid, exoergic change.

First, a star that wanders into an area where the ratio of specific heats, γ, is less than 4/3 will experience core collapse and release of gravitational potential energy on a timescale $R/V_{\rm sound}$ (milliseconds). A low ratio of specific heats can result from (a) electron-positron pair production at high temperature and low density, (b) photo-dissociation of iron nuclei at high temperature and intermediate density, (c) electron

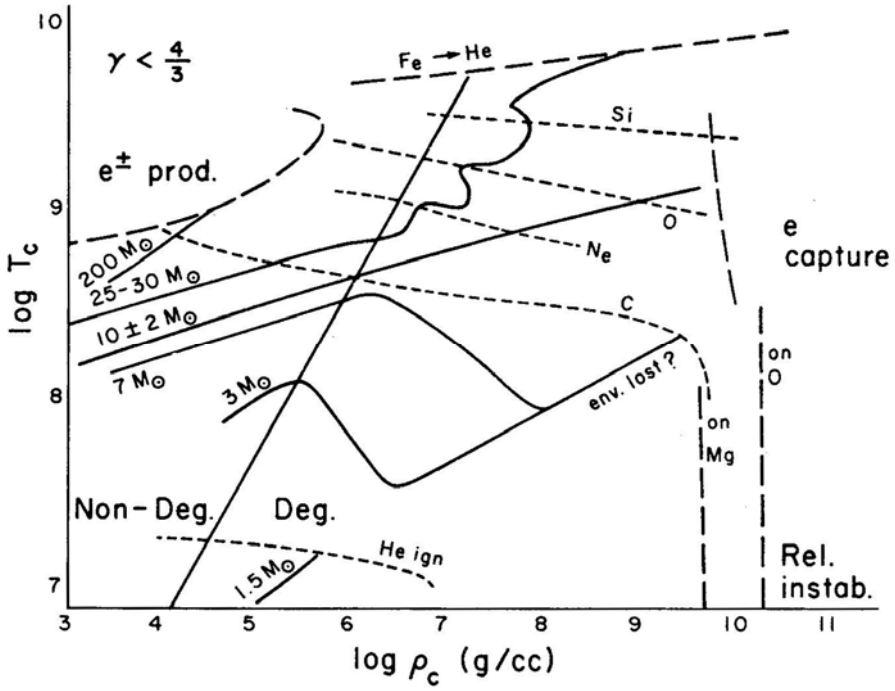


Figure 1. Stellar evolution in the central temperature *vs.* central density plane. The straight, diagonal solid line divides the plane into degenerate and non-degenerate regions. The nearly-horizontal, short-dashed lines indicate loci on which helium, carbon, neon, oxygen, and silicon ignite (*i.e.* energy production by their burning exceeds losses by neutrinos above these lines). The long-dashed lines mark off regions where the ratio of specific heats, γ, is less than $\frac{4}{3}$, so that core collapse will occur. The cause of the soft equation of state is electron-positron pair production in the upper left corner; photodisintegration of iron at the top; electron capture (at slightly different densities for different nuclides) to the right, and general relativistic instabilities at the extreme right. The curving solid lines are evolutionary tracks for stars of the masses indicated. Stars progress in the direction of higher central temperature and density. Those which cross the long-dashed lines should experience core collapse. Those which cross short-dashed lines in the degenerate regime will experience explosive nuclear burning, which can (especially for carbon) disrupt the star. Stars of less than 6–8 $M_\odot$ probably lose their envelopes and stop burning before reaching the carbon-ignition line. The figure was assembled from the references on single-star evolution cited by Trimble (1982a).

capture on intermediate weight nuclei at high density and a range of temperatures, or (d) general relativistic effects at very high densities. The results will depend both on the amount of energy released and on the mechanism of its propagation outward through the rest of the star.

Second, a star with a degenerate core that crosses the line on which a nuclear fuel ignites (*i.e.* energy production exceeds neutrino losses) will experience nuclear energy release at a rate that grows exponentially with time until the central temperature rises high enough to relieve the degeneracy. Fuels that may ignite this way include helium, carbon, neon, oxygen, and silicon. The results will depend primarily on the ratio of energy released to gravitational binding energy of the core.

The sample evolutionary tracks in the figure show that several different classes of stars can enter the assorted instability zones. First, supermassive stars ($M \gtrsim 100\ M_\odot$)

invariably cross into the pair production regime before oxygen burning is complete (Bond, Arnett & Carr 1984). The most massive ones leave black holes, the others explode. Such stars, if they exist at all in our galaxy, are too rare to be responsible for more than 1 per cent or so of the observed SN rate, though one extragalactic event, SN 1961v, the proto-Type V, has been blamed on a supermassive star (Chevalier 1981).

Next, standard models of 12–100 $M_\odot$ stars (Arnett 1977; Lamb, Iben & Howard 1976; Woosley & Weaver 1982) cross into the iron-disintegration regime, and those of 8–12 $M_\odot$ into the electron-capture regime (Nompto 1984). These groups of stars are generally thought to be responsible for the bulk of nucleosynthesis and the majority of pulsar births respectively. Current birthrates of 8–100 $M_\odot$ stars in galaxies match, at least roughly, the rates of Type II Supernovae they are believed to produce (Trimble 1982a; Kennicutt 1984).

Finally, some fuel may ignite degenerately. For the massive stars that burn Ne, O, and Si, the energy release is sufficiently less than the gravitational binding energy that nothing much happens (Arnett 1977). In low-mass stars that ignite helium degenerately, on the other hand, the degeneracy is raised when only a small fraction of the helium has burned (Cole & Deupree 1981), and the star, shaken but structurally still sound, continues its life as a horizontal branch or clump object. But carbon ignition, like the Baby Bear's porridge, is just right, occurring at an intermediate density such that nearly complete burning is needed to raise the degeneracy and the energy release exceeds the gravitational binding.

Arnett (1969) showed that a model 5 $M_\odot$ star without mass loss would evolve to degenerate carbon ignition and disrupt itself completely, producing something rather like a Type I supernova; and Paczyński (1970) demonstrated that all mass-conserving models in the range 3–8 $M_\odot$ should converge to the same sort of event. But real stars must not do this, or we would see an SN I every five years or so, far in excess of the observed rate, and would be drowning in iron-peak elements, made at a rate of ~ 1 $M_\odot$ from every such event. A wide range of observational data and theoretical arguments now indicate (Morris, Jura & Zuckerman 1984) that the vast majority of stars below 6–8 $M_\odot$ lose so much mass during the asymptotic giant (double shell burning) phase that nuclear reactions cease before the CO core reaches ignition conditions. The material shed appears as circumstellar shells and planetary nebulae; the core cools to a normal white dwarf.

It is not implausible that a very few intermediate mass stars with anomalously large mixed cores (due perhaps to rapid rotation) shed only their hydrogen envelopes and produce carbon-detonation SN I's among relatively young stellar populations in spiral and irregular galaxies (Shklovskiĭ 1983; Saio & Wheeler 1980; Tinsley 1980). But this mechanism cannot account for SN I seen among the old stars in elliptical galaxies (van den Bergh 1980, 1982). For these, processes in close binaries provide the only currently popular models.

2. Massive binaries, neutron stars, and Type II Supernovae

2.1 *Wide Systems*

Truly single stars are, in general, quite rare (Abt 1983); and this seems also to be true (Stone 1981, 1982; Garmany, Conti & Massey 1980) for unevolved O and early B stars

that are expected to produce SN II's via iron photodisintegration and electron capture respectively. Any system with large enough velocity amplitude to be picked up as a spectroscopic binary (Garmany, Conti & Massey 1980; Batten, Fletcher & Mann 1978) also has a small enough semi-major axis that the stars will interact before the primary completes its evolution. This is true even for rather extreme cases like VV Cep, consisting of two ~ 18 $M_\odot$ stars in a 20-year orbit, which nevertheless shows spectroscopic evidence for gas streaming from the supergiant to and around the B star. These interacting binaries belong to the next section.

About one-third of the rather small sample of normal Population I O stars studied by Stone (1981) seem to have only visual companions. Another third have spectroscopic companions, and almost a third have both. Although no O or B visual binaries have first-class orbit determinations (Popper 1980), their periods should be $\gtrsim$ 30 yr and their separations larger than the stellar radii will ever be. Both primaries and secondaries in these systems should evolve like model single stars.

We are still not, however, quite out of the woods. The current majority viewpoint is that all or most single and wide binary OB stars shed their entire hydrogen-rich envelopes during post-main-sequence evolution, becoming first Of, then Wolf-Rayet (WR) stars (Conti *et al.* 1983; Falk & Mitalas 1983; Bertelli, Bressan & Chiosi 1984). This is very worrisome. Although 0.1–0.3 $M_\odot$ of hydrogen probably suffices to produce the spectral lines observed in SN II's (D. Branch 1984, personal communication), we need something like 35 $M_\odot$ of extended hydrogen envelope to get the light curve right (T. A. Weaver 1984, personal communication; Woosley & Weaver 1984).

If the pre-SN envelope is compact, most of the shock energy driving the event goes into expanding the envelope rather than into radiation, yielding a very dim SN. And if there is too little hydrogen in the envelope, its recombination cannot produce the prolonged plateau phase seen in most Type II light curves. Core collapses in stars with very small remaining hydrogen envelopes are perhaps responsible for the ~ ⅓ of SNII's with no plateau phase (Litvinova & Nadyozhin 1983). If Flamsteed's star of 1685 gave birth to Cas·A, it may have been an anemic event of this type (Chevalier 1976).When the envelope has been completely stripped after a WR phase, we could still expect a normal supernova remnant eventually to be energized by the ejecta (Arnett 1982), and there might in principle be enough Ni^{56} made to produce a faint Type I event, but we surely cannot get a standard SN II (Woosley & Weaver 1984).

I regard this problem as a fairly serious one (Trimble 1984). It can possibly be resolved within the standard picture if (*pace* Conti *et al.* 1983) only O main-sequence stars $\gtrsim$15 $M_\odot$ are stripped to make Wolf-Rayets (followed by at most very dim supervonae and ejection of several solar masses of heavy elements per event). Then B main-sequence stars of 6 or 8–15 $M_\odot$ might retain enough hydrogen envelope to look like SN II's (for either mechanism of core collapse). A hint that this may be so comes from the Cepheid variables, most of which, whether single or in wide binaries like α UMi, manage to get at least as far as core helium burning with most of their main-sequence mass (~ 10 $M_\odot$) intact (Iben 1983; Carson & Stothers 1984).

Less conventionally, Underhill (1983, 1984; Underhill & Bhatia 1984) has proposed that the rest of us have badly misinterpreted OB star winds (which she believes carry away very little mass) and Wolf-Rayet spectra (which she models with essentially normal He/H ratios ~ 0.1). Even Underbill's WRs have radii much less than the few $\times 10^{13}$ cm needed to make a standard Type II light curve. Lest we despair completely, it is worth remembering that there do exist massive, bright, highly evolved, extended, cool

supergiants, whose spectra are still dominated by hydrogen (H^- continuous opacity), among single stars (Betelgeuse), visual binaries (Antares), and spectroscopic binaries (KQ Per).

We have no direct information on which stars in wide binaries will leave neutron stars (NS) and which black holes (BH) (Woosley & Weaver 1984), though the difficulties in getting core bounce ejection to work for any but the smallest iron cores (Takahara & Sato 1983) would suggest O → BH; B → NS. There remains the long-standing problem of the poor correlation between real neutron stars (pulsars or otherwise) and real supernova remnants. Neither wide nor close companions seem to help with this in any way, and I shall not address it further here.

In any case, demise of the primary in massive wide binaries normally disrupts the system. We expect this on theoretical grounds, because without mass transfer between the components, the primary is still the more massive star when it dies (thus can eject more than half the total mass, the criterion. for unbinding a system by spherically symmetric, impulsive mass loss), and because the system is only loosely bound. Post-disruption velocities will be $\lesssim 20$ km s^{-1} like the predisruption orbital velocities. Observational confirmation comes from the pulsars and OB runaway stars. The four known binary pulsars all have periods $\leqslant 3$ yr and can be modelled as the products of initially close, interacting binaries (van den Heuval & Taam 1984). No wider systems are known, although the same techniques that reveal secular period changes as small as $\dot{P}/P = 10^{-7}$ yr^{-1} should be sensitive to changes in orbital velocity associated with periods up to at least 500 yr. Among the OB runaways, about half show low-mass spectroscopic companions (Stone 1982), at least half of which, in turn, are probably neutron stars whose formation accelerated the systems without disrupting them (de Cuyper 1982). None shows a visual companion (Stone 1981, 1982). But many of the predecessors should have been triples, implying that the visual systems were disrupted when the neutron stars formed. The same remarks apply to the absence of known visual companions to X-ray binaries.

In summary, then, companions in massive wide binaries are normally mere innocent bystanders to the supernova explosions of their primaries and are usually liberated in the process.

2.2 *Close Systems*

At least 30 per cent of unevolved O stars are spectroscopic binaries (Garmany, Conti & Massey 1980), most of those detected having mass ratios fairly close to one (distinct from later types, and apparently not entirely an observational selection effect; Abt 1983). Such stars necessarily interact during their evolution, the best-known process being mass transfer when the primary (and later the secondary) expands to fill its inner Lagrangian surface (Roche lobe). Evolutionary models of such interacting massive binaries go back almost 20 years (Paczyński 1966, 1967; Plavec 1967; Kippenhahn & Weigert 1967; Snezhko 1967). More recent ones are exceedingly numerous, often include the effects of mass and angular momentum lost to the system, and typically follow the stars to the bitter end of the two compact (or disrupted) remnants (Doom & De Grève 1983; Yungel'son & Masevich 1983; Vanbeveren 1983; Kornilov & Lipunov 1984; van den Heuvel 1981a, b; van den Heuvel & Taam 1984; de Loore & Sutantyo 1984; and many others).

From the point of view of supernova production, the differences from single star evolution are largely bad! Mass transfer strips the primary (and probably also the secondary) even more thoroughly than single-star winds. And when the stripping begins during hydrogen burning, it makes building a core in excess of the Chandrasekhar limiting mass (needed for most SN mechanisms) considerably more difficult. This raises the cut between white dwarf producers and neutron star producers to 1215 $M_\odot$ Webbink 1979). The result is a 30–50 per cent reduction in the Type II supernova rate expected from a stellar population like that currently being formed in our Galaxy (Scalo 1984).

The stripped, though still intact, primary (and later secondary) should look for a time like a Wolf-Rayet star, for initial masses from at least 17 (Doom & De Grève) to 100 $M_\odot$ (Stickland *et al.* 1984 on CQ Cep). This was, in fact, the first plausible mechanism suggested for making WRs (Paczyński 1967). Unfortunately, known binary WRs look much the same as those for which no companion can be detected; and the fraction of WRs that are close binaries is close enough to that among their main-sequence ancestors (~ 50 per cent, Hidayat, Admiranto & van der Hucht 1984; Lamontagne, Moffat & Seggewiss 1983) to suggest that wind stripping dominates, at least in the later phases, and that the companion is, once again, an innocent bystander.

In contrast to the single star and wide binary cases, we know a bit about the neutron-star/black-hole mass cut for close systems. An analysis of the black-hole candidate system LMC X-3 (van den Heuvel & Habets 1984) suggests that a main-sequence primary of at least 50 ± 10 $M_\odot$ is needed for black hole production. It would be quite self-consistent to say that this does not differ between wide and close binaries, but it could also be larger for the close systems, like the WD/NS cut.

Also in contrast to the wide binary case, the first supernova event frequently does not disrupt a close, massive system. We expect this on theoretical grounds (de Cuyper 1982), because mass transfer and loss in the system guarantee that the star that explodes first is, by then, the less massive component, so that no spherically symmetric explosion can unbind the system (Trimble & Rees 1971), and because the potential well is much deeper than for wide systems. The explosion of the secondary, in contrast, can be expected to disrupt frequently. Observational confirmation again comes from OB runaway stars, pulsars, and X-ray binaries.

About half the OB runaways (Stone 1982) and some WR runaways (Isserstedt, Moffat & Niemala 1983) have low-mass spectroscopic companions, of which, in turn, at least half are probably neutron stars or other compact configurations (de Cuyper 1982). The few known binary pulsars have orbital periods of 0.32,1.03, ~ 120, and 1232 days, and all are most convincingly evolved from close systems, two of initially high mass and two low (van den Heuvel & Taam 1984). Next, the observed numbers and expected lifetimes of massive X-ray binaries require that most existing massive OB + WR systems must evolve into them without disruption (van den Heuvel 1981a, b; Doom & De Grève 1983; *etc.*). Finally, the great preponderance of single pulsars suggests that the first NS produced typically has its low-frequency emission quenched by gas from its close companion, while the transformation of the secondary to a second pulsar unbinds the system.

In summary, then, the companions in close, massive binary systems generally affect the evolution of their primaries in ways inimical to the production of Type II Supernovae that resemble observed events. The systems generally remain bound through the first neutron-star or black-hole formation process, and not the second.

3. Low-mass binaries, cataclysmic variables, and Type I supernovae

3.1 *Wide Systems*

There seems to be nothing to say about these systems except that (1) if the primary manages to produce a supernova, it will be of the single-star, carbon-deflagration sort mentioned in Section I, with the companion only an on-looker, and (2) since such events completely disrupt the star concerned, they will, *α fortiori*, disrupt the system.

3.2 *Close Systems—the Recent Concensus*

We ended Section I with the point that single-star Type I Supernovae require a combination of events that might not occur anywhere and cannot occur among the old, low-mass stars that dominate elliptical galaxies. The idea that one could overcome the difficulties with close-binary, mass-transfer models caught on rather suddenly in the early 70's (Wheeler & Hansen 1971; Truran & Cameron 1971; Hartwick 1972; Whelan & Iben 1973; Mazurek 1973).

The general picture is that a primary of any mass 1–7 or more $M_\odot$ produces a white dwarf, which then waits patiently for the secondary to evolve away from the main sequence and transfer mass back onto it. The white dwarf's mass grows until it (1) collapses by electron capture on O, Ne, and Mg to a neutron star, (2) ignites helium off centre, detonating the helium layer and leaving a CO core behind still at white-dwarf densities, (3) ignites carbon off centre, so that dual-detonation waves propagate in ward and outward, incinerating and disrupting the whole star, or (4) ignites carbon at the centre, so that a deflagration runaway burns only the central part of the star, but the whole thing is disrupted (Woosley, Axelrod & Weaver 1984 and earlier references therein). Which of these happens depends on (1) the mass and composition of the primary white dwarf ($\lesssim 0.45\ M_\odot$ of He, 0.45–1.1 $M_\odot$ of CO, or 1.1–1.4 $.M_\odot$ of Ne-O-Mg), (2) how long the WD cools before back transfer begins (Isern, Labay & Canal 1984), (3) the rate at which fresh material arrives, and (4) the composition of the accreted material, via the amount of heating produced when it burns.

One charm of this scenario is that the supernova event follows star formation after a time set by the lifespan of the lower-mass secondary, which can be long, yet the system has the mass of the primary to draw on, greatly improving the chances of getting close enough to 1.4 $M_\odot$ for one of the four instabilities mentioned above to set in. Additional advantages are: First, there exists a fairly numerous class of objects, the cataclysmic variables (CVs, including novae, dwarf novae, recurrent novae, nova-like variables, symbiotic stars, and polars), which can be claimed as *en route* to producing such SN I's. Second, models based on this picture (summarized in Wheeler 1980, Trimble 1982a, Rees & Stoneham 1982, and Iben & Tutukov 1984) provide fairly good matches to the spectra and light curves observed for Type I Supernovae.

The spectra near maximum light consist of an underlying blackbody from a ~ 8000 K photosphere expanding at ~ 10000 km s^{-1}, plus broad P Cygni lines of common elements in roughly solar proportions, apart from a complete absence of hydrogen (Branch 1982). For instance, the products of the deflagrations studied by Nomoto, Thielemann & Wheeler (1984) provide the right line intensities for SN 1981b if layers of the star are well mixed (Branch 1984 and personal communication). Spectra

well past maximum light are dominated by iron lines. There is not yet complete agreement upon their interpretation, but one possibility is a large excess of iron and cobalt (Axelrod 1980). This is important because the model light curves (Weaver, Axelrod & Woosley 1980; Woosley, Weaver & Taam 1980) have two main energy inputs, instantaneous release as carbon and oxygen burn to iron-peak elements, especially Ni^{56}, and gradual release as the Ni^{56} beta decays via Co^{56} to Fe^{56} (half lives 6 and 77 days). Next, such events, when the white dwarf disrupts, contribute significant amounts of oxygen-burning products (both abundant and rare species) to the galactic supply. This will be especially important if very massive stars typically trap their entire iron cores in black holes.

Finally, the case where the white dwarf collapses gently rather than exploding is (apart from assorted capture mechanisms) seemingly the only way to make the low mass X-ray binaries like Her X–1 (Webbink, Rappaport & Savonije 1983; de Loore & Sutantyo 1984; van den Heuvel 1981a, b). In fact, one must be a bit careful not to let this happen too often and overproduce such systems (Taam & Fryxell 1984; Iben & Tutukov 1984). The implication is that an accreting white dwarf deflagrates or detonates and disrupts (either itself or at least the system) in all but perhaps those very few cases where its initial mass was very close to the Chandrasekhar limit (as in the systems discussed by Law & Ritter 1983). Van den Heuval & Taam (1984) use this process also to account for the two binary pulsars with low-mass companions. Such triggered collapses occur much slower than the free-fall timescale, and so are unlikely to give shockwave ejection or make SNRs (Lipunov 1983).

A very large fraction of the people currently working on Supernovae are convinced that some version of this close-binary scenario is responsible for many, most, or all Type I events (but see Imshennik & Nadëzhin 1983 for contrasting views). There are, however, several problems with the scheme, some affecting only details and some possibly fundamental. Section 3.3 addresses these.

3.3 *Problems with the Concensus Model*

The objections initially voiced to all carbon detonation Supernovae were (Ostriker, Richstone & Thuan 1974) that, if they were at all common (a) the pulsar production rate would not be sufficient to keep up the observed supply, and (b) we would be drowning in iron. The first of these problems has somewhat changed its form in the intervening decade. Many supernova remnants, including those associated with the 1572 and 1604 (Tycho and Kepler) events, simply do not contain neutron stars of the same age as the remnants (Helfand & Becker 1984). Thus we cannot object, *α priori*, to a supernova mechanism that leaves no compact core! The lack of correlation between SNe, SNRs, and NSs remains a puzzle (Srinivasan, Bhattacharya & Dwarakanath 1984, and many others), and we are not going to explain it here.

The iron problem is still with us. Sutherland & Wheeler (1984) note that the maximum amount of iron we can tolerate from Type I Supernovae occurring every 50–100 yr in our galaxy is about 0.7 $M_\odot$, and that this is just about the minimum needed to give the disrupted star the observed expansion velocity. This much burning makes the event intrinsically rather bright, corresponding to a large extragalactic distance scale. The authors go so far as to say that, if the scale should be established at $H_0 \geqslant 100$ km s^{-1} Mpc^{-1} by other means, then the carbon detonation/deflagration model of SN I would have to be abandoned.

Woosley, Axelrod & Weaver (1984) conclude, somewhat more gently, that no single model simultaneously yields believable element and isotope ratios in its burning products while matching typical light curves and spectra. They, however, suggest possible ways out through variations from one event to another and/or departures from the bare, spherically symmetric white dwarfs of their models.

Another difficulty with the concensus scenario is a statistical one: are there enough progenitors? Recent discussions, both simple (Greggio & Renzini 1983; Trimble 1982b) and elaborate (Iben & Tutukov 1984; Patterson 1984) conclude that our own galaxy is rather close to the ragged edge. Making enough cataclysmic variables is easy—either a few per cent of the low-to-intermediate mass binaries might function that way for $\sim 10^9$ yr each, or they might all do it for a few per cent of their lifetimes. But the SN I's are more difficult—nearly all binaries capable of growing an explosive (~ Chandrasekhar mass) white dwarf must do so without more mass being lost from the system than would be shed by similar single stars. I find this somewhat unlikely-sounding, given that the process of bringing the stars close enough together to get a cataclysmic system requires considerable angular momentum (hence mass) to be lost in a common envelope phase (Paczyński 1976).

There are, however, a good many factors of two to play with. And, if we look at measured masses of CVs (24 systems tabulated by Patterson 1984, 39 tabulated by Ritter 1984, with considerable overlap), we see that about a quarter of them already have white dwarf masses $\gtrsim 1\ M_\odot$ and about 40 per cent have total masses in excess of $1.4\ M_\odot$ These could all become SN I's if no further mass were lost, and they may be just about enough (Patterson 1984, *etc.*), except that the novae, at least, expel material from the system at least as fast as the secondary tries to give it to the primary.

Patterson (1984) also worries about the eventual fate of CVs that do not give rise to Supernovae. It now seems likely (Nather 1984) that continued mass transfer and ejection erodes them down to very small, short-period binary white dwarfs, of which we currently know three examples (AM CVn, GP Com, PG 1346 + 082).

The preceding three paragraphs apply to population I stars in our own galaxy. We do not directly know the formation rate of binaries as a function of stellar mass and separation for any other galaxy, except for the very brightest stars in Andromeda and the Magellanic Clouds, which eclipse about as often as similar Milky Way stars (Herczeg 1982a, b). In particular, there is no information on the giant elliptical galaxies, for which these binary models seem most vital.

We might, therefore, be tempted to assume without further worry that the binary formation rate, like the initial mass function (Scalo 1984) varies rather little from place to place, were it not for an apparent severe deficit of binaries among galactic Population II stars, which are as old as giant elliptical populations, though much poorer in heavy elements.

The precise extent of the deficit is debated from time to time, but there are no confirmed eclipsing (main sequence or giant) binaries among the globular clusters (Hogg 1973; Webbink 1980) or in the dwarf spheroidal Draco (Herczeg 1982a, b). In addition, several searches for radial velocity variability among globular cluster stars (Mayor *et al.* 1984 on 47 Tuc, with references to earlier work) have found only atmospheric effects and no spectroscopic binaries, though corresponding investigations of open clusters found many. The tightness of the main sequence in many globular cluster colour-magnitude diagrams (Richer & Fahlman 1984, on M 4; Sandage & Katem 1983, on M 92) says that at most a few per cent of the stars are binaries with mass

ratios $\gtrsim 0.7$ (among the commonest sorts in the solar neighbourhood). Even among field subdwarfs, colours suggest a lower-than-average binary incidence (Eggen 1983; Carney 1983). Finally, although the globular clusters have their fair share and more of cataclysmic variables and low-mass X-ray binaries, these were probably formed by capture processes among previously single neutron stars, white dwarfs, and main-sequence stars (Hertz & Grindlay 1983, 1984). Thus, we really cannot guess, even to order of magnitude, how many binary progenitors are available to make Type I Supernovae in elliptical galaxies. Shklovskiĭ (1978) has associated the declining SN I rate along the galaxy type sequence Sc–Sb–Sa–S0–E with declining binary frequency.

The problems noted thus far—non-production of pulsars, over-production of iron, and possible shortage of progenitors—apply to essentially all versions of the concensus model. Still, none of them sounds absolutely fatal. There are, however, two additional problems, potentially more serious, which a recent modification of the standard model enables us to avoid. If cataclysmic variables are the immediate predecessors of Type I Supernovae, then (1) the material being transferred to the white dwarf is necessarily mostly hydrogen, and (2) the maximum available time from star formation to explosion is the nuclear-burning lifetime of the secondary.

Having hydrogen around is, on the whole, a bad thing (Sutherland & Wheeler 1984), since type I spectra do not show any, and, unless the hydrogen is accreted at a carefully selected rate (which the secondary may not know about), it burns in violent flashes every 10^{4-5} yr, making nova explosions, (Sion, Acierno & Tomsczyk 1979). These blow off everything that was accreted, and perhaps some material from the white dwarf as well, so the white dwarf mass does not increase with time. The timescale problem arises because the secondary must be massive enough to be a useful donor, but small enough to live $\sim 10^{10}$ yr. In this connection, it is of interest that the Type I event 1983n, whose radio emission (because of the kind of circumstellar shell needed to make it) appears to imply initial masses near $8 + 6.5\ M_\odot$ (Sramek, Panagia & Weiler 1984; Chevalier 1984), occurred in a spiral galaxy with considerable current star formation (M 83). On the other hand, some of the SN I's whose infrared emission seems to be a 'light echo' from dust in similar shells (Evans *et al.* 1983) were in ellipticals.

The hydrogen and timescale problems both go away if the progenitor's biography has a chapter in which the secondary completes its evolution becoming a second white dwarf (with or without CV phase), and the two degenerate stars then "spiral together as angular momentum is drained from the system by gravitational radiation (Dyson 1963; Kraft, Mathews & Greenstein 1962) or magnetic braking (Taam 1983; Patterson 1984). The transferred material will be largely helium or carbon and oxygen, with a normal admixture of heavier elements. This will neither flash nor contaminate the eventual spectrum. And the time available is expanded by however long it takes the pair to spiral together—anything from 10^8 to $> 10^{10}$ yr, for plausible initial separations, with the longer times perhaps more likely.

In addition, the merger process can provide a good deal of extra heating, so that considerably less than a Chandrasekhar mass may be able to detonate. The white-dwarf-red-giant death spiral scenario of Sparks & Stecher (1974) shares this last advantage. And models where the donor star has already been stripped to a helium star avoid the hydrogen shell flash problem (Fujimoto & Sugimoto 1982). But only the double degenerate dwarf version has all three virtues. Webbink (1979) mentioned the possibility in a single sentence, while Tutukov & Yungel'son (1979) approached it peripherally. I first heard, of the scenario from B. Paczyński (1982, personal

communication and 1983) in connection with the question of identifying pre-explosion systems. Recent detailed discussions have been published by Iben & Tutukov (1984) and Webbink (1984).

Among the unanswered questions are how can we identify the precursor pairs and how many of them exist. The current white dwarf formation rate in the galaxy is about 0.51.0 yr (Guseinov, Novruzova & Rustamov 1983; Weidemann & Koester 1983), so we need 25 per cent of existing white dwarfs (in steady state) to be reasonably massive binaries with timescales for angular momentum loss $\lesssim 10^{10}$ yr to keep up the SN I rate. The number presently known is zero. The three very close double white dwarfs (prototype AM CVn) all have total masses $\lesssim 0.5\ M_{\odot}$, ,which will not do (Nather 1984). And known more massive systems (like the Sanduleak-Pesch object, Greenstein, Dolez & Vauclair 1983; and G 107 – 70, Harrington, Christy & Strand 1981) are visual binaries, with spiraling-in timescales rather in excess of the lifetime of the proton. Double degenerate dwarfs, even with orbital periods $\lesssim 3$ h, are exceedingly unlikely to eclipse. Thus the proper phenomena to look for are variable radial velocity or spectral peculiarities (B. Paczyński 1982, personal communication). Eggen (1984) has reported one white dwarf (CoD – 48°3636) with possibly double lines (though I suspect these could be single broad lines with emission cores). And Greenstein & Trimble (1967) tabulated a handful of white dwarfs whose velocities were discordant on two or more apparently good 200-inch prime focus spectrograph plates. Observers with access to suitable instruments are invited to look for radial velocity variations $\gtrsim 100$ km s^{-1} on timescales $\lesssim 3$ h for these stars and any others that appeal to them.

4. Conclusions

Among the massive stars expected to produce Type II (hydrogen-rich) Supernovae, the presence of a close companion seems to be largely negative. It can increase the main-sequence mass needed to yield a collapsing core, and, owing to mass transfer from the primary to the secondary (and later, back again), the companion enhances the stripping of the stellar hydrogen envelope produced by single star winds, thus making it harder for the star to give rise to a typical SN II light curve.

Among the less massive stars that we think make Type I (hydrogen-free) Supernovae, a close companion could be an innocent bystander to carbon detonation/deflagration in the primary, or it can be a vital participant, transferring material to a white dwarf primary and so driving it to explosive conditions. A widely discussed recent scenario allows both stars first to become degenerate dwarfs, which then spiral together, giving rise to a hydrogen-poor explosion arbitrarily long after the star formation event, even for quite massive binaries.

Acknowledgements

My views on the late phases of single and double star evolution have been particularly influenced by the work of W. David Arnett, E. P. J. van den Heuvel, Bohdan Paczyński, and Stanford E. Woosley. Drs David Branch and Thomas Weaver generously and carefully answered questions-by-telephone on the subject of hydrogen in Type II's.

References

Abt, H. A 1983, *A. Rev. Astr. Astrophys.*, **21,** 343.
Abt, H. A., Levy, S. G. 1976, *Astrophys. J. Suppl. Ser.*, **30,** 273.
Abt, H. A., Levy, S. G. 1978, *Astrophys. J. Suppl. Ser.*, **36,** 241.
Arnett, W. D. 1969, *Astrophys. Space Sci.*, **5,** 180.
Arnett, W. D. 1977, *Astrophys. J. Suppl. Ser.*, **35,** 145.
Arnett, W. D. 1982, in *Supernovae: A Survey of Current Research*, Eds M. J. Rees & R. J. Stoneham, D. Reidel, Dordrecht, p. 221.
Axelrod, T. S. 1980, in *Type I Supernovae*, Ed. J. C. Wheeler, Univ. Texas & McDonald Obs., Austin, p. 80.
Baade, W., Zwicky, F. 1934, *Proc. Natl. Acad. Sci. Am.*, **20,** 254, 259.
Batten, A. H., Fletcher, J. M., Mann, P. J. 1978, *Publ. Dom. Astrophys. Obs.*, **15,** 121.
Bertelli, G., Bressan, A. G., Chiosi, C. S. 1984, *Astr. Astrophys.*, **130,** 279.
Bond, J. R., Arnett, W. D., Carr, B. J. 1984, *Astrophys. J.*, **280,** 825.
Branch, D. 1982, in *Supernovae: A Survey of Current Research*, Eds M. J. Rees & R. J. Stoneham, D Reidel, Dordrecht, p. 267.
Branch, D. 1984, in *W. A. Fowler Conf. on Nucleosynthesis*, Univ. Chicago Press (in press).
Carney, B. W. 1983, *Astr. J.*, **88,** 610, 623.
Carson, T. R., Stothers, R. B. 1984, *Astrophys. J.*, **276,** 593.
Chevalier, R. A. 1976, *Astrophys. J.*, **207,** 872.
Chevalier, R. A. 1981, *Fund. Cosmic Phys.*, **7,** 1.
Chevalier, R. A. 1984, *Astrophys. J. Lett.* (in press).
Cole, P. W., Deupree, R. G. 1981, *Astrophys. J.*, **247,** 607.
Conti, P., Garmany, C., de Loore, C., Vanbeveren, D. 1983, *Astrophys. J.*, **274,** 302.
de Cuyper, J.-P. 1982, in *IAU Coll.* 69: *Binary and Multiple Stars as Tracers of Stellar Evolution*, Eds Z. Kopal & J. Rahe, D. Reidel, Dordrecht, p. 417.
de Loore, C., Sutantyo, W. 1984, *Astrophys. Space Sci.*, **99,** 335.
Doom, C., De Grève, J. P. 1983, *Astr. Astrophys.*, **120,** 97.
Dyson, F. 1963, in *Interstellar Communication*, Ed. A. G. W: Cameron, Pergamon, New York, p. 115.
Eggen, O. J. 1983, *Astr. J.*, **88,** 813.
Eggen, O. J. 1984, *Astr. J.*, **89,** 389.
Evans, D. S. *et al.* 1983, *Nature*, **304**, 709.
Falk, D. W., Mitalas, R. 1983, *Mon. Not. R. astr. Soc.*, **202,** 19.
Fujimoto, M. Y., Sugimoto, D. 1982, *Astrophys. J.*, **257,** 291.
Garmany, C. D., Conti, P. S., Massey, P. 1980, *Astrophys. J.*, **242,** 1063.
Greenstein, J. L., Dolez, N., Vauclair, G. 1983, *Astr. Astrophys.*, **121,** 25.
Greenstein, J. L., Trimble, V. L. 1967, *Astrophys. J.*, **149,** 283.
Greggio, L., Renzini, A. 1983, *Astr. Astrophys.*, **118,** 217.
Guseinov, O. Kh., Novruzova, H. I., Rustamov, Yu. S. 1983, *Astrophys. Space Sci.*, **96,** 1.
Harrington, R. S., Christy, J. W., Strand, K. Aa. 1981, *Astr. J.*, **86,** 909.
Hartwick, F. D. A. 1972, *Nature, Phys. Sci.*, **237,** 137.
Helfand, D., Becker, R. 1984, *Nature*, **307,** 215.
Herczeg, T. J. 1982a, in *IAU Coll.* 69." *Binary and Multiple Stars as Tracers of Stellar Evolution*, Eds Z. Kopal & J. Rahe, D. Reidel, Dordrecht, p. 145.
Herczeg, T. J. 1982b, in *Landolt-Börnstein*, New Series, Eds K. Schaifers & H. H. Voigt, Springer-Verlag, Berlin, **2,** 381.
Hertz, P. 1984, *Astrophys. J.*, **275,** 105.
Hertz, P., Grindlay, J. E. 1983, *Astrophys. J.*, **267,** L83.
Hidayat, B., Admiranto, A. G., van der Hucht, K. A. 1984, *Astrophys. Space Sci.*, **99,** 175.
Hogg, H. S. 1973, *Publ David Dunlap Obs.*, **3,** No. 6.
Iben, I. 1983, *Solar Phys.*, **82,** 457.
Iben, I., Tutukov, A. V. 1984, *Astrophys. J. Suppl. Ser.*, **54,** 335.
Imshennik, V. S., Nadëzhin, D. K. 1983, *Astrophys. Space Phys. Rev.*, **2,** 75.
Isern, J., Labay, J., Canal, R. 1984, *Nature*, **309,** 431.
Isserstedt, J., Moffat, A. F. J., Niemala, V. S. 1983, *Astr. Astrophys.*, **126,** 183.

Kennicutt, R. C. 1984, *Astrophys. J.*, **277,** 361.
Kippenhahn, R., Weigert, A. 1967, *Z. Astrophys.*, **65,** 251.
Kornilov, V. G., Lipunov, V. M. 1984, *Soviet Astr.*, **27,** 163.
Kraft, R. P., Mathews, J., Greenstein, J. L. 1962, *Astrophys. J.*, **136,** 312.
Lamb, S. A., Iben, I. A., Howard, W. M. 1976, *Astrophys. J.*, **207,** 209.
Lamontagne, R., Moffat, A. F. J., Seggewiss, W. 1983, *Astrophys. J.*, **269,** 596.
Law, W. Y., Ritter, H. 1983, *Astr. Astrophys.*, **123,** 33.
Lipunov, V. M. 1983, *Astrophys. Space. Sci.*, **97,** 121.
Litvinova, I. Yu., Nadyozhin, D. K. 1983, *Astrophys. Space Sci.*, **89,** 89.
Mayor, M. *et al*. 1984, *Astr. Astrophys.*, **134,** 118.
Mazurek, T. J. 1973, *Astrophys. Space. Sci.*, **23,** 365.
Morris, M., Jura, M., Zuckerman, B. (Eds) 1984, in *Mass Loss from Red Giants*, D. Reidel, Dordrecht (in press).
Nather, R. E. 1984, in *Interacting Binaries*, Eds J. Pringle & P. Eggelton, D. Reidel, Dordrecht (in press).
Nomoto, K. 1984, *Astrophys. J.*, **277,** 791.
Nomoto, K., Thielemann, F.-K., Wheeler, J. C. 1984, *Astrophys. J.*, **279,** L23.
Ostriker, J. P., Richstone, D. O., Thuan, T. X. 1974, *Astrophys. J.*, **188,** L87.
Paczyński, B. 1966, *Acta Astr.*, **16,** 231.
Paczyński, B. 1967, *Acta Astr.*, **17,** 1, 193, 287 & 355.
Paczyński, B. 1970, *Acta Astr.*, **20,** 40.
Paczyński, B. 1976, in *IAU Symp.* 73: *Structure and Evolution of Close Binary Systems*, Eds P. Eggleton, S. Mitton & J. Whelan, D. Reidel, Dordrecht, p. 75.
Paczyński, B,1983, in *7th North-American Workshop on Cataclysmic Variables and Low-Mass X-*ray *Binaries*, D. Reidel, Dordrecht (in press).
Patterson, J. 1984, *Astrophys. J. Suppl. Ser.*, **54,** 443.
Plavec, M. 1967, *Comm. Obs. R. Belgique, Uccle*, **B17,** 83.
Popper, D. M. 1980, *A. Rev. Astr. Astrophys.*, **18,** 115.
Rees, M. J., Stoneham, R. J. (Eds) 1982, *Supernovae: A Survey of Current Research*, D. Reidel, Dordrecht.
Richer, H. B., Fahlman, G. 1984, *Astrophys. J.*, **277,** 227.
Ritter, H. 1984, *Astr. Astrophys. Suppl. Ser.* (in press).
Saio, H., Wheeler, J. C. 1980, *Astrophys. J.*, **242**, 1176.
Sandage, A. R., Katem, B. 1983, *Astr. J.*, **88**, 1146.
Scalo, J. 1984, *Fund. Cosmic Phys.* (in press).
Shklovskiĭ, I. S. 1978, *Soviet Astr.*, **22,** 413.
Shklovskiĭ, I. S. 1983, *Soviet Astr. Lett.*, **9,** 250.
Sion, E. M., Acierno, M. J., Tomsczyk, S. 1979, *Astrophys. J.*, **230**, 832.
Snezhko, L. I. 1967, *Perem. Zvezdy*, **16,** 253.
Sparks, W. M., Stecher, T. P. 1974, *Astrophys. J.*, **188,** 149.
Sramek, R. A., Panagia, N., Weiler, K. W. 1984, *Astrophys. J. Lett.* (in press).
Srinivasan, G., Bhattacharya, D., Dwarakanath, K. S. 1984, *J. Astrophys. Astr.*, **5,** 403.
Stickland, D. J., Bromage, G. E., Budding, E., Burton, W. M., Howarth, I. D., Jameson, R., Sherrington, M. R., Willis, A. J. 1984, *Astr. Astrophys.*, **134,** 45.
Stone, R. C. 1981, *Astr. J.*, **86,** 544.
Stone, R. C. 1982, *Astr. J.*, **87,** 90.
Sutherland, P. G., Wheeler, J. C. 1984, *Astrophys. J.*, **280,** 282.
Taam, R. E. 1983, *Astrophys. J.*, **268,** 361.
Taam, R. E., Fryxell, B. 1984, *Astrophys. J.*, **279,** 166.
Takahara, F., Sato, K. 1983, *Prog. theor. Phys.*, **71,** 524.
Tammann, G. A. 1982, in *Supernovae: A Survey of Current Research*, Eds M. J. Rees & R. J. Stoneham, D. Reidel, Dordrecht, p. 371.
Tinsley, B. M. 1980, in *Type I Supernovae*, Ed. J. C. Wheeler, Univ. Texas & McDonald Obs., Austin, p. 196.
Trimble, V. 1982a, *Rev. Mod. Phys.*, **54,** 1183.
Trimble, V. 1982b, *Observatory*, **102,** 133.
Trimble, V. 1984, in *Mass Loss from Red Giants*, Eds M. Morris, M. Jura & B. Zuckerman, D. Reidel, Dordrecht (in press).

Trimble, V., Rees, M. J. 1971, *Astrophys. J.,* **166,** L85.
Truran, J. W. Cameron, A. G. W. 1971, *Astrophys. Space Sci.,* **14,** 179.
Tutukov, A. V., Yungel'son, L. R. 1979, *Acta Astr.,* **23,** 665.
Underhill, A. B. 1983, *Astrophys. J.,* **265,** 933.
Underhill, A. B. 1984, *Astrophys. J.,* **276,** 583.
Underhill, A. B., Bhatia, A. K. 1984, *Bull. Am. astr. Soc.,* **16,** 492.
Vanbeveren, D. 1983, *Astr. Astrophys.,* **119,** 239.
van den Bergh, S. 1980, in *Type I Supernovae,* Ed. J. C. Wheeler, Univ. Texas & McDonald Obs., Austin, p. 11.
van den Bergh, S. 1982, *Bull. Astr. Soc. India,* **10,** 199.
van den Heuvel, E. P. J. 1981a, in *IAU Symp.* 93: *Fundamental Problems in the Theory of Stellar Evolution,* Eds D. Sugimoto, D. Q. Lamb & D. N. Schramm, D. Reidel, Dordrecht, p. 155.
van den Heuvel, E. P. J. 1981b, in *IAU Symp.* 95: *Pulsars,* Eds W. Sieber & R. Wielebinski, D. Reidel, Dordrecht, p. 379.
van den Heuvel, E. P. J., Habets, G. 1984, *Nature,* **309,** 598.
van den Heuvel, E. P. J., Taam, R. E. 1984, *Nature,* **309,** 235.
Weaver, T. A., Axelrod, T. S., Woosley, S. E. 1980, in *Type I Supernovae,* Ed. J. C. Wheeler, Univ. Texas & McDonald Obs., Austin, p. 113.
Webbink, R. F. 1979, in *IAU Coll* 53: *White Dwarfs and Variable Degenerate Stars,* Eds H. Van Horn & V. Weidemann, Univ. Rochester Press, p. 426.
Webbink, R. F. 1980, in *IAU Symp.* 88: *Close Binary Stars: Observations and Interpretation,* Eds M. J. Plavec, D. M. Popper & R. Ulrich, D. Reidel, Dordrecht, p. 561.
Webbink, R. F. 1984, *Astrophys. J.,* **277,** 355.
Webbink, R. F., Rappaport, S., Savonije, G. J. 1983, *Astrophys. J.,* **270,** 678.
Weidemann, V., Koester, D. 1983, *Astr. Astrophys.,* **121,** 77.
Wheeler, J. C. (Ed.) 1980, *Type I Supernovae,* Univ. Texas & McDonald Obs., Austin.
Wheeler, J. C., Hansen, C. 1971, *Astrophys. Space Sci.,* **11,** 373.
Whelan, J. A. J., Iben, I. 1973, *Astrophys. J.,* **186,** 1007.
Woosley, S. E., Axelrod, T., Weaver, T. A. 1984, *Stellar Nucleosynthesis,* Eds C. Chiosi & A. Renzini, D. Reidel, Dordrecht (in press).
Woosley, S. E., Weaver, T. A. 1982, in *Essays on Nuclear Astrophysics,* Eds C. Barnes, D. Clayton & D. N. Schramm, Cambridge Univ. Press, p. 377.
Woosley, S. E., Weaver, T. A. 1984, Preprint.
Woosley, S. E., Weaver, T. A., Taam, R. E. 1980, in *Type I Supernovae,* Ed. J. C. Wheeler, Univ. Texas & McDonald Obs., Austin, p. 96.
Yungel'son, L. R., Masevich, A. G. 1983, *Astrophys. Space Phys. Rev.,* **2**, 29.

J. Astrophys. Astr. (1984) **5,** 403–423

On the Supernova Remnants Produced by Pulsars

G. Srinivasan, D. Bhattacharya* & K. S. Dwarakanath
Raman Research Institute, Bangalore 560080

Received 1984 June 4; accepted 1984 August 16

Abstract. We conclude that pulsar-driven supernova remnants (SNRs) are extremely rare objects. Indeed an analysis of the known sample of plerions suggests a very low birthrate ~ 1 in 240 years. Long-lived and bright plerions like the Crab nebula are likely to be produced only when the pulsar has an initial period ~ 10–20 milliseconds and a field ~ 10^{12} G. Such pulsars inside rapidly expanding shell remnants should also produce detectable plerions. The extreme rarity of SNRs with such hybrid morphology leads us to conclude that these pulsars must have been born with an initial period larger than ~ 35–70 milliseconds.

Key words: supernova remnants—plerions—pulsars

1. Introduction

It is 50 years since the publication of the historic paper by Baade & Zwicky (1934) in which they advanced the hypothesis that Supernovae (SN) are the result of formation of neutron stars in the centres of ordinary stars. Detailed stellar evolution calculations done in recent years have confirmed this brilliant conjecture; it is now generally accepted that this is indeed the origin of Type II Supernovae. On the other hand, according to the current consensus no stellar remnant is left behind in a Type I supernova; the star completely disrupts (see for example, Trimble 1983). In spiral galaxies of morphology similar to ours the frequency of Type I and Type II SN are roughly equal (Tammann 1974).

Though no supernova has been sighted in our galaxy since the time of Kepler, it is generally believed that they occur once in about 30 years as suggested by historical observations (Clark & Stephenson 1977a, b). At any rate, Supernovae do leave behind relatively long-lived remnants (SNRs). In all about 140 SNRs are known in the Galaxy. Most of them have the morphology of shells with hollow interiors such as Tycho, Kepler and SNR 1006. However, the best studied SNR, namely the Crab nebula, has a distinctly different morphology: it has a filled-centre appearance with no limb-brightening. For a long time, the Crab nebula was unique in this respect. Weiler (1969) and Weiler & Seielstad (1971) first drew attention to the fact that 3C58 has a morphology similar to that of the Crab. Since then, the list of such filled-centre remnants, which have come to be known as ‘plerions’, has grown to a modest number of 7 or 8 (Weiler 1983). Several others (Radhakrishnan & Srinivasan 1978, 1980a; Weiler &

* Joint Astronomy Program, Department of Physics, Indian Institute of Science, Bangalore 560012.

Shaver 1978; Weiler & Panagia 1980) have suggested that these plerions may, like the Crab nebula, be produced and maintained by an active central pulsar.

Even if neutron stars are associated with only Type II supernova events, it is remarkable that until recently pulsars were associated with only two SNRs, namely the Crab and Vela X, both of which, curiously, are of *filled-centre* morphology. The standard explanation for the poor pulsar–SNR association invoked statistical factors such as beaming of pulsars, interstellar smearing of the pulses, low fluxes, *etc.* (Manchester & Taylor 1977). Radhakrishnan & Srinivasan (1980a) suggested that the above-mentioned statistical arguments were unsatisfactory, since an active pulsar inside a shell remnant will produce a centrally-condensed nebula like the Crab, which should be seen from any viewing geometry. They argued that the hollowness of the interiors of young shell SNRs is consistent with the *absence* of a central pulsar in them. This did not, of course, rule out the possibility that there could be a central neutron star which for some reason was not an active pulsar. A possible reason for this was suggested by Shukre & Radhakrishnan (1982) who proposed that a neutron star may not function as a pulsar unless its magnetic field lies in a narrow 'window' centred around the Crab value. An alternative possibility that the magnetic fields of neutron stars are built up after their birth over a long period of time has also been recently discussed in literature (Blandford, Applegate & Hernquist 1983; Woodward 1984). In this scenario the SNR would have faded away before the neutron star turned on as a pulsar.

Recently, however, a third pulsar–SNR association was found in the Galaxy, but this time the SNR MSH 15–52 had a shell morphology (Seward & Harnden 1982). There is no central radio emission surrounding the pulsar, although there is an extended X-ray synchrotron nebula. In view of this latest pulsar–SNR association one must admit the possibility that in all the shell remnants there are perhaps functioning pulsars which we do not see for the statistical reasons mentioned above, and *which do not produce plerions of sufficient surface brightness.* This might happen, for example, if pulsars inside shell remnants have relatively long periods (Radhakrishnan & Srinivasan 1983).

If pulsars are associated with every supernova explosion, then the birthrate of pulsars must be consistent with the frequency of Supernovae, and the birthrate of SNRs. Current estimates of pulsar birthrate of 1 in 20–40 yr (Taylor & Manchester 1977; Vivekanand & Narayan 1981) are indeed consistent with the previously mentioned supernova rate, and the recent estimates of the birthrate of shell remnants of 1 in ~ 30 yr (Srinivasan & Dwarakanath 1982; Mills 1983). However, in view of the fact that a pulsar and/or a plerion has been seen in only two shell remnants (MSH 15–52 and G 326.3 –1.8) (Weiler 1983), and the general absence of point X-ray sources (Helfand 1983) within the shells, the 'agreement' between the birthrate of shell SNRs and that of the pulsars appears puzzling.

On the other hand, since one expects an active pulsar in plerions, it is important to confront the birthrate of pulsars with the birthrate of plerions. In this paper, we shall address this important question. In Section 2, we derive a birthrate for Crablike SNRs assuming that all of them are similar to the Crab nebula in every respect, namely, the pulsars powering them have the same characteristics as the Crab pulsar, and their *initial* velocity of expansion is the same as that of the Crab nebula. Given these assumptions, the relative lifetime of such nebulae will depend on the density of the interstellar medium into which they are expanding. Using a slight variant of the model of the interstellar medium given by McKee & Ostriker (1977), we derive a mean birthrate of plerions ~ 1 in 240 yr.

In Section 3, we relax the assumption that the Crab nebula is a prototype. Pulsars are allowed to have a range of initial periods and fields. But we model all the plerions in analogy with the Crab nebula, namely that *their boundaries were accelerated by the energy lost by the pulsar* (Trimble & Rees 1970). We conclude that pulsar-driven supernova explosions are very rare events, and that long-lived and bright remnants like the Crab are even more rare.

Since the conclusion from Section 2 and 3 is that Crablike supernova remnants are extremely rare, in Section 4 *we move away from the pulsar-driven scenario to the standard model in which the supernova ejecta are accelerated by a shock wave*; the pulsar plays no dynamical role. We evolve plerions produced by pulsars inside rapidly expanding shells and compare the expected number of such Plerions implied by the generally accepted pulsar birthrate with observations. This forces us to the conclusion that the initial periods of pulsars must be much greater than 20 ms.

In Section 6, we estimate the characteristics of pulsars in the historical shell SNRs. From limits on their central surface brightness we conclude that their initial periods must have been larger than ~ 35–70 ms.

2. Birthrate of crablike remnants

Weiler (1983) has listed possible and probable SNR candidates with a filled-centre morphology. Of these, some have a surrounding shell. From this list, we have selected the remnants given in Table 1 for a birthrate calculation which will be done in this

Table 1. The adopted sample of plerions.

Source	Flux (S) at 1 GHz Jy	Distance (d) kpc	Luminosity Jy kpc^2	Ref.
G 21.5 – 0.9	6.4	4.8	147	1, 2
G 74.9 + 1.2	8.6	12	1238	1, 3
Crab	1000	2	4000	
Vela X	1100	0.5	275	
3C 58	33	8[a]	2112	1
		2.6	223	4
G 326.3 – 1.8 (centre)	40	2[b]	160	1, 5
		4.6	846	
MSH 15 – 52	0.1	4.2	1.6	7, 8, 5
G 5.3 – 1.1	37	3	333	1, 6
G 328.4 + 0.2	15	20	6000	1, 5

[a] The distance to 3C 58 remains highly controversial, as does its association with SN 1181. Following Weiler (1983) we adopt a distance of 8 kpc.

[b] Caswell *et al.* give a distance of 1.5 kpc, although they do not rule out a larger distance of 4.6 kpc. They regard the latter distance as unreliable without independent confirmation. The Σ-D relation for Galactic SNRs given by Mills (1983) yields a distance of 2.2 kpc. Hence we shall assume a distance of ~ 2 kpc.

References:

1. Weiler (1983)
2. Becker & Szymkowiak (1981)
3. Kazes & Caswell (1977)
4. Green & Gull (1983)
5. Caswell *et al.* (1975)
6. Milne & Dickel (1971)
7. Manchester & Durdin (1983)
8. Caswell, Milne & Wellington (1981)

section. A few comments are in order as to why the following remnants listed by Weiler have been excluded from our sample.

RCW 103 : No *extended* central emission is seen either in X-ray or in radio. A compact X-ray source is seen but its nature is not clear.

W 28 : There is no reliable distance estimate to the source.

W 50 : Though there is a condensed star at the centre, it is almost certainly not a standard pulsar.

CTB 80 : There seems to be some doubt as to whether the radio morphology is compatible with the identification as a plerion.

W 44 : Again, we feel that there is no clear evidence of centrally peaked emission within the shell.

Since we will be allowing for an incompleteness factor of 3, even if some of the sources we have rejected are 'legitimate' plerions it should not affect the birthrate derived.

In order to proceed with an estimate of the birthrate of plerions, one must have an evolutionary scenario for them from which one can derive their ages.

2.1 *The Evolution of Plerions*

In their pioneering paper, Pacini & Salvati (1973; hereinafter PS) discussed the evolution of the magnetic field, particle content and luminosity of the nebula produced and maintained by a central pulsar. After the initial phase, which relates to the explosion itself, there are two distinct phases of evolution:

(1) $t < \tau_0$: where $\tau_0 = P_0 / 2\dot{P}_0$ is the initial characteristic slowdown time of the pulsar. For the Crab pulsar $\tau_0 \sim 300$ yr.

(2) $t > \tau_0$: in this phase the nebular radius increases, the pulsar output decreases, and consequently the nebular luminosity decreases. Many of the observed properties of the Crab nebula can be successfully accounted for by this model.

PS assumed that the nebular boundary was expanding freely, even for $t > \tau_0$. This is certainly so for the Crab nebula *at the present time.* But if one wants to evolve the Crab nebula to a much older age, then one must modify the evolutionary scenario of PS to take into account the deceleration of the expansion at later times. This was first done by Weiler & Panagia (1980), and more recently by Reynolds & Chevalier (1984). Weiler and Panagia argued that the boundary of the nebula will decelerate and enter the adiabatic phase of expansion at $t \sim \tau_0$, the time when the pulsar would have lost half its rotational energy. According to them, the two youngest plerions, namely the Crab nebula and 3C 58 (probably the remnant of SN 1181) are entering, or are already in the adiabatic phase. However, in our opinion there is no immediate connection between the initial characteristic slowdown time of the pulsar and the time when the freely expanding filamentary shell will be significantly decelerated. Though the pulsar ceases to have a significant effect on the dynamics of the shell beyond $t > \tau_0$— *if ever it did*– the question of deceleration is determined by the mass in the ejecta and the density of the interstellar medium (ISM) into which it is expanding. It is generally accepted that the expanding shell will enter the adiabatic or Sedov phase only when the mass swept up far exceeds the mass ejected (Woltjer 1972). The time when this will occur depends on the mass ejected, the initial velocity of expansion, and the density of the

ISM. Since the expansion velocity of the Crab nebula (1700 km s^{-1}) is much less than the expected initial velocity of the shell SNRs($\sim 10^4$ km s^{-1}) it will take a much longer time for the former to sweep up a given amount of mass. At any rate, observations indicate that the filaments in the Crab nebula have not decelerated measurably.

In the standard model, the ISM consists of cold dense clouds in pressure equilibrium with the warm intercloud medium with a density $n_w \sim 0.3 cm^{-3}$ (Spitzer 1978). According to McKee & Ostriker (1977), however, the intercloud medium is a hot, low-density gas ($n_H \sim 0.003$ cm^{-3}). Although there is ample evidence for the existence of such a low-density coronal gas, there are strong observational reasons to believe in the presence of a denser intercloud medium also. Radhakrishnan & Srinivasan (1980b) have argued that a substantial fraction of the volume of the intercloud medium must be occupied by the denser component ($n_w \sim 0.3$ cm^{-3}). If one accepts this picture, then one is led to the conclusion that a fraction of SNRs must be expanding in the denser medium and must therefore suffer significant deceleration. Recent analyses of the evolution of shell-type SNRs also lend support to the above picture of the ISM (Higdon & Lingenfelter 1980; Srinivasan & Dwarakanath 1982).

Let us now estimate the time t_0 at which an expanding remnant like the Crab nebula will experience deceleration. Various observations suggest that the mass in the filaments of the Crab is $\sim 1\ M_\odot$ (Henry & MacAlpine 1982). If the Crab is expanding in the coronal gas, t_0 will be $\gtrsim 8000$ yr, while it will be $\gtrsim 1700$ yr if it is expanding in the denser component of ISM (at $t = t_0$ the mass swept up equals the mass ejected). For $t \gg t_0$ the radius of the nebula will increase as t^η with $\eta = 0.4$. With this modification one can easily extend the results of PS, as was done by Weiler & Panagia (1980).

In what follows we shall confine our attention to radio observations of plerions. For completeness, we give below the formulae for the radio spectral luminosity (for $\nu < \nu_c$).

$$\tau_0 < t < t_0: \qquad L_\nu \propto t^{-2\gamma} \nu^{(1-\gamma)/2}, \tag{1}$$

$$t \gg t_0 > \tau_0: \qquad L_\nu \propto t^{-2\eta\gamma} \nu^{(1-\gamma)/2}. \tag{2}$$

In Equations (1) and (2) γ is the exponent of the particle spectrum injected into the nebula by the pulsar (see PS and Weiler & Panagia 1980). The particle spectral index γ may be related to the radio spectral index a_R through the relation $\gamma = 1 + 2a_R$. For the Crab nebula, $a_R = 0.3$, implying $\gamma = 1.6$.

In Fig. 1, we have plotted Equations (1) and (2) which describe the decay of the radio spectral luminosity as a function of the age of the nebula. We have normalized the curve to the observed spectral luminosity of Crab nebula at 1 GHz. The solid curve is appropriate for the observed radio spectral index of the Crab nebula and the dashed curve for $\alpha_R \simeq 0.0$, such as for G 74.9 +1.2 or Vela X. Initially the luminosity drops as $t^{-2\gamma}$ and then flattens as the nebula decelerates. The sharp break in the curve is an artefact of the approximation that the remnant expands freely upto $t = t_0$ and according to the Sedov solution beyond t_0. In reality, of course, the evolution of luminosity will be described by a smooth curve. The curves labelled n_w are appropriate if the nebula were expanding in the warm, dense intercloud medium and those labelled n_H describe the evolution if it were expanding in the hot, low-density gas.

The above discussion of a smooth transition to the ISM-dominated phase ignores a subtle effect pointed out by Reynolds & Chevalier (1984). They have argued that during this transition, a reverse shock wave is likely to compress the pulsar bubble resulting in

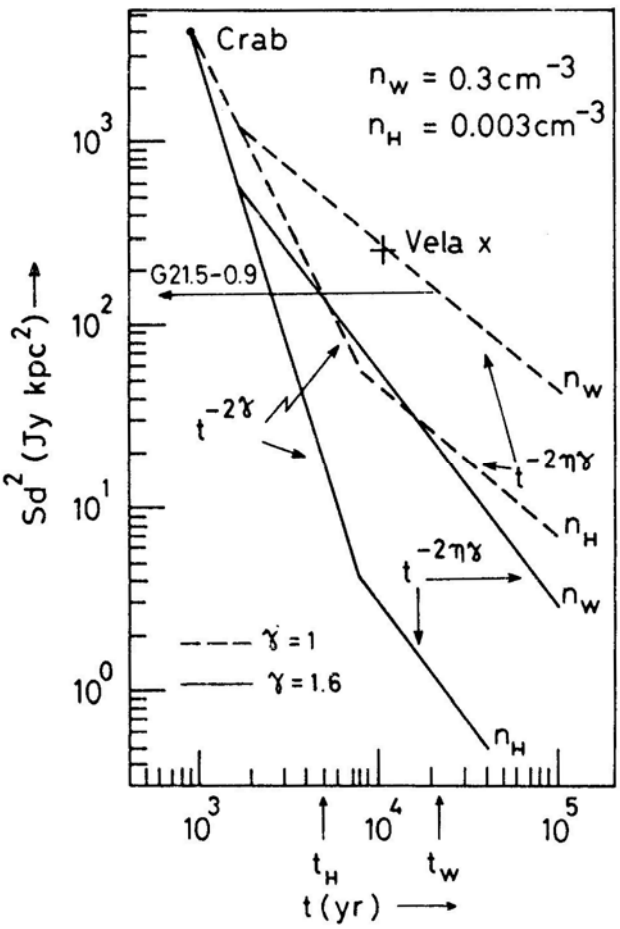


Figure 1. The secular decrease of the radio spectral luminosity of plerions at 1 GHz ($L_\nu \propto Sd^2$). The evolutionary tracks have been normalized to the luminosity of Crab nebula at an age of 1000 yr. $\gamma = 1 + 2a_R$ where α_R is the radio spectral index of the nebula; $a_R = 0.26$ for the Crab and 0.0 for G 21.5–0.9. The nebulae have been assumed to expand into regions with two typical densities, n_w and n_H; tracks corresponding to $\gamma = 1$ and $y = 1.6$ are shown. The estimated age of G 21.5 – 0.9 is ~ 4800 yr and ~ 23000 yr in the rarer and the denser media respectively.

a discontinuous increase in the plerion luminosity. Also, in the model of Reynolds & Chevalier (1984), the plerion radius increases as $t^{0.3}$ instead of $t^{0.4}$. However, the conclusions drawn from our model will in no way be altered by the above-mentioned effects, for these further increase the lifetime of the plerion.

2.2 *Birthrate*

We will now estimate the birthrate of plerions under the following assumption, namely that *the pulsars in all of them are identical to the Crab pulsar and their expansion velocities are the same as that of the Crab nebula.*

Of the plerions listed in Table 1, the oldest one is presumably G 21.5–0.9 since except for MSH 15–52 (centre) it is the least luminous one. One knows that the

plerionic component of MSH 15–52 cannot be older than 1600 years, the characteristic age of the pulsar. Its low surface brightness must be due to reasons other than old age and will be discussed in Section 5. Of course, a source more luminous than G 21.5 – 0.9 *need not be younger if the latter is expanding in the hot medium and the former in the denser medium* (See Fig. 1) but we shall correct for this below.

It now remains to estimate the age of the least luminous source. Its age estimated from Fig. 1 (the γ=1 tracks since its spectral index is 0.0) is ~ 4800 yr (if expanding in the hot medium) and ~ 23000 yr (if expanding in the warm medium). Since not all the plerions have a spectral index of 0.0, we shall assume an 'average' value of $\alpha_R \simeq 0.15$ implying $\gamma = 1.3$. If one uses the evolutionary track corresponding to this 'average' γ, the above estimate of the age of G 21.5 – 0.9 gets modified to 3600 yr and 13000 yr, respectively. These numbers represent the lifetimes above this luminosity in the two media.

If f_H and f_w are the filling factors of hot and warm media respectively, then

$$N(>) = \frac{1}{\tau}(t_H f_H + t_w f_w); \qquad f_H + f_w = 1 \tag{3}$$

where N (>) is the number of plerions with spectral luminosities greater than that of a given source, t_H and t_w are the lifetimes in the hot and the warm media respectively and τ the mean interval between Supernovae that produce plerions. From Table 1, we see that there are 9 sources more luminous than G 21.5 – 0.9. Using this number and the age estimates t_H and t_w given above for G 21.5–0.9 Equation (3) can be simplified to read

$$\tau = (1444 - 1044 f_H)\ \text{yr}. \tag{4}$$

Though it is felt that the sample of plerions is a reasonably complete one (Weiler 1983), since most of the observed plerions are relatively close by we shall take a conservative attitude and allow for an *incompleteness factor of* 3. Thus,

$$\tau \gtrsim (481 - 348 f_H)\ \text{yr}. \tag{5}$$

This should be regarded as an *upper limit* to the birthrate. If the low-density interstellar medium has a filling factor f_H= 0.7, as suggested by McKee & Ostriker (1977), Equation (5) gives a birthrate of *one in* 240 *yr.* The filling factor for coronal gas remains a highly uncertain one. Chevalier (1978), for example, has questioned the global nature of the coronal gas. A smaller filling factor will yield a lower birthrate.

2.3 *Implications*

It should be noted that even with a very large filling factor for the coronal gas one gets a birthrate of plerions much smaller than the pulsar birthrates in the literature which lie in the range of one in 10 years to one in 40 years (Phinney & Blandford 1981; Taylor & Manchester 1977; Vivekanand & Narayan 1981). It should be emphasized that we have already allowed for an incompleteness in the sample of plerions by a factor of 3. This makes the above discrepancy a very glaring one.

The first conclusion that suggests itself is that the pulsar birthrate must be grossly in error. It has been long recognized that it could be in error due to uncertainties in the beaming factor, the interstellar electron density, selection effects in pulsar searches, *etc.* M. Vivekanand (1984, personal communication) has made a systematic study of each

one of these factors and has put strong limits on the pulsar birthrate around a mean of one in ~ 40 yr. The other possible conclusion is that the evolutionary scenario used above to estimate the ages of plerions is questionable. It should be recalled that the basic assumption we have made is that Crab pulsar and the nebula are prototype objects. To be more explicit, we have assumed

(1) The initial period of *all* pulsars is the same as that of the Crab pulsar.
(2) The surface magnetic field of *all* pulsars is the same as that of the Crab pulsar.
(3) The expansion velocity of the nebular boundary, in every case, is the same as that of the Crab nebula.

In the next two sections, we shall relax all of the above three assumptions.

3. Pulsar-driven supernova remnants

As may be seen from PS, the spectral luminosity of a nebula *at a given age* depends on the expansion velocity and the initial luminosity of the pulsar, which in turn, is determined by its initial period and the surface magnetic field. To illustrate this, we rewrite below the expression derived by PS for the radio spectral luminosity for times $t > \tau_0$ (Equation 5.7 of PS) explicitly displaying the pulsar parameters.

$$L_\nu(t) \propto B_*^{(3-5\gamma)/2} P_0^{2(\gamma-2)} V^{-3(1+\gamma)/4} t^{-2\gamma} \nu^{(1-\gamma)/2}. \quad (6)$$

In the above equation B_* is the surface magnetic field of the pulsar, P_0 its initial period and V the expansion velocity. As was mentioned before, $\gamma = 1.6$ for the Crab nebula, and the above formula will read as

$$L_\nu \propto B_*^{-2.5} P_0^{-0.8} V^{-1.95} t^{-3.2} \nu^{-0.3}. \quad (7)$$

It can be seen from Equation (7) that the dependence of the luminosity on the pulsar field and the velocity of expansion is quite strong. In view of this, in estimating the luminosity of a nebula for a given age, *one should not assume that the Crab nebula and its pulsar are prototypes.* We discuss this point in greater detail below.

3.1 *Expansion Velocity of Plerions*

One of the most remarkable aspects of Crab nebula is its very low expansion velocity compared to expansion velocities of ejecta in typical Supernovae. It has been well established that the kinetic energy of expansion of the filamentary shell, as well as the acceleration experienced by it in the past, can be understood in terms of the energy being derived from the *stored rotational energy of the newly born pulsar.* The pressure of the relativistic 'wind' from the pulsar and the magnetic field frozen into it pushed out the remaining mass and accelerated it to the present velocity. It was through such arguments that one was able to estimate the initial period of the Crab pulsar (Trimble & Rees 1970; PS; Bees & Gunn 1974). It is natural, therefore, to assume that the same is true of all the plerions, namely, that the boundary is expanding with a velocity which was given to it by the central pulsar while it still had a dynamical effect on it. Let $E_0^{\mathrm{R}} = \frac{1}{2} I \omega_0^2$ be the initial stored rotational energy of the pulsar. Here I is the moment of inertia of the neutron star and ω_0 is the initial angular frequency of rotation. Within the

initial characteristic slowdown time $\tau_0 = P_0/2\dot{P}_0$, the pulsar would have dumped approximately half this energy in the form of relativistic particles and magnetic field. If M_{ej} is the mass accelerated, then the velocity imparted to it by the pulsar can be estimated from the relation

$$\tfrac{1}{2} M_{ej} V^2 \simeq \tfrac{1}{2} E_0^R \tag{8}$$

In what follows *we shall assume that the mass ejected in all cases is roughly the same as in the Crab nebula, and hence the expansion velocity* $V \propto 1/P_0$.

3.2 *The Initial Characteristics of the Pulsars*

Although it is believed that the initial period of the Crab pulsar was 16 ms, according to conventional wisdom most pulsars at birth will be spinning much more rapidly with $P_0 \sim$ a few milliseconds. *For the present we shall adopt the conventional viewpoint that the initial period of pulsars can be anywhere between* 1 *to* 20 *ms with equal probability.*

In the previous section we assumed that all pulsars have the same surface magnetic field as that of the Crab pulsar. However, one knows that there is a wide distribution in the derived magnetic fields of pulsars, which range from 10^{11}–$10^{13\cdot5}$ G and there are strong reasons to believe that very few pulsars have magnetic fields very much less than $10^{12\ 5}$ G *at birth* (Radhakrishnan 1982). The pulsars with $B < 10^{12}$ G are presumably several millions of years old and consequently their field would have decayed. If this were not the case, it is very hard to understand why no pulsar has been found with a field less than the Crab value and whose period is < 150 ms, since with such low fields it will take a long time before their periods lengthen to 150 ms, and consequently the chance of detection is significant. [The binary pulsar PSR 1913 + 16 and the two recently discovered millisecond pulsars are believed to have low fields and short periods because of their evolution in binary systems (Radhakrishnan & Srinivasan 1981; Radhakrishnan & Srinivasan 1982)]. In the calculations to follow, *we shall therefore assume that the magnetic fields of pulsars at birth can lie anywhere in the range* $10^{12} - 10^{13\,5}$ *G with equal probability in equal logarithmic intervals.*

3.3 *The Evolution of the Nebula*

We are now ready to discuss the evolution of the luminosity of such pulsar-driven nebulae. Combining Equations (6) and (8) one gets

$$L_\nu \propto B_*^{(3-5\gamma)/2} P_0^{(11\gamma-13)/4} t^{-2\gamma} \nu^{(1-\gamma)/2}. \tag{9}$$

It will be recalled that this formula is valid only for $t > \tau_0$. As long as one stuck to the initial period of the Crab pulsar and its magnetic field, one was mainly interested in this phase. But since we will now allow pulsar periods and fields to take a range of values we will need the full evolutionary curve, both for $t < \tau_0$ and $t > \tau_0$.

This is shown in Fig. 2 where we have compared the evolution of the radio luminosity of different nebulae with the central pulsars having different fields and initial periods. Since we are now dealing with pulsar-driven nebulae *we have taken into account the acceleration of the nebular boundary during* $t < \tau_0$. In this phase, the expansion velocity of the nebula is not constant and is proportional to $t^{1/2}$; consequently a slight modification of the formulae given in PS is needed. Since this is fairly straightforward

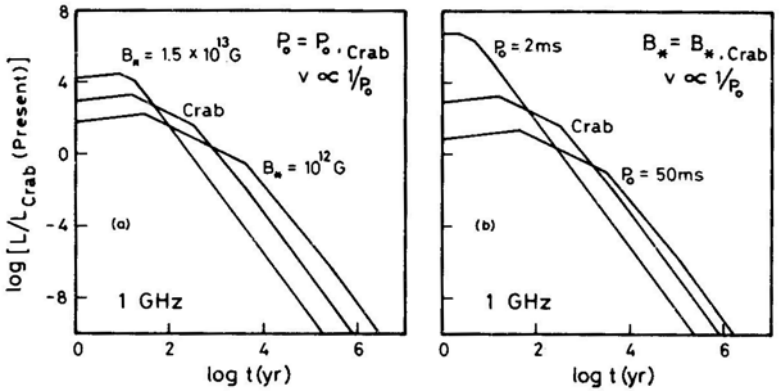


Figure 2. The evolution of radio spectral luminosity for pulsar-driven SNRs, in units of the present value for the Crab nebula, (a) The evolutionary tracks of two such nebulae powered by pulsars with the same initial period as the Crab pulsar, but with differing magnetic fields, are compared with the evolution of the Crab nebula, (b) Here the pulsars are assumed to have the same magnetic field but different initial periods.

we refer to Maceroni, Salvati & Pacini (1974) and Reynolds & Chevalier (1984) for further details. It will be seen from the figures that nebulae of the same age can have widely differing luminosities depending upon the characteristics of the central pulsar. The same information is displayed in a more concise form in Fig. 3. What is shown are *contours of constant luminosities for a given age* in the $B_* - P_0$ plane. All pulsars with initial characteristics which lie on a given contour will produce nebulae of the *same luminosity* at a given age. In Fig. 3a, the different contours correspond to different luminosities but the same age, whereas in Fig. 3b, different contours correspond to different ages for the same luminosity. It is clear from Figs 2 and 3 that it is not meaningful to assert that nebulae with luminosities greater than that of a given one

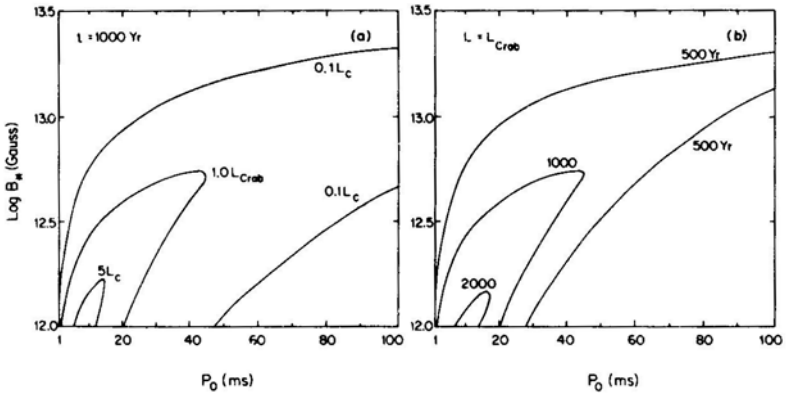


Figure 3. Contours of constant luminosity for pulsar-driven SNRs are shown in the $B_* - P_o$ plane; here B_* is the surface magnetic field and P_o the initial period of the pulsars. All pulsars born on a given contour will have the same luminosity at a specified age. (a) The three contours correspond to three different luminosities (measured in units of the present luminosity of Crab) and an age of 1000 yr. (b) The contours correspond to different ages, but the same luminosity, *viz.*, the present luminosity of Crab.

are necessarily younger, as was assumed in Section 2. We already saw the case of MSH 15 – 52 in which hardly any central radio emission was found even though it is of roughly the same age as the Crab nebula! Thus, if we relax the assumption made in Section 2, namely that the Crab pulsar and its nebula are prototypes, it is not possible to estimate the ages of plerions and therefore their birthrates.

3.4 *Expected Number of Plerions*

What one can ask is the following. Given a pulsar birthrate, and range of initial periods and magnetic fields, how many nebulae does one expect to see with luminosities above a specified value.

One has to now set a luminosity limit such that if a nebula has a luminosity greater than that, one is unlikely to miss it anywhere in the Galaxy. The flux from the Crab nebula will be 10 Jy at 1 GHz if placed at a distance of 20 kpc. The flux from a source with 1/l0th the luminosity of the Crab will be 1 Jy at the same distance. It is reasonable to suppose that many sources with flux greater than 1 Ay are unlikely to have been missed in surveys at frequencies around 1 GHz. It must be kept in mind that the plerions are likely to be more or less uniformly distributed in the inner Galaxy and that this flux limit corresponding to $L = 0.1\ L_{\mathrm{Crab}}$ refers to an extreme distance of 20 kpc. Therefore in what follows we shall take $0.1\ L_{\mathrm{Crab}}$ as the luminosity cut-off above which one should, in principle, be able to detect all sources in the Galaxy.

In Fig. 4, we have plotted several contours all corresponding to the above-mentioned

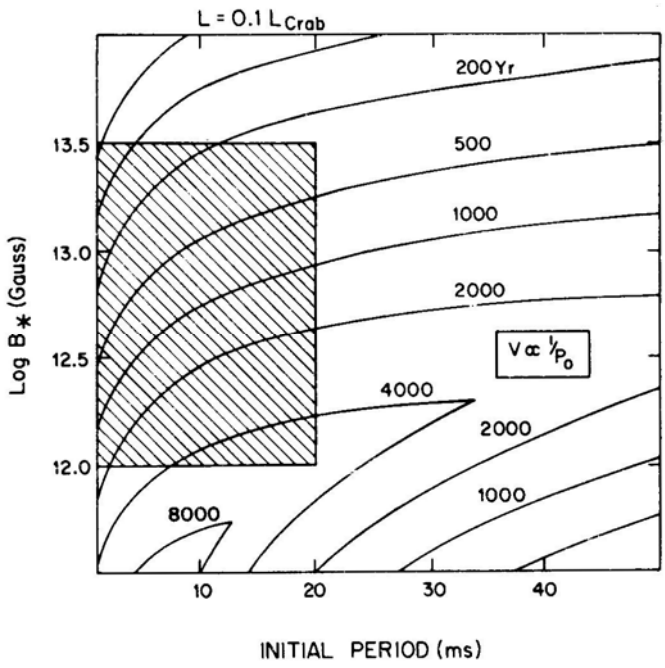


Figure 4. Pulsar-driven plerions. The contours of diffrent ages for a luminosity of $0.1\ L_{\mathrm{Crab}}$. In estimating the expected number of plerions with luminosities greater than the above-mentioned value we have assumed that pulsars are born anywhere inside the shaded region (see Section 3).

luminosity, namely, 0.1 $L_{\rm Crab}$. The labels on them represent the duration for which the nebulae are more luminous than the specified value, or in other words, their *lifetimes.* If τ is the mean interval between the birth of pulsars, then the number N of nebulae that one expects to see above the threshold luminosity is given by

$$N(>) = \frac{1}{\tau}\int_0^\infty t f(t)\,\mathrm{d}t. \tag{10}$$

Here $f(t)\,\mathrm{d}t$ is the probability that the nebula will have a lifetime between t and $t + \mathrm{d}t$. This formula is analogous to Equation (3) of Section 2. As we discussed above, the lifetime of the nebula depends on the initial parameters of the pulsar.

We shall assume a pulsar birthrate of 1 in 40 years, and that they are born with periods anywhere between 1 to 20 ms and log B_* (G) between 12 to 13.5 with equal probability.

Let $P(>t)$ be the probability that a nebula will have a *lifetime greater* than *t.* This is related to $f(t)$ we introduced in Equation (10) through

$$P(>t) = \int_t^\infty f(t')\,\mathrm{d}t'$$

or

$$f(t) = -\frac{\mathrm{d}P(>t)}{\mathrm{d}t}. \tag{11}$$

Let $a(t)$ be the area enclosed by the contour corresponding to age t and *within* the area A specified above in the plane (the hatched area in Fig. 4). Then, clearly,

$$P(>t) = a(t)/A. \tag{12}$$

From Fig. (4) and Equation (10), we find that there should be 35 nebulae whose luminosities are greater than 1/10th that of Crab nebula, or in other words, whose fluxes should be greater than 1 Jy *even if placed at* 20 *kpc.* However, as can be seen from Table 1 there are at most 4 objects with luminosities above 0.1 $L_{\rm Crab}$. There is, of course, a remote possibility that the sample is grossly incomplete, but this is extremely unlikely (see Weiler 1983). Thus we are once again faced with a gross discrepancy between the pulsar birthrate and the observed number of Crablike supernova remnants. There is, of course, the possibility that pulsar birthrates available in the literature are seriously in error. But it is very unlikely that this is so by a factor of eight! Another possibility is that most pulsars are born with fields less than 10^{12} G or greater than $10^{13\,5}$ G, or that their initial periods are much greater than 20 ms. The former may be ruled out since it is inconsistent with pulsar observations (Radhakrishnan 1982). The latter possibility must be taken seriously.

The most straightforward conclusion that one might draw is that *pulsar-driven supernova explosions,* such as SN 1054 A.D., *are very rare events.* This conclusion has also been arrived at independently by Bandiera, Pacini & Salvati (1984), Reynolds & Chevalier (1984) and Weiler (1983). It must be remarked that this conclusion is consistent with the one drawn above, namely, that the initial periods of pulsars might be much greater than 20 ms. Pulsars with such long initial periods will have very little stored rotational energy and in addition will take a very long time to get rid of it ($\tau_0 \propto P_0^2/B_*^2$). Hence they are unlikely to have any dynamical effect on the mass surrounding the newly born neutron star.

According to the standard picture of Supernovae the energy of the explosion is not derived from the stored rotational energy of the central pulsar but rather from a shock wave driven by the core bounce during the formation of the neutron star (Arnett 1980). The velocity of the shell is determined by the strength of the shock wave and the mass in the envelope, and is expected to be ~ 10,000 km s^{-1}. It is immediately obvious that a pulsar in the centre of such a rapidly expanding shell will produce a much weaker plerion. In the next section we shall discuss this scenario.

4. Pulsars inside rapidly expanding shells

We shall assume a typical expansion velocity of 10,000 km s^{-1} for the shell and a pulsar birthrate of 1 in ~ 40 yr. Once again, we shall allow the initial periods of pulsars to lie anywhere in the range 1–20 ms and their fields between 10^{12} to $10^{13.5}$ G. Since we have now decoupled the velocity of the shell from the initial period of the pulsar, the formulae derived in PS can once again be used to calculate the luminosity of the central nebula produced by the pulsar, as a function of its age. We shall now estimate the number of such plerions with luminosities greater than 0.1 L_{Crab}

In Fig. 5, we have plotted contours corresponding to the luminosity mentioned above for different ages. Following the procedure outlined in detail in Section 3, we arrive at the following conclusion. *There should be at least* 16 *plerions with luminosities greater*

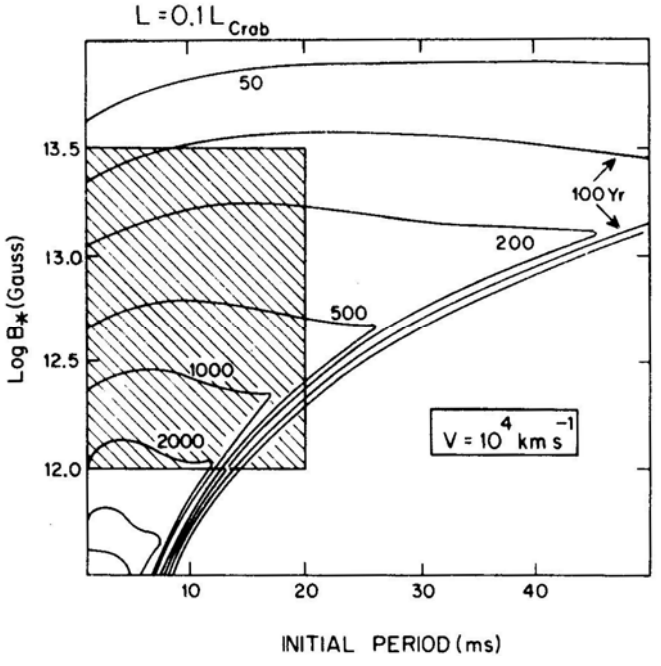


Figure 5. Plerions produced by pulsars inside standard shell SNRs expanding with a velocity 10^4 km s^{-1}. Once again the contours correspond to a luminosity of 0.1 LCrab. In Section 4, we have estimated the number of such plerions allowing the initial parameters of the pulsars to lie anywhere in the shaded region.

than 0.1 $L_{\rm Crab}$ *inside rapidly expanding shells.* This number will be even greater if the velocity of the shell is smaller than the assumed value of 10^4 km s^{-1}.

Although this number is less than the 35 predicted in the pulsar-driven scenario, the discrepancy with observations is even more glaring. From Table 1 we see that there are only four plerions above our luminosity limit. Of these, Crab clearly does not belong to the scenario in discussion. This leaves G 328.4 + 0.2, G 74.9 +1.2 and 3C 58. We shall now argue that *even* these three do not correspond to the present scenario of a pulsar inside a fast moving shock. When the shock sweeps up sufficient interstellar matter, one expects a pronounced radio and thermal X-ray shell. However, none of the three remnants mentioned above show any limb-brightening in the radio or an X-ray shell (Weiler 1983; Becker, Helfand & Szymkowiak 1982). One might argue that the radio shell is not pronounced because of very high central surface brightness due to the plerion. But one certainly expects to see an X-ray shell since the X-ray plerion will have a fairly small spatial extent compared to the diameter of the shell.

One is therefore once again faced with a dilemma! This can of course be reconciled with a much lower pulsar birthrate. But once again we reject it. We also reject the possibility that no stellar remnant is left behind in Supernovae that produce well-defined shells for the following reasons. For one thing, a pulsar *has* been detected in a standard shell (MSH 15 – 52). Even if the Type I Supernovae do not leave behind stellar remnants but well-defined shells, one still expects pulsars in at least half the shell SNRs, since the frequency of Type II Supernovae, which are believed to leave behind neutron stars, is roughly equal to the frequency of Type I Supernovae (Tammann 1974). The only alternative is to say that all pulsars are born in a very different scenario, such as the instability of accreting white dwarfs. Though this is a very distinct possibility, van den Heuvel & Taam (1984) have argued that pulsars born in such a manner will belong to a very different class and will be a minority.

We feel that the only resolution is the following. Namely, that the majority of pulsars are born outside the region we have considered in the B_*–P_o plane. This implies relatively long initial periods ($P_0 \gg$ 20 ms), contrary to conventional wisdom, and/or fields > $10^{13.5}$ G or <10^{12} G. We have already remarked that the statistics of pulsars is inconsistent with initial fields less than 10^{12} G and only a small fraction have fields greater than 10^{13} G (Radhakrishnan 1982). *Thus the only viable conclusion is that the initial periods of pulsars must be much greater than* 20 *ms.* It will be seen from Equation (7) that the dependence of the luminosity on the initial period is weaker than on the initial velocity of expansion. Consequently, the initial periods must be *substantially greater* than 20 ms, since increasing the expansion velocity of the boundary of the pulsar bubble from ~ 10^3 km s^{-1} to 10^4 km s^{-1} has not resolved the issue! The situation will be much worse in some current models of Supernovae (Chevalier 1977; Reynolds & Chevalier 1984) in which the pulsar bubble always expands with a relatively small velocity irrespective of the velocity of the expanding shell. These models will constrain the lower limit on the initial period much more.

5. The case of MSH 15–52

This is the third pulsar-SNR association in the Galaxy. Although the standard age of the shell is very large (~ 10^4 yr) compared to the characteristic age of the pulsar (1600 yr), the age derived from the Σ-t relation given by Srinivasan & Dwarakanath

(1982) is in excellent agreement with the pulsar age, thus confirming the pulsar–SNR association (Srinivasan, Dwarakanath & Radhakrishnan 1982). There is of course a possibility that it is an accidental superposition, as has been suggested by van den Bergh & Kamper (1984). But in this section we shall assume that the pulsar is in fact associated with the SNR.

Despite its young age, there is hardly any radio emission surrounding the pulsar (Manchester & Durdin 1983) but there is a pronounced X-ray nebula. Srinivasan, Dwarakanath & Radhakrishnan (1982) argued that the observed X-ray and radio luminosities are consistent with an initial period of the pulsar ~ 70 ms. They assumed that there was an inner shell which contained relativistic particles and which was expanding with a velocity similar to the filaments in the Crab nebula. In their paper, published soon after the discovery of the pulsar, the X-ray luminosity of the plerion was taken to be ~ 1/15th that of the Crab nebula. More recent estimates, however, give a value ~ 1/100 (Seward *et al.* 1984). This would modify the estimate of the initial period to ~ 115 ms, given the same assumptions, implying an age of 660 years for the pulsar and an average expansion velocity for the shell of ~ 24 000 km s^{-1}, an unacceptably large value. Hence we reject this estimate for the initial period.

As was already emphasized by Srinivasan, Dwarakanath & Radhakrishnan (1982), the assumption of an inner shell is a serious one. We now feel that it is more likely that the boundary of the plerion is the observed shell itself. This is the scenario discussed in Section 4.

Using the formalism of PS, we have calculated the evolutionary track for the radio and X-ray spectral luminosities appropriate for an expansion velocity ~ 10000km s^{-1} (as suggested by the characteristic age of the pulsar). These are shown in Fig. 6. It can be seen that the radio luminosity will be almost 10^4 times smaller than that of the Crab nebula for a whole range of initial periods. The predicted

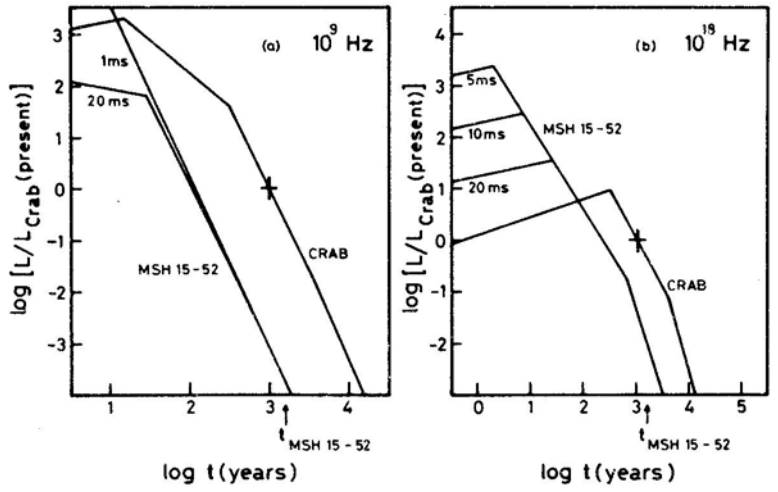


Figure 6. The expected luminosity of the central plerion in MSH 15–52 for several possible initial periods of the central pulsar. The measured magnetic field strength of the pulsar and an expansion velocity of 10^4 km s^{-1} were used; (a) refers to the radio and (b) to the X-ray luminosity; both are plotted in units of the present luminosity of the Crab at the respective frequencies. Shown for comparison are the evolutionary tracks of the Crab nebula.

X-ray luminosity ~ 10^{-2} that of the Crab nebula is also in good agreement with the observed value.

We see from Fig. 6 that the predicted luminosity of the plerion is very insensitive to the initial period of the pulsar; consequently we cannot draw any conclusions in this regard, as was done by Srinivasan, Dwarakanath & Radhakrishnan (1982) who used formulae appropriate for $t \lesssim \tau_0$.

6. What kind of pulsars may be present in historical shell SNRs?

According to the prevalent view historical shells such as Kepler, Tycho and SNR 1006 may be the remnants of Type I Supernovae which leave no compact remnants (Clark & Stephenson 1977a; Trimble 1983). Indeed, the absence of point thermal X-ray sources in them may be consistent with the above picture. In what follows, however, we shall assume that there are pulsars present and ask what kind of initial periods and fields would they have had? None of these shells show significant emission from the centre From the published maps, one can get an *actual* estimate of the central emission only for the case of Kepler. The brightness temperature in the central region is less than 2 K, while the limb has an average brightness temperature ~ 10 K. Central emission at such a low level is consistent with an optically thin shell whose thickness is, say, 1/5 the radius. But by attributing all of it to a possible central plerion, one can get an *upper limit* to its luminosity. From this we can put bounds on the parameters of a central pulsar. In the case of the other remnants, since no central emission is detected, one can get weaker limits on the pulsar parameters by postulating a central plerion with a surface brightness l/5th the average value for the remnant. Since we know the ages of these remnants, we can estimate their average expansion velocities. The fluxes and distances used are summarised in Table 2. In Fig. 7, we have plotted for each of these remnants contours corresponding to $\Sigma_{\text{plerion}} = f\,\Sigma_{\text{average}}$ The meaning of these contours is the following. Consider the one labelled 'Kepler'. If there is an active pulsar in its centre, then it could not have had an initial period and a field in the region enclosed by the contour. We have also included RCW 103, since there is a point X-ray source inside it. Its age was estimated to be 740. yr using the Σ–t relation given by Srinivasan & Dwarakanath (1982). It should be remarked that for all the remnants except Kepler the limit on the excluded region for the pulsar is very weak; one has been generous in admitting a central plerion with as large a surface brightness as $f\,\Sigma_{\text{average}}$ *A more realistic value for the central surface brightness will increase the excluded region considerably*, bringing them closer to the contour for Kepler. We see from Fig. 7 that if there are pulsars in these remnants, they must all have fields significantly greater than 10^{13} G or lie to the right of the contours. Although one is dealing with a very small sample of historical remnants, it is striking that the conclusion is quantitatively similar in each case. Hence this may be statistically significant and suggest that pulsars in all the shells must have been born outside such an excluded region. In a prescient paper, Pacini (1972) arrived at the remarkable conclusion that the hollowness of the historical shells is consistent with the presence of very high field pulsars in them (~ 10^{14} G). This may indeed be so in specific cases. But this cannot apply to the majority of shells for the following reason. In Fig. 7 we have shown a histogram of the distribution of pulsar fields at birth. This has been derived from Fig. 13 of Radhakrishnan (1982). It is seen that very few pulsars have fields greater than 10^{13} G. This would imply that for the

Table 2. Historical shells considered.

Source	Distance d kpc	Angular diameter arcmin	Size pc	Age yr	Average velocity of expansion km s^{-1}	$\frac{L_{\text{Plerion}}*}{L_{\text{Crab}}}$	Ref.
SNR 185	2.5	39	28	1800	7800	< 0.016	1, 2
SNR 1006	1.3	34	12.8	980	6500†	< 0.002	1, 3
RCW 103	3.3	9.4	9	740[a]	6000	< 0.014	1, 4
Kepler	3.5[b]	3.2	3.3	380	4300†	$\lesssim 0.012$	1, 5, 6
Tycho	3	7.9	6.9	410	8400†	< 0.026	1, 7, 8, 9

* Luminosity attributed to a possible central plerion.

[a] This is not a historical remnant. Nevertheless it is an important one for our discussion since a point X-ray source has been detected in it. We have estimated its age using the Σ–t relation given by Srinivasan & Dwarakanath (1982).

[b] Danziger & Goss (1980) have significantly improved upon the standard estimate of $d \sim 10$ kpc.

† These average velocities tend to be much greater than the measured values (Hesser & van den Bergh 1981; van den Bergh & Kamper 1977; Kamper & van den Bergh 1978). It is conceivable that these remnants have decelerated in recent times.

References:

1. Clark & Caswell (1976)
2. Caswell, Clark & Crawford (1975)
3. Milne (1971)
4. Caswell *et al.* (1980)
5. Danziger & Goss (1980)
6. Gull (1975)
7. Duin & Strom (1975)
8. Gorenstein, Seward & Tucker (1983)
9. Storm, Goss & Shaver (1982)

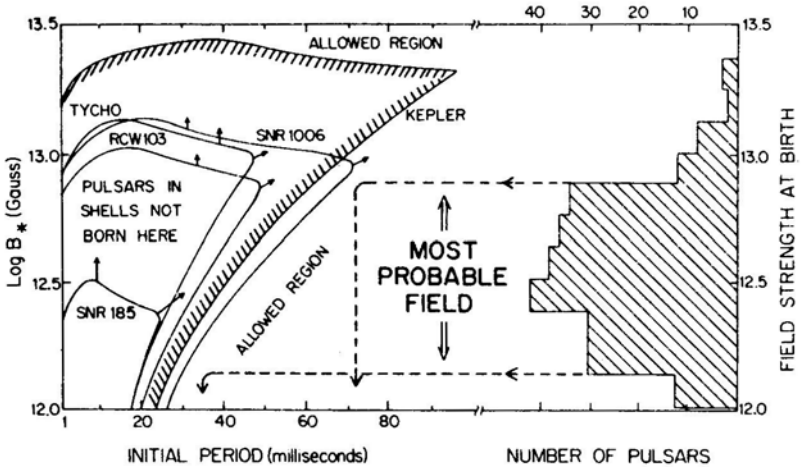


Figure 7. Pulsars inside the historical shells. The contours correspond to surface brightness of an assumed central plerion equal to fth the average surface brightness of the SNR; the appropriate ages and expansion velocities inferred from their sizes were used. If pulsars in these shells were born inside the region enclosed by the contours, then the plerions produced by them should have been easily detected. The arrows on the contours indicate that the excluded region is likely to be much larger. Also shown is the histogram of pulsar fields at birth. Although pulsars in these remnants could have been born anywhere outside the region enclosed by the contours, the histogram suggests that for the majority of them the initial periods will be greater than 30 to 70 ms.

majority of pulsars that could be there in shells the period at birth must have been longer than 35–70 ms. We wish to regard this as a lower limit for the initial periods since the analysis was done not on the basis of measured fluxes from their centres, but on the basis of upper limits on them. Since the shell SNRs constitute more than 80 per cent of the sample of SNRs, *the above conclusion applies to the majority of all pulsars.*

7. Another way out?

In addressing the question of the poor pulsar–SNR association we have adopted the point of view that there are functioning pulsars in all SNRs or in at least roughly half of them (produced by Type II Supernovae). We have then attempted to put constraints on the initial characteristics of the pulsars consistent with the absence of pronounced plerions in the shells.

There is however an alternative way out of this dilemma, and that is to say that there are neutron stars in all young SNRs, but not substantially endowed with fossil fields, and therefore not functioning as pulsars. A variety of mechanisms have been recently suggested for thermally driven magnetic field generation in neutron stars after their birth (Woodward 1978, 1984; Blandford, Applegate & Hernquist 1983 and references therein). According to the latter the timescale to build up the magnetic field to typical values observed in pulsars is $\sim 10^5$ yr. Long before this the SNRs would have faded away. It would appear that this scenario neatly explains the small number of SNRs of hybrid morphology. Two pulsars, however, are an embarrassment to the above picture, *viz.*, the Crab pulsar and PSR 1509 – 58 in MSH 15–52 (in the latter case a field of

1.5×10^{13} G has been presumably built up in at most ~ 9000 yr). Blandford, Applegate & Hernquist (1983) have suggested that in a rapidly rotating neutron star one may be able to tap the rotational energy to generate additional heat flux, thus enabling a rapid build-up of the field. Woodward (1984) has suggested that initially the magnetic fields of neutron stars may be built up by a Hall-field-limited battery effect. In this mechanism a saturation field which is proportional to the angular frequency of rotation (ω) will be built up in a timescale which will be proportional to ω^{-1} (Roxburgh 1966).

If such field build-up mechanisms are taken seriously, the fact that there are very few shell SNRs with central plerions seems to suggest that the neutron stars in them are unlikely to have been born spinning rapidly. For, otherwise, they will build up strong magnetic fields before the SNR disappears (like in MSH 15 – 52). It therefore seems to us that even if the magnetic fields of neutron stars are built up *after* their birth, one is forced to the conclusion that the majority of them must be born spinning slowly.

8. Conclusions

1. Our analysis of the sample of Crablike supernova remnants indicates that they are extremely rare objects. At present, only 7 or 8 such objects are known in the Galaxy, even though over 120 SNRs of shell morphology have been detected. In analogy with the Crab nebula and Vela X, it is reasonable to suppose that they are all powered by a central pulsar but whose beams are presumably missing us. If one assumes that all of them are remnants of Supernovae similar to SN 1054 AD, then the observed number of plerions yields a birthrate of 1 in ~ 240 yr. This is in fact an upper limit since we have allowed 7 out of 10 explosions to occur in the extremely low-density coronal gas component of the interstellar medium. Remnants expanding in this medium will have a very short life-time and hence will increase the estimated birthrate. A smaller filling factor for the coronal gas than the one we have assumed will further drastically reduce the birthrate quoted above.

2. In analogy with the remnant of SN 1054 AD, we have assumed in Section 3 that the boundaries of all plerions are accelerated by the central pulsar, but allowed the initial periods and fields to have any value in a domain $1 \lesssim P_0 \lesssim 20$ ms and $10^{12} \lesssim B_* \lesssim 10^{13.5}$ G. It was found that given a pulsar birthrate of one in 40 years, there should be about 35 nebulae with luminosities greater than 0.1 $L_{\rm Crab}$. But in the sample of plerions, only four satisfy this criterion. Hence the majority of pulsars must be born outside the domain mentioned above, implying initial periods much greater than 20 ms. We dismiss the possibility of the majority having fields outside the range considered as inconsistent with pulsar observations. The alternative is to say that only in rare cases the pulsar accelerates the nebular boundary. Even if it does, only when the rare combination of 10 ms $\lesssim P \lesssim$ 20 ms and $B_* \sim 10^{12}$ G obtains, can the pulsar driven nebula be expected to be 'long-lived' and 'bright'. The particular nature of the Crab nebula must be understood in terms of the Crab pulsar having just these characteristics as surmised by Pacini (1972) a long time ago.

3. If the energy of the supernova is not derived from the pulsar, then it must be derived from the energy released in the formation of the neutron star. Hence we have studied the evolution of the luminosity of pulsar-produced nebulae inside rapidly expanding shells. Even in this case one should find 16 plerions with luminosities greater than 0.1 $L_{\rm Crab}$ inside standard shell SNRs. But there is not even a single such example.

Of the three sources (other than the Crab nebula) above this luminosity limit none show the expected X-ray shell or limb brightening in the radio. We conclude from this that pulsars inside shell SNRs must have initial periods substantially greater than 20 ms to be consistent with observations.

4. We have estimated the characteristics of pulsars in the historical shells from (generous) limits on the surface brightness of associated plerions (Fig. 7). In all cases, the bounds are similar, forcing us to the conclusion that pulsars in shell SNRs are born with periods greater than 35–70 ms. This provides strong support for the conclusion arrived at by Vivekanand & Narayan (1981) from an analysis of the periods and period derivatives of pulsars that the majority of them make their 'appearance' with periods $\gtrsim 100$ ms.

These conclusions once again raise but leave unanswered the fundamental question as to what determines when pulsars play a role in the acceleration of the nebular boundary.

Acknowledgements

We wish to thank V. Radhakrishnan, K. W. Weiler and L. Woltjer for stimulating comments and discussion. Critical comments on an earlier version of the manuscript by R. Chevalier, R. Nityananda and V. Trimble are gratefully acknowledged. DB acknowledges financial support by National Council of Educational Research and Training, New Delhi and thanks Raman Research Institute for extending research facilities.

References

Arnett, W. D. 1980, *Ann. NY. Acad. Sci.*, **336,** 366.
Baade, W., Zwicky, F. 1934, *Phys. Rev.*, **45,** 138.
Bandiera, R., Pacini, F. , Salvati, M. 1984, *Astrophys. J.*, (in press).
Becker, R. H., Helfand, D. J., Szymkowiak, A. E. 1982, *Astrophys. J.*, **255,** 557.
Becker, R. H., Szymkowiak, A. E. 1981, *Astrophys. J.*, **248,** L23.
Blandford, R. D., Applegate, J. H., Hernquist, L. 1983, *Mon. Not. R. astr. Soc,* **204,** 1025.
Caswell, J. L, Clark, D. H., Crawford, D. F. 1975, *Austr. J. Phys., Astrophys. Suppl. No.* 37, 39.
Caswell, J. L, Haynes, R. F., Milne, D. K., Wellington, K. J. 1980, *Mon. Not. R. astr. Soc,* **190,** 881.
Caswell, J. L., Milne, D. K., Wellington, K. J. 1981, *Mon. Not. R. astr. Soc,* **195,** 89.
Caswell, J. L, Murray, J. D., Roger, R. S., Cole, D. J., Cooke, D. J. 1975, *Astr. Astrophys.*, **45**,239.
Chevalier, R. A. 1977, in *Supernovae*, Ed. D. N. Schramm, D. Reidel, Dordrecht, p. 53.
Chevalier, R. A. 1978, *Mem. Soc astr. ItaL,* **49**, 497.
Clark, D. H., Caswell, J. L. 1976, *Mon. Not. R. Astr. Soc,* **174,** 267.
Clark, D. H., Stephenson, F. R. 1977a, *Mon. Not. R. astr. Soc,* **179**, 87p.
Clark, D. H., Stephenson, F. R. 1977b, *The Historical Supernovae,* Pergamon Press, Oxford.
Danziger, I. J., Goss, W. M. *l9W,Mon. Not. R. astr. Soc,* **190,** 47p.
Duin, R. M., Strom, R. G. 1975, *Astr. Astrophys.*, **39**, 33.
Gorenstein, P., Seward, F., Tucker, W. 1983, in *IAU .Symp.* 101: *Supernova Remnants and their X-ray Emission,* Eds J. Danziger & P. Gorenstein, D. Reidel, Dordrecht, p. 1.
Green, D. A., Gull, S. F. 1983, in *IAU Symp.* 101; *Supernova Remnants and their X-ray Emission,* Eds J. Danziger & P. Gorenstein; D. Reidei, Dordrecht, p. 335.
Gull, S. F. 1975, *Mon. Not. R. astr. Soc.*, **171**, 237.
Helfand, D. J. 1983, in *IAU Symp.* 101: *Supernova Remnants ana their X-ray Emission,* Eds J. Danziger & P. Gorenstein, D. Reidei, Dordrecht, p. 471.
Henry, R. B. C, MacAlpine, G. M. 1982, *Astrophys. J.*, **258,** IL

Hesser, J. E., van den Bergh, S. 1981, *Astrophys., J.*, **251,** 549.
Higdon, J. C, Lingenfelter, R. E. 1980, *Astrophys., J.,* **239,** 867.
Kamper, K. W., van den Bergh, S. 1978, *Astrophys. J.,* **224,** 851.
Kazes, I, Caswell, J. L. 1977, *Pulsars,* Freeman, USA.
Maceroni, C, Salvati, M., Pacini, F. 1974, *Astrophys. Space Sei.,* **28,** 205.
Manchester, R. N., Durdin, J. M. 1983, in *IAU Symp.* 101: *Supernova Remnants and their X-ray Emission,* Eds J. Danziger & P. Gorenstein, D, Reidel, Dordrecht, p. 421.
Manchester, R. N., Taylor, J. H. 1977, *Pulsars,* Freeman, USA.
McKee, C. F., Ostriker, J. P. 1977, *Astrophys. J.*, **218,** 148.
Mills, B. Y. 1983, in *IAU Symp.* 101: *Supernova Remnants and their X-ray Emission,* Eds J. Danziger & P. Gorenstein, D. Reidel, Dordrecht, p. 551.
Milne, D. K. 1971, *Austr. J. Phys.,* **24,** 757.
Milne D. K., Dickel, J. R. 1971, *Nature Phys. Sei.,* **231,** 33.
Pacini, F. 1972, in *The Physics of Pulsars,* Ed. A. M. Lenchek, Gordon & Breach, London, p. 119.
Pacini, F., Salvati, M. 1973, *Astrophys. J.,* **186,** 249 (PS).
Phinney, E. S., Blandford, R. D. 1981, *Mon. Not. R. astr. Soc,* **194,** 137.
Radhakrishnan, V. 1982, *Contemp. Phys.,* **23,** 207.
Radhakrishnan, V., Srinivasan, G. 1978, paper presented at *Asian-South-Pacific Regional Meeting in Astronomy,* Wellington, unpublished.
Radhakrishnan, V., Srinivasan, G. 1980a, /. *Astrophys. Astr.,* **1,** 25.
Radhakrishnan, V., Srinivasan, G. 1980b, /. *Astrophys. Astr.,* **1**, 47.
Radhakrishnan, V., Srinivasan, G. 1981, paper presented at *2nd Asian-Pacific Regional Meeting IAU,* Bandung.
Radhakrishnan, V., Srinivasan, G. 1982, *Current Sei.,* **51,** 1096.
Radhakrishnan, V., Srinivasan, G. 1983, in *IAU Symp.* 101: *Supernova Remnants and their X-ray Emission,* Eds J. Danziger & P. Gorenstein, D. Reidel, Dordrecht, p. 487.
Rees, M. J., Gunn, J. E. 1974, *Mon. Not. R. astr. Soc,* **167,** 1.
Reynolds, S. P., Chevalier, R. A. 1984, *Astrophys. J.,* **278,** 630.
Roxburgh, I. W. 1966, *Mon. Not. R. astr. Soc,* **132**, 201.
Seward, F. D., Harnden, F. R. 1982, *Astrophys. J.,* **256,** L45.
Seward, F. D., Harnden, F. R., Szymkowiak, A., Swank, J. 1984, *Astrophys. J.,* **281,** 650.
Shukre, C. S., Radhakrishnan, V. 1982, *Astrophys. J.,* **258,** 121.
Spitzer, L. 1978, *Physical Processes in the Interstellar Medium,* Wiley-Interscience, New York.
Srinivasan, G., Dwarakanath, K. S. 1982, *J. Astrophys. Astr.,* **3,** 351.
Srinivasan, G., Dwarakanath, K. S., Radhakrishnan, V. 1982, *Current Set,* **51,** 596.
Strom, R. G., Goss, W. M., Shaver, P. A. 1982, *Mon. Not. R. astr. Soc,* **200,** 473.
Tammann, G. A. 1974, in *Supernovae and Supernova Remnants,* Ed. C. B. Cosmovici, D. Reidel, Dordrecht, p. 155.
Taylor, J. H., Manchester, R. N. 1977, *Astrophys. J.,* **215,** 885.
Trimble, V. L., Rees, M. J. 1970, *Astrophys. Lett.,* **5,** 93.
Trimble, V. 1983, *Rev. mod. Phys.,* **55,** 511.
van den Bergh, S., Kamper, K. W. 1977, *Astrophys. J.,* **218,** 617.
van den Bergh, S., Kamper, K. W. 1984, *Astrophys. J.,* **280,** L51.
van den Heuvel, E. P. J., Taam, R. E., 1984, *Nature,* **309,** 235.
Vivekanand, M., Narayan, R. 1981, *J. Astrophys. Astr.,* **2,** 315.
Weiler, K. W. 1969, *PhD Thesis,* California Institute of Technology.
Weiler, K. W. 1983, in *IAU Symp.* 101: *Supernova Remnants and their X-ray Emission,* Eds J. Danziger & P. Gorenstein, D. Reidel, Dordrecht, p. 299.
Weiler, K. W., Panagia, N. 1980, *Astr. Astrophys.,* **90,** 269.
Weiler, K. W., Seielstad, G. A. 1971, *Astrophys. J.,* **163,** 455.
Weiler, K. W., Shaver, P. A. 1978, *Astr. Astrophys.,* **70,** 389.
Woltjer, L. 1972_S *A. Rev. Astr. Astrophys.,* **10,** 129.
Woodward, J. F. 1978, *Astrophys.* J.,**225,** 574.
Woodward, J. F. 1984, *Astrophys.*J.,**279,** 803.

J. Astrophys. Astr. (1984) **5,** 425–428

HI Absorption Observations of two Radio Sources near the Supernova Remnant G 127.1 + 0.5

W. M. GOSS *Kapteyn Astronomical Institute, P.O. Box 800, 9700 AV Groningen, The Netherlands* and *National Radio Astronomy Observatory*, P.O. Box 0, Socorro, New Mexico 87801, USA*

J. H. van Gorkom *National Radio Astronomy Observatory*, P.O. Box 0, Socorro, New Mexico 87801, USA*

Received 1984 April 3; accepted 1984 July 27

Abstract. The compact source 0125 + 628 in the centre of the galactic supernova remnant G 127.1+ 0.5 has been re-observed in HI absorption using the Westerbork Synthesis Radio Telescope (WSRT). The outer arm HI absorption at V= –95km s^{-1} has been confirmed. The absorption spectrum is similar to that of the nearby extragalactic source 0123 + 633. We discuss the arguments concerning an extragalactic origin of 0125 + 628 and conclude that it is most likely extragalactic and not an SS 433 type object.

Key words: HI absorption—galactic supernova remnants

1. Introduction

The nature of the compact nonthermal source in the centre of the galactic supernova remnant (SNR) Gl27.1 +0.5 has remained controversial. The debate has been whether the source is galactic and associated with the SNR or whether it is an extragalactic object. Arguments in favour of a galactic origin are:

(1) It is located almost precisely at the centre of the SNR (Caswell 1977).

(2) The source shows VLBI structure, in contrast to a nearby source (displaced by 8°) which is dominated by the effects of interstellar scattering (Geldzahler & Shaffer 1982; hereinafter GS).

(3) The SNR has morphological features which align with the Very Long Baseline Interferometry (VLBI) (core-dual-jet) structure (GS).

Arguments in favour of an extragalactic origin are:

(1) The optical spectrum is similar to that of a radio galaxy with a redshift of ~ 0.02 (Kirshner & Chevalier 1978; Spinrad, Stauffer & Harlan 1979).

(2) The HI absorption spectrum places it behind essentially all the galactic hydrogen at this longitude (Pauls *et, al.* 1982).

If 0125 + 628 is in fact galactic and associated with the SNR, it would be classified as an SS 433 type object and thus be quite rare. Thus a classification of the object is

* The National Radio Astronomy Observatory is operated by Associated Universities Inc. under contract with the National Science Foundation.

important. An ideal observational test would be a direct comparison of the HI absorption towards the point source and the SNR. The latter observation is, however, very difficult due to the extremely low surface brightness of the SNR at 21 cm. We have thus decided to follow the same method that was used for SS 433 (van Gorkom *et al.* 1982) and compare the HI absorption spectrum of 0125 + 628 with that of a nearby point source (0123 + 633), which is almost certainly extragalactic. This test only provides information about the distance of 0125 + 628 and provides no direct evidence concerning the distance of the SNR.

In this paper we present new HI absorption observations in the directions of 0125 +628 and 0123 + 633. The previous HI data (Pauls *et al.* 1982) had poor signal-to-noise ratio at the more extreme velocities ($V < -90 \mathrm{kms}^{-1}$) arising from the outer arm of the Galaxy. It is precisely these velocities which are crucial for any distinction between a galactic and an extragalactic source. The new data have a factor of three improvement in sensitivity. The previous results are confirmed, thus supporting an extragalactic origin of 0125 + 628. This led us to re-examine the arguments for and against a galactic origin of the object; contrary to GS we conclude that an extragalactic origin is more likely.

2. Observations

The 21cm HI observations were carried out in the autumn of 1982 using the Westerbork Synthesis Radio Telescope (WSRT). The observing procedure is similar to that described by Pauls *et al.* (1982) and van Gorkom *et al.* (1982). The 10 fixed telescopes were correlated with the more distant movable telescopes C and D. The spacings range from 1368 m to 2736 m for 0125+628 and 1350 to 2718 m for 0123 + 633 at intervals of 72 m. (The spacings below ~ 1300 m were not observed in order to avoid the effects of HI emission.)

Both 0123 + 633 and 0125 + 628 were observed for 12 hours with orthogonal linear polarizations. The total bandwidth was 0.625 MHz with 63 frequency channels and a velocity resolution of 2.5km s^{-1}. The rms noise in a single channel is 7mJy as compared to 21 mJy in the earlier observations reported by Pauls *et al.* The flux density of 0125 + 628 was 400 mJy in close agreement with the value measured by Pauls *et al.* (1982) in 1979. The source 48 W8 (Pauls *et al.*) or 0123 + 633 has a 21 cm flux density of 210mJy. The position at 21cm is $\alpha(1950) = 0.1^{h}23^{m}18^{s}.29 \pm 0^{s}.06$, $\delta(1950) = 63°19'54''.5 \pm 0''4$ [The position quoted by Caswell (1977) differs by 1.5arcmin.]

3. Results

In Fig. 1 we show the WSRT HI spectra of 0123 + 633 and 0125 + 628. Because of the choice of the central velocity of -60 kms^{-1} the strong absorption near 0 km s^{-1} (Pauls *et al.* 1982) is not fully covered. The spectrum of 0125 + 628 is in good agreement with the earlier WSRT data and the prominent line at -95 km s^{-1} ($\tau \sim 0.1$) is confirmed. The absorption spectrum of 0123 + 633 (which is 0°.5 displaced on the sky) is quite similar in form to 0125 + 628. In particular there are prominent lines at $V = -70$ kms^{-1} and $-97 \mathrm{kms}^{-1}$ ($\tau \sim 0.2$). The HI emission spectrum in this direction (Pauls *et al.* 1982) shows prominent lines up to a velocity of —100 km s^{-1}.

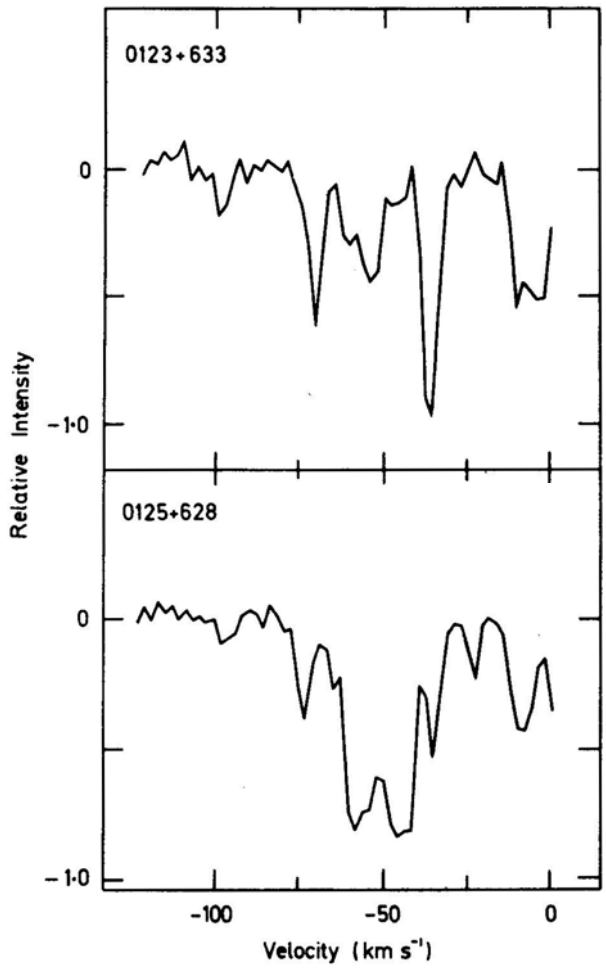


Figure 1. HI absorption spectra of 0123 + 633 (above) and 0125 + 628 (below) obtained with the WSRT. Velocity is with respect to the local standard of rest. The velocity resolution is 2.5 km s^{-1}.The intensity scale is relative intensity (–1 is complete absorption and 0 is no absorption). The flux densities of 0123 + 633 and 0125 + 628 are 210 and 400 mJy, respectively.

4. Discussion

The HI absorption data show that both 0125 + 628 and 0123 + 633 are behind essentially all the galactic HI in this direction. This provides a lower limit of 8 kpc to the distance. The comparison point source (< 10 arcsec in size) 0123 + 633 is probably an extragalactic source observed in projection near the SNR G 127.1 + 0.5. The spectral index between 49 and 21 cm is –0.33 ($S \propto \nu^{\alpha}$). The similarity in HI spectra between this source and 0125 + 628 suggests that 0125 + 628 is also probably extragalactic. The present result removes one of the stronger arguments of GS indicating that 0125 + 628 is galactic. GS also make a comparison with a nearby source 0241 + 622, which is 8° away. They find the structure on VLBI size scales for 0125 + 628 while the size of 0241 + 622 is dominated by the effects of scattering. GS conclude that either the path

length to 0125 + 628 through the galaxy must be substantially less than toward 0241 + 622, or there is a 'hole' in the interstellar scattering medium. The HI data show that the latter explanation must be correct, since the path length is at least 8 kpc. 'Holes' in the interstellar scattering medium have been found to be quite common in the galactic anti-centre direction (Dennison *et al.* 1984) as determined by low-frequency VLBI studies of measured angular sizes of background sources.

The possibility does remain that both 0125 + 628 and the SNR are at a distance of 8 kpc, although the surface-brightness diameter (Σ–D) relation (Mills *et al.* 1984) indicates a distance of 4 kpc for the SNR. However, if there is an internal energy source in the SNR the conventional Σ–D relation may not be valid. Although this possibility cannot be ruled out, there does seem to be compelling evidence that 0125 + 628 is not an SS 433 type object. The most important differences are:

(1) SS 433 has a stellar appearance in the optical and (2) has a composite spectrum consisting of a narrow emission-line system at rest superimposed on the broad-line fast-moving system (Margon *et al.* 1979). 0125 + 625 on the other hand has (1) a diffuse optical appearance and (2) the emission lines resemble a radio galaxy with a constant redshift of ~ 0.02 (Kirshner & Chevalier 1978; Spinrad *et al.* 1979). The SNR W 50 has enhanced bright regions ('ears') that are aligned with the jets from SS 433. G 127.1 + 0.5 is spherical with numerous hot spots and breaks in the shell. As reference to the map published by Salter, Pauls & Haslam (1978) indicates, there are many orientations in the SNR which would lead to alignment with the VLBI structure of 0125 + 638.

In summary, although SS 433 and 0125 + 628 do have many similarities in their radio properties, the current evidence suggests that the latter source is extragalactic.

Acknowledgements

The Westerbork Synthesis Radio Telescope (WSRT) is operated by the Netherlands Foundation for Radio Astronomy with the financial support of the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

References

Caswell, J. L. 1977, *Mon. Not. R. astr. Soc.*,**181,** 789.

Dennison, B., Thomas, M., Booth, R. S., Brown, R. L., Broderick, J. J., Condon, J. J. 1984, Preprint.

Geldzahler,R. J.,Shaffer,D.B.1982, *Astrophys. J.,* **260,** L69(GS).

Kirshner, R. P., Chevalier, R. A. 1978, *Nature*, **276,** 480.

Margon, B.,Ford, H.C.,Kate, J.I.,Kwitter,K.B., Ulrich, R.K.,Stone,R.P. S., Klemola,A. 1979, *Astrophys. J.***, 230,**L41.

Mills, B. Y., Turtle, A. J., Little, A. G., Durdin, J. M. 1984, *Austr. J. Phys.*, (in press).

Salter, C. J.,Pauls, T.,Haslam C.G.T.1978, *Astr. Astrophys.*, **66,** 77.

Spinrad, H.,Stauffer, J.,Harlan, E. 1979, *Pub. astr. Soc. Pacific,* **91,**619.

Pauls, T., van Gorkom, J. H., Goss, W. M., Shaver, P. A., Dickey, J. M., Kulkarni, S. 1982, *Astr. Astrophys.*, **112,** 120.

van Gorkom, J.H., Goss,W.M.,Seaquist, E.R.,Gilmore, W.S.1982, *Mon. Not. R. astr. Soc.,* **198,** 757.

J. Astrophys. Astr.(1984) **5,** 429–473

Extragalactic Sources with Asymmetric Radio Structure I. Observations of 17 Sources.

D. J. Saikia & P. Shastri *Radio Astronomy Centre, Tata Institute of Fundamental Research, Post Box 1234, Bangalore 560012*

R. P. Sinha *Systems and Applied Science Corporation, 5809 Annapolis Road, Hyattsville, MD 20784, USA*

V. K. Kapahi *Radio Astronomy Centre, Tata Institute of Fundamental Research, Post Box 1234, Bangalore 560012*

G. Swarup *Radio Astronomy Centre, Tata Institute of Fundamental Research, Post Box 8, Udhagamandalam 643001*

Received 1984 April 17; accepted 1984 June 8

Abstract. We present total-intensity and linear-polarization observations with the Very Large Array (VLA) at λ6 and 2 cm of 17 sources, almost all of which were suspected to have extended emission only on one side of the nucleus. Five of them are still one-sided, three appear unresolved, while seven have radio lobes on both sides of the nucleus. The outer components in the double-lobed sources, however, have significantly different surface brightness or are very asymmetrically located with respect to the nucleus.

Key words: extragalactic radio sources—asymmetric radio structure—linear polarization

1. Introduction

The vast majority of powerful, extended, extragalactic radio sources have relatively symmetric double structure with lobes on opposite sides of the optical or nuclear component. However, a small fraction show highly asymmetric structure having a compact component coincident with the optical object and a single extended lobe on one side. These sources are often referred to as D2-type (Miley 1971) double sources. A list of 49 such objects was compiled by Kapahi (1981a; hereafter referred to as K 81). This class of sources does not include the relatively lower luminosity ($P_{178} \lesssim 2 \times 10^{25}$ W Hz^{-1} sr^{-1}) head-tail radio sources generally found in clusters of galaxies, although these often show one-sided radio structure. The large majority of sources in the K 81 list are identified with quasars or blue stellar objects.

Most of the sources listed by K 81 were identified to be one-sided from observations made with either the Westerbork telescope, the NRAO interferometer or the Cambridge 5-km array. To confirm their classification from observations with better resolution and sensitivity, we have observed a large number of sources from the K 81 list with the Very Large Array (VLA). A systematic study of the properties of those which still appear one-sided and possible explanations for their observed morphology

Table 1. The list of sources.

Source	Alternative name	Optical Identification	Optical position						Ref. for position	Redshift	LAS
			Right Ascension			Declination					
			h	m	s	°	′	″			(arcsec)
(1)	(2)	(3)		(4)			(5)		(6)	(7)	(8)
0232 − 042	4C − 04.06	Q	02	32	36.550	−04	15	10.24	C83	1.436	13.2
0309 + 411	NRAO 128	G	03	09	44.81	41	08	48.9	EKM		u
0717 + 170	3C 176	U									105.7
0740 + 380	3C 186	Q	07	40	56.817	38	00	31.04	C83	1.063	100.4
0742 + 376	4C 37.19	BSO	07	42	22.63	37	38	33.6	P		62.3
0821 + 394	4C 39.23A	Q	08	21	37.26	39	26	28.0	PW	1.216	u
0836 + 195	4C 19.31	Q	08	36	15.00	19	32	24.4	MH	1.691	31.2
0932 + 022	4C 02.27	Q	09	32	42.93	02	17	39.6	MH	0.659	44.8
1007 + 417	4C 41.21	Q	10	07	26.13	41	47	24.4	C77	0.611	32.0
1047 + 096	4C 09.37	Q	10	47	48.95	09	41	47.7	MH	0.786	21.1
1055 + 201	4C 20.24	Q	10	55	37.59	20	07	55.3	MH	1.110	21.2
1320 + 299	4C 29.48	BSO	13	20	40.47	29	57	23.2	F79		51.4
1347 + 539	4C 53.28	BSO	13	47	42.66	53	56	08.0	C77		u
1354 + 195	4C 19.44	Q	13	54	42.14	19	33	42.6	MH	0.720	43.7
1419 + 315	4C 31.45	BSO	14	19	19.39	31	32	43.7	F79		13.1
1636 + 473	4C 47.44	Q	16	36	19.17	47	23	28.7	P	0.740	20.0
1729 + 501	4C 50.43	Q	17	29	49.26	50	09	44.3	C77	1.107	20.4

References:

C77	Cohen *et al.* (1977)	MH	Miley & Hartsuijker (1978)
C83	Clements (1983)	P	Present work
EKM	Edwards, Kronberg & Menard (1975)	PW	Potash & Wardle (1979)
F79	Fanti *et al.* (1979b)		

will be discussed in a forthcoming paper. The present paper, which is the first of a series, describes C-array observations of all but three of the sources inthe K81 list with largest angular sizes $\geqslant 5$ arcsec. Two sources were omitted as they had been already studied with high resolution and sensitivity. These are 3C 273 (Perley 1981; Conway *et al.* 1981) and 3C293 (Bridle, Fomalont & Corn well 1981). Bridle, Fomalont & Cornwell have shown that 3C293, a galaxy, is not truly one-sided, but has an extremely asymmetric brightness distribution. The third source not included in the present set of observations is 2041 – 149. To the remaining sample of 14 sources were added the QSOs 0232 – 042 Mley & Hartsuijker 1978; hereinafter MH78) and 1007 + 417 (Kapahi 1981b), and the radio source 3C176 (Joshi 1981), all of which were suspected of having one-sided radio structure and angular sizes $\gtrsim 15$ arcsec.

The final list of 17 sources is presented in Table 1, which is arranged as follows.

Columns 1 and 2: The source name in the coordinate designation, together with an alternative name.

Column 3: Optical identification; Q: quasar, G: galaxy, BSO: blue stellar object, U: unidentified.

Columns 4, 5 and 6: Position of the optical object (epoch 1950) and a reference for the position.

Column 7: Redshift The values are from Hewitt & Burbidge (1980).

Column 8: The largest angular size (LAS) of the radio structure in arcsec; u: unresolved. These values are from the present observations.

The VLA observations possess sufficient resolution to give at least three beamwidths across each source at a wavelength of 6 cm. The sensitivity of the observations permitted fairly good dynamic ranges to be obtained. This was of particular importance to check if a much weaker outer component (hereinafter referred to as OC) was present on the opposite side of the nucleus in addition to the dominant OC.

2. Observations and analyses

The present observations were made on 1980 July 13 and 14, and were among the first made with the completed VLA (Thompson *et al.* 1980). The array was in its C configuration giving a maximum baseline of about 3.4 km. Observations were made at *C* band (λ6 cm) and *U* band (λ2 cm). The observational parameters of the system are shown in Table 2.

For each source, ten-minute observations were made at a number of hour angles (typically three) at each frequency. The *C*- and U-band observations were interleaved

Table 2. Parameters of the observations.

	λ6 cm	λ2 cm
Frequency (MHz)	4885	15035
Bandwidth (MHz)	50	50
System temperature (K)	60	300
Typical HPBW (arcsec)	5	1.5
Typical largest structure 'visible' (arcsec)	125	40
Typical number of antennas available	23	22

Table 3. Flux densities of secondary calibrators.

Source	S_{4885} (Jy)	S_{15035} (Jy)
0316 + 161	2.95	0.71
0316 + 413	57.0	51.3
0711 + 356	1.10	0.47
0839 + 187	0.97	0.53
1404 + 286	2.90	1.41
1739 + 522	0.92	1.05

with the calibrators which were observed about every twenty-five minutes. 3C 286 was the primary calibrator with assumed flux densities of 7.41 Jy at λ6 cm and 3.48 Jy at λ2 cm. Six secondary calibrators were used and their derived flux densities are given in Table 3.

The data were edited and calibrated via standard VLA DEC-10 computer programmes. Initial total power maps were made from the untapered visibility data and CLEANed on the PDP 11/70 system. Although the data were not tapered, attenuation due to bandwidth smearing is not significant for any of the sources in the present sample. The data were further processed using a self-calibration technique to minimize random phase errors (Schwab 1980). At *U* band, the highly extended source 3C176 had insufficient signal-to-noise ratio to permit the use of self-calibration.

The linear polarization characteristics of the sources were also mapped at both frequencies using 3C286 as a calibrator, for which a polarization percentage of 11.3 at position angle (PA) 33° was adopted at *C* band and 11.6 at 33° at *U* band. Correction for the instrumental polarization was made using the extensive observations of the secondary calibrators. To obtain the most accurate possible estimates of polarization percentages on the sources, a further set of total power maps was made using only those baselines for which polarization information was available.

3. Results

The λ6- and 2-cm observational parameters for the individual sources are given in Table 4. For each source at each frequency, the root-mean-square (rms) noises on the final total power and polarization maps are given, along with the half-power beamwidths (HPBW)and orientations of the restored elliptical Gaussian beams used in the CLEAN process.

The self-calibrated total power maps of the sources are shown in Figs 1–14. The figures have not been corrected for attenuation by the primary polar diagram. The maps for the sources 0309 + 411, 0821 + 394 and 1347 + 539 are not included as they showed only a point source coincident with the optical object. For the same reason the λ2-cm map of 1354 + 195 is not reproduced. Also shown in the figures are the λ6-cm maps of linear polarization. Vectors representing polarized intensity are superposed on total power contours containing information only from those baselines which possessed polarization information (see Section 2). When considering the total power structures of the sources at λ6 cm, Figs l(a) to 14(a) should be used exclusively. Due to insufficient signal-to-noise ratio at λ2 cm all components showing linear polarization at

Table 4. The observational parameters for individual sources.

Source	λ6 cm					λ2 cm				
	σ_{tp} (mJy/ beam)	σ_{pol} (mJy/ beam)	HPBW major (″)	HPBW minor (″)	PA (°)	σ_{tp} (mJy/ beam)	σ_{pol} (mJy/ beam)	HPBW major (″)	HPBW minor (″)	PA (°)
0232 − 042	2.3	0.3	5.24	3.94	171	1.7	3.4	1.77	1.42	19
0309 + 411	1.0	0.2	4.96	3.88	72	2.3	2.2	1.65	1.27	71
0717 + 170	0.4	0.6	5.36	4.28	55	1.7	2.4	1.68	1.51	70
0740 + 380	0.9	0.2	5.00	3.96	95	1.2	2.4	1.70	1.33	92
0742 + 376	0.8	0.2	5.21	3.98	95	1.4	2.1	1.77	1.36	95
0821 + 394	6.4	0.4	4.59	4.32	82	8.1	2.5	1.96	1.40	100
0836 + 195	0.3	0.2	4.72	4.37	98	1.1	1.6	1.53	1.48	75
0932 + 022	0.6	0.2	6.15	4.75	116	1.2	2.6	1.65	1.34	30
1007 + 417	0.9	0.2	4.84	3.95	104	1.4	2.3	1.57	1.29	85
1047 + 096	0.3	0.3	4.67	4.14	175	1.1	1.9	1.53	1.37	17
1055 + 201	3.1	0.3	4.37	4.01	16	6.6	2.5	1.48	1.31	176
1320 + 299	1.2	0.2	4.74	4.09	58	1.7	1.9	1.54	1.34	51
1347 + 539	1.8	0.2	4.88	3.84	106	2.6	3.9	1.66	1.27	125
1354 + 195	2.7	0.6	4.40	4.24	175	4.5	3.7	1.56	1.41	2
1419 + 315	0.8	0.2	4.42	3.91	63	1.1	1.9	1.51	1.35	50
1636 + 473	0.8	0.2	4.32	3.77	97	2.9	1.6	1.61	1.22	70
1729 + 501	1.0	0.6	4.44	3.87	71	1.3	1.8	1.43	1.23	68

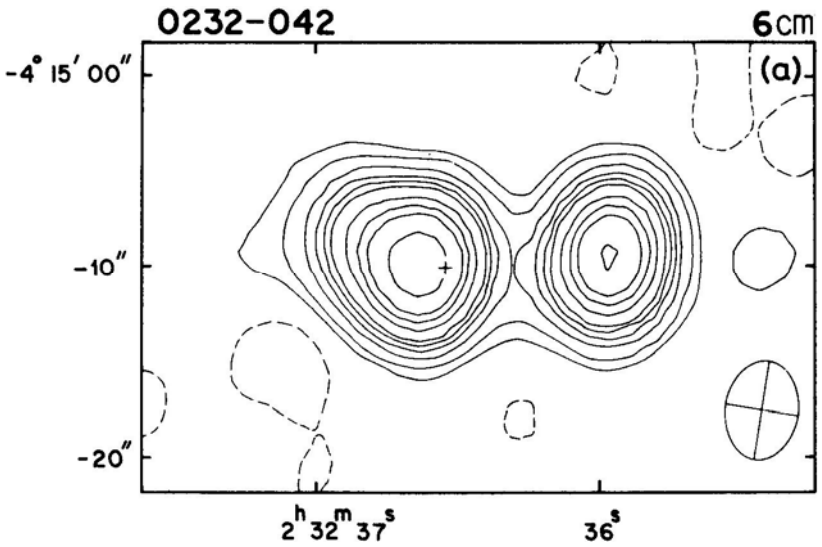


Figure 1.* (a) 0232–042. Contours: 230 × (–0.04, –0.02, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40,0.50, 0.75).

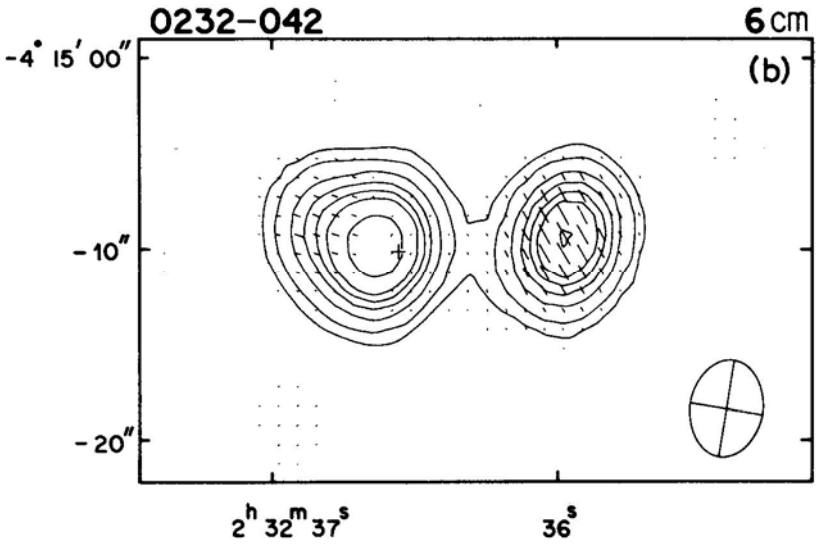


Figure 1. (b) 0232–042. Contours: 240 × (–0.10, –0.05, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.75).

* In Figs 114, the crosses mark the positions of the optical objects while the ellipses indicate the sizes of the CLEAN beams listed in Table 4. Contours are listed as fractions of the peak flux density on the map in units of mJy/beam. The figures have not been corrected for attenuation by the primary polar diagram. The flux density values corrected for attenuation are listed in Tables 5a and 5b. The λ6–cm total-power maps using data from all baselines are shown in Figs l(a) to 14(a), while total-power contours containing information only from those baselines which possessed polarization information are shown in Figs l(b) to 14(b). In these figures, vectors representing polarized intensity are superposed on the total-power contours. The λ2 cm maps are shown in Figs l(c) to 14(c).

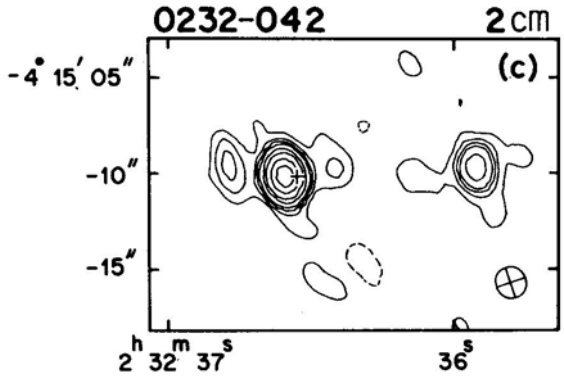


Figure 1. (c) 0232–042. Contours: 121 × (–0.06, –0.03, 0.03, 0.06, 0.09, 0.12, 0.20, 0.30, 0.50, 0.75).

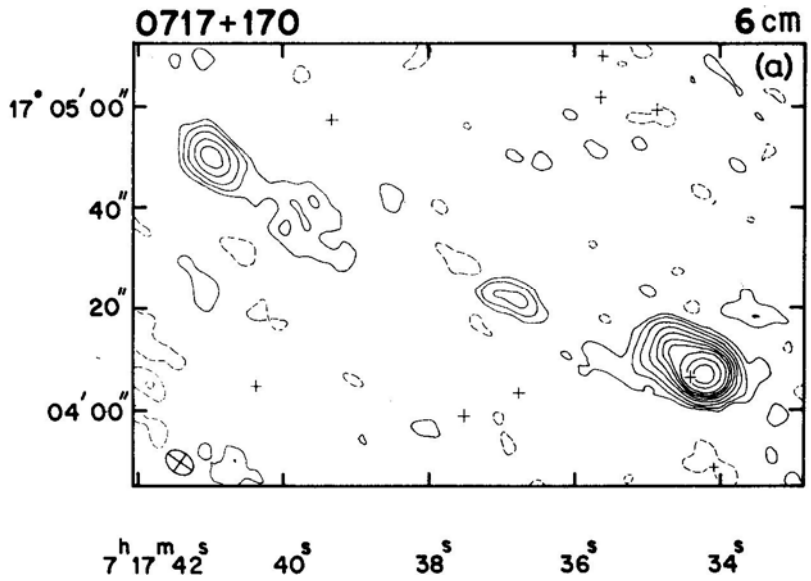


Figure 2. (a) 0717 +170. Contours: 101 × (–0.015, –0.0075, 0.0075, 0.015, 0.03, 0.06, 0.10, 0.15, 0.20, 0.30, 0.50, 0.75).

a 2σ or higher level were either unresolved, or only marginaly resolved. The $\lambda 2$ cm maps of linear polarization are therefore not included and the relevant information can be found in Table 5b.

The detailed information on the individual objects at $\lambda 6$ cm is to be found in Table 5a, with the corresponding data for $\lambda 2$ cm being contained in Table 5b. Table 5a is arranged as follows.

Column 1: The source name.

Column 2: Sub-component identification.

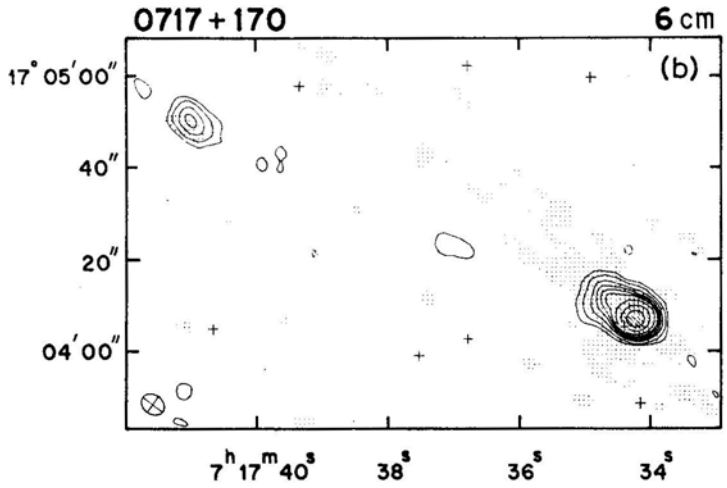


Figure 2. (b) 0717 +170, entire source. Contours: 94 × (–0.04, –0.02, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.50, 0.75).

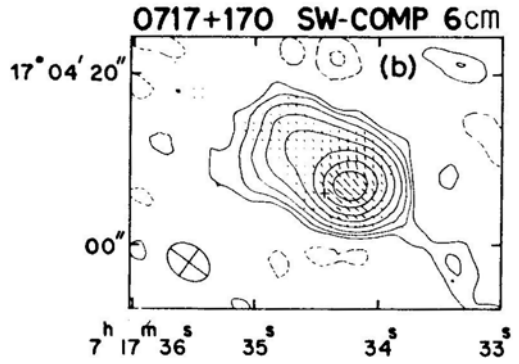


Figure 2. (b) 0717 +170, southwestern component. Contours: 94 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.16, 0.30, 0.50, 0.75).

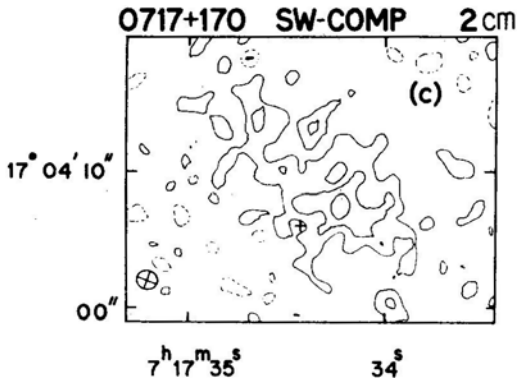


Figure 2. (c) 0717 +170. Contours: 12 × (–0.50, –0.25, 0.25, 0.50, 0.75).

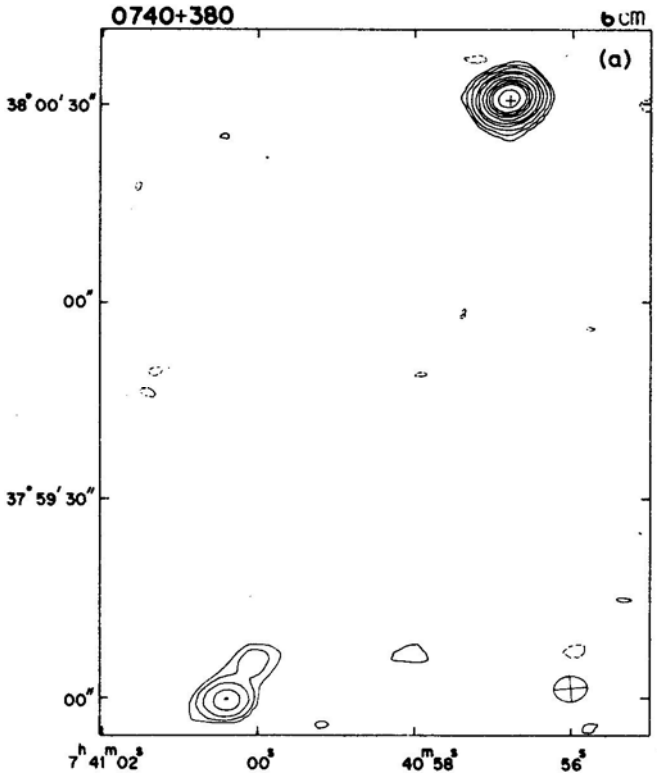


Figure 3. (a)0740 + 380. Contours: 210 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.20, 0.30, 0.40, 0.50, 0.75).

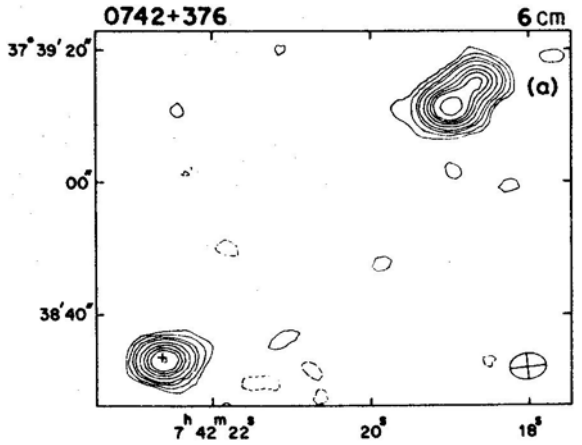


Figure 4. (a) 0742 + 376. Contours: 76 × (–0.04, –0.02, 0.02, 0.04, 0.08, 0.12, 0.20, 0.30, 0.40, 0.50,0.75).

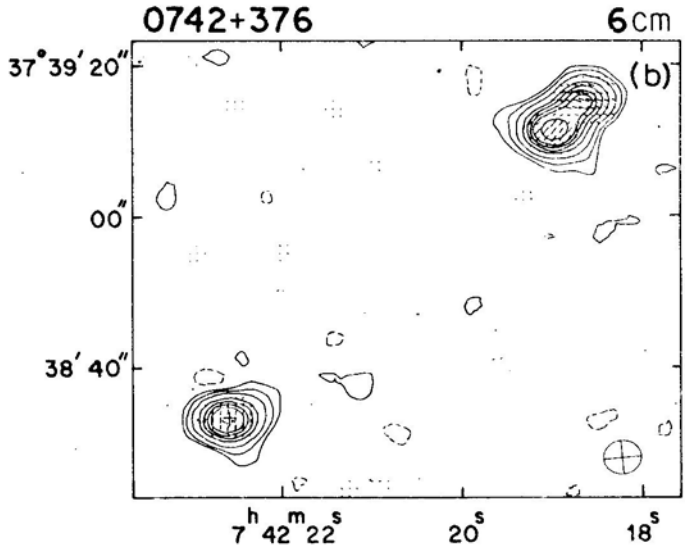


Figure 4. (b) 0742 + 376. Contours: 72 × (0.05, 0.025, 0.025, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.75).

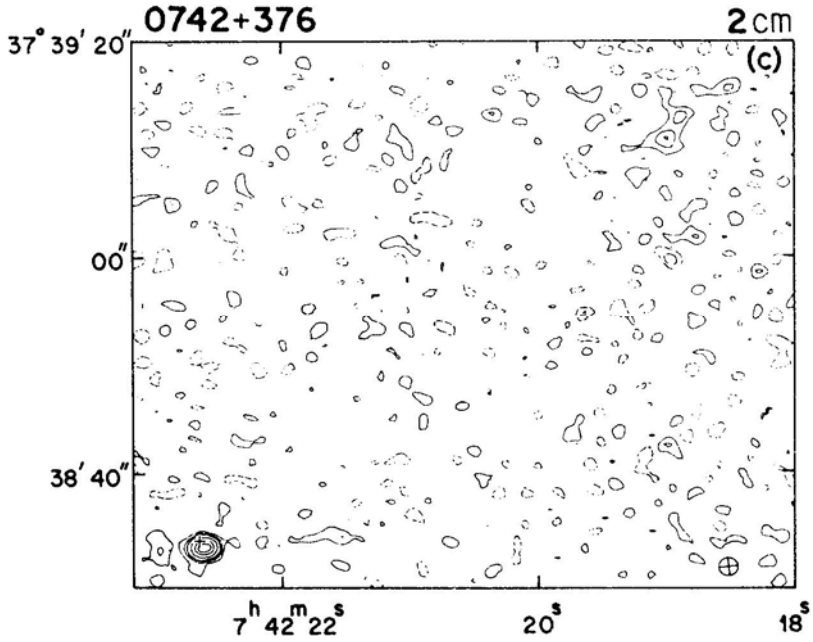


Figure 4. (c) 0742 + 376. Contours: 47 × (–0.10, –0.05, 0.05, 0.10, 0.15, 0.30, 0.50, 0.75).

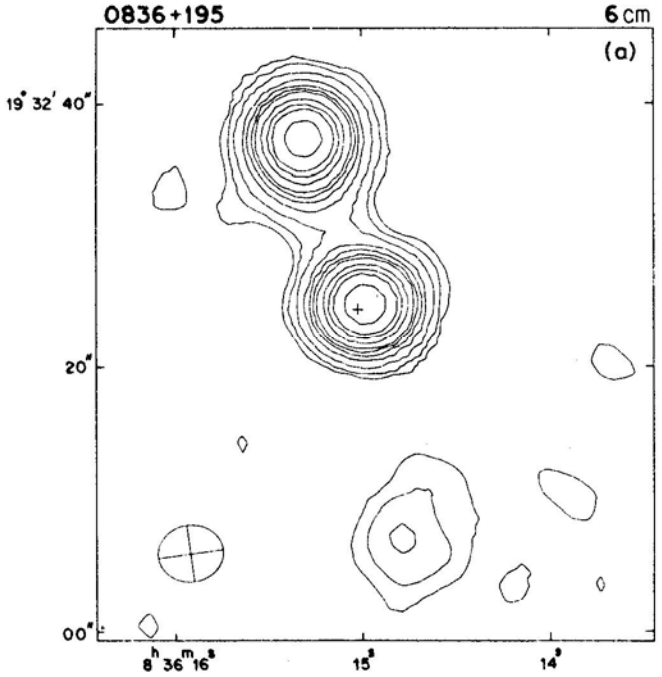


Figure 5. (a) 0836 + 195. Contours: 71 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

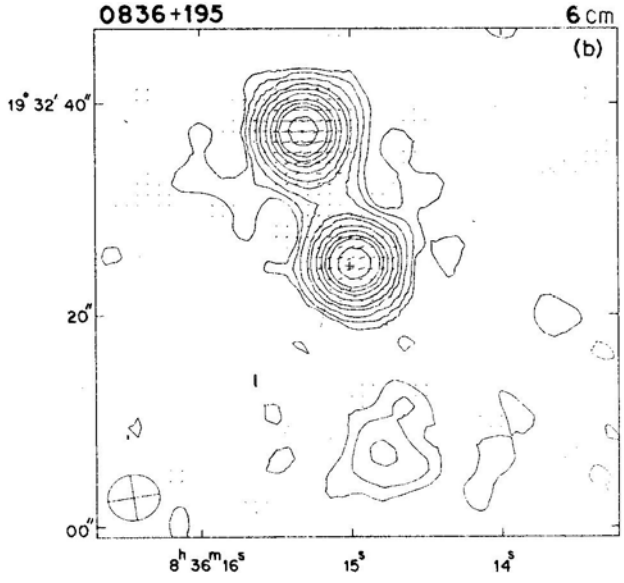


Figure 5. (b) 0836 +195. Contours: **72 × (–**0.02, –0.01, 0.01**,** 0.02, 0.04, 0.08, 0.12, 0.20, 0.30, 0.40, 0.50, 0.75).

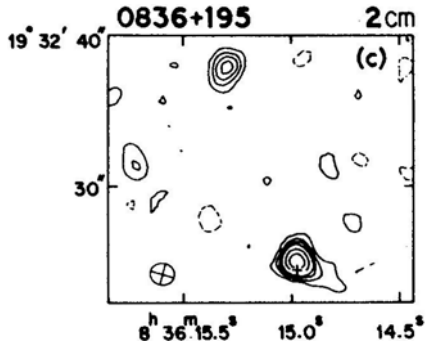


Figure 5. (c) 0836+195. Contours: 28 × (–0.15, –0.075, 0.075, 0.15, 0.225, 0.30, 0.50, 0.75).

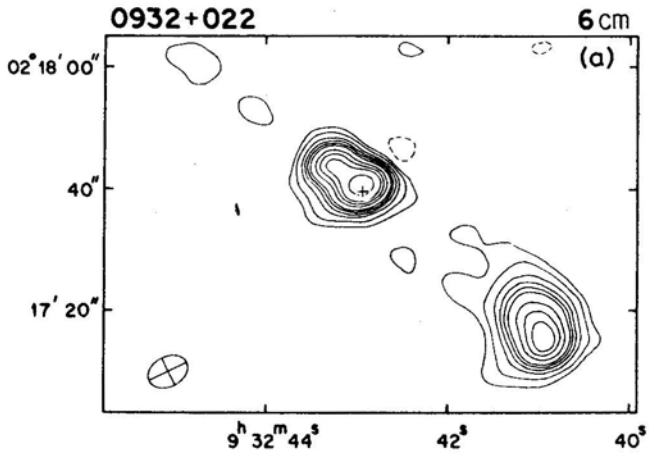


Figure 6. (a) 0932 + 022. Contours: 82 × (–0.04, –0.02, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

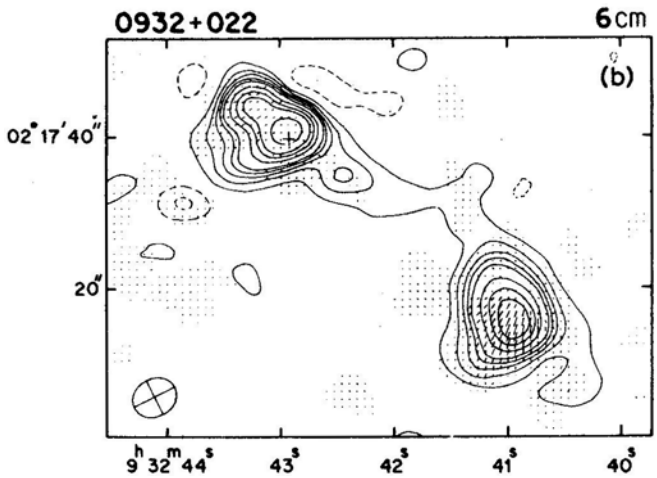


Figure 6. (b) 0932 + 022. Contours: 72 × (– 0.08, –0.04, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75)

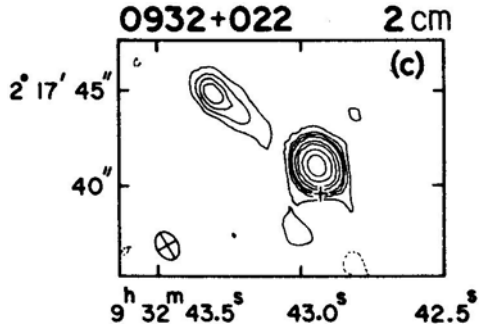


Figure 6. (c) 0932 + 022. Contours: 72 × (–0.06, –0.03, 0.03, 0.06, 0.09, 0.12, 0.20, 0.30, 0.50, 0.75).

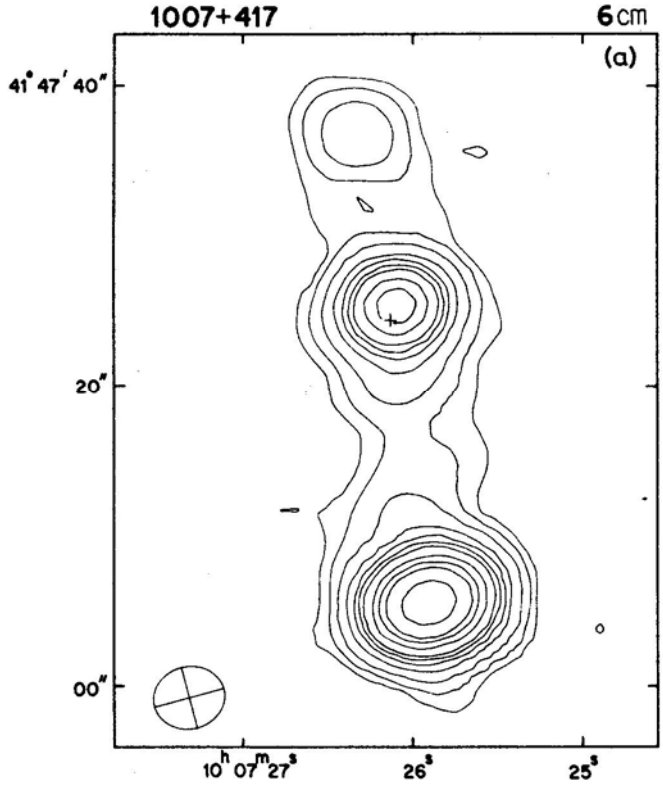


Figure 7. (a) 1007 + 417. Contours: 304 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

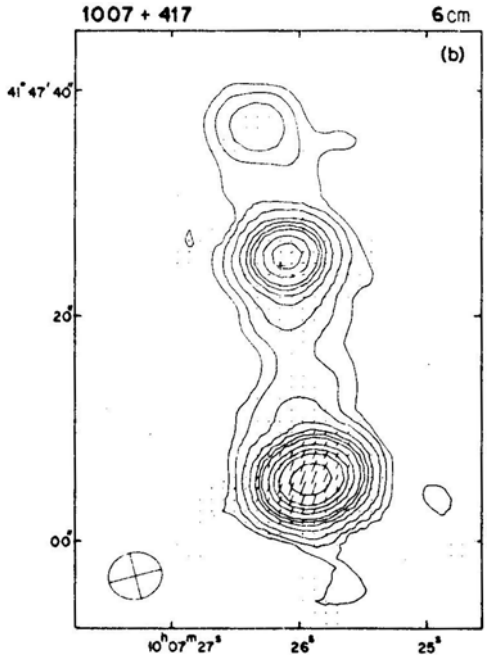


Figure 7. (b) 1007 + 417. Contours: 302 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

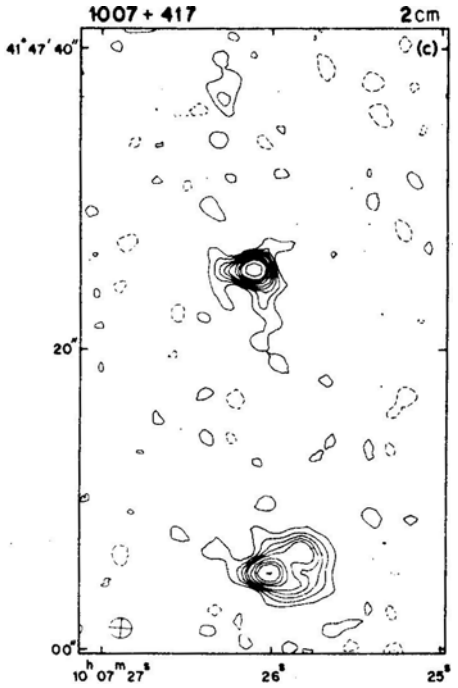


Figure 7. (c) 1007 + 417. Contours: 67 × (–0.08, –0.04, 0.04 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

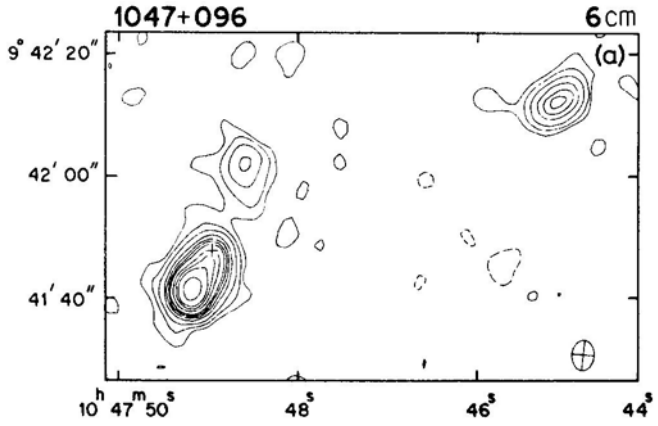


Figure 8. (a) 1047 + 096. Contours: 59 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

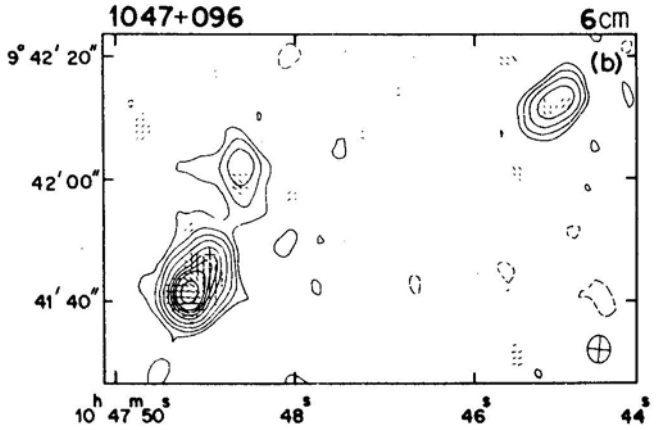


Figure 8. (b) 1047 + 096. Contours: 58 × (–0.025, –0.0125, 0.0125, 0.025, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.75).

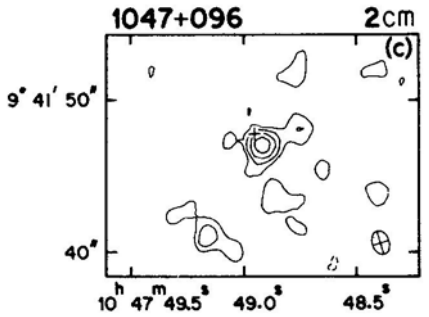


Figure 8. (c) 1047 + 096. Contours: 16 × (–0.30, –0.15, 0.15, 0.30, 0.50, 0.75).

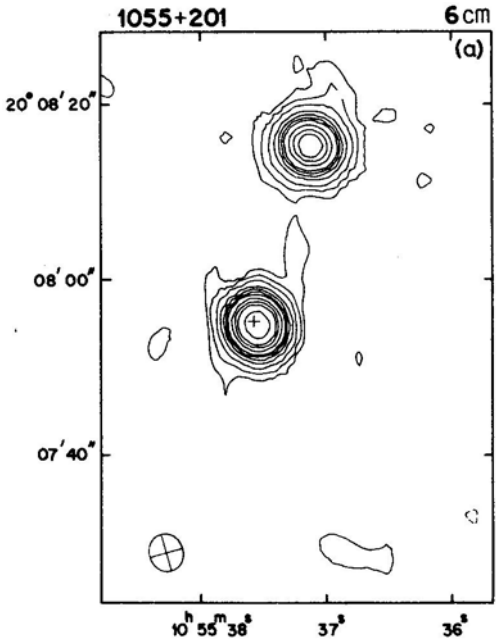


Figure 9. (a) 1055 + 201. Contours: 792 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

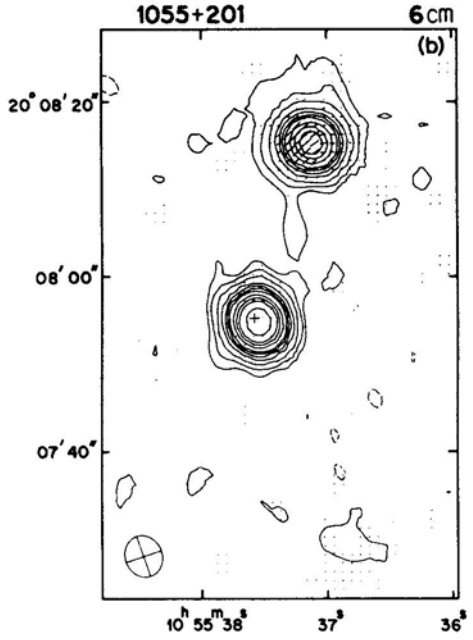


Figure 9. (b) 1055 + 201. Contours: 787 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

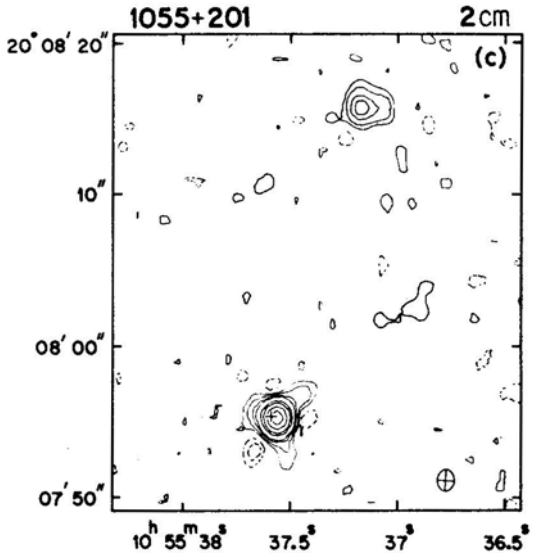


Figure 9. (c) 1055 + 201. Contours: 819 × (–0.03, –0.015, 0.015, 0.03, 0.06, 0.09, 0.20, 0.30, 0.50, 0.75).

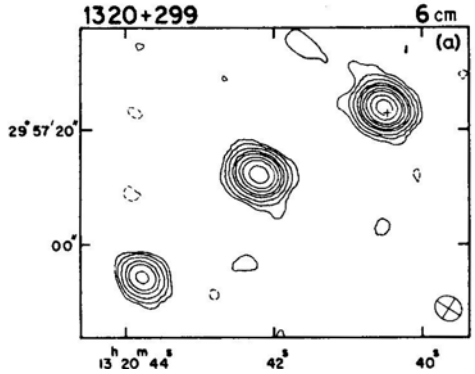


Figure 10. (a) 1320+299. Contours: 277 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.20, 0.30, 0.50, 0.75).

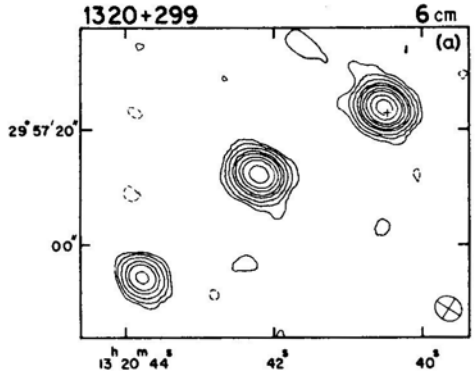


Figure 10. (b) 1320 + 299. Contours: 277 x (—0.02, —0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

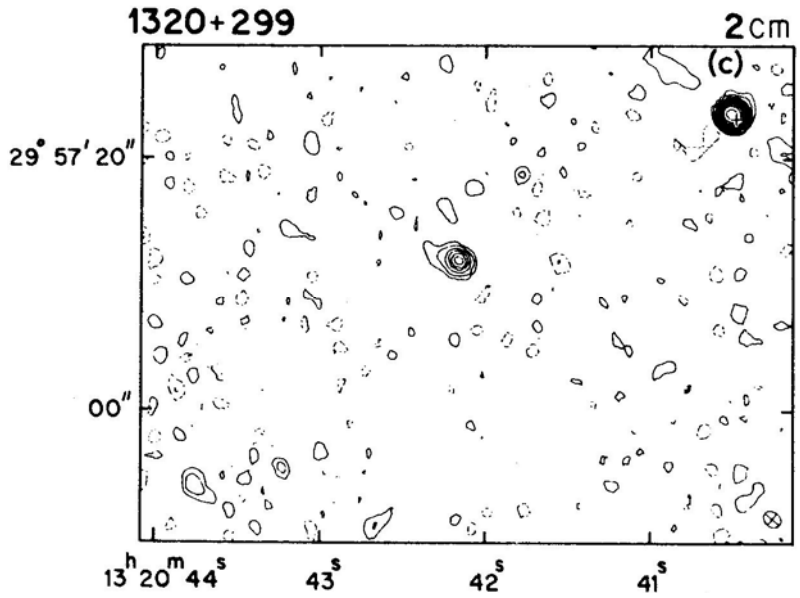


Figure 10. (c) 1320 + 299. Contours: 156 × (–0.04, –0.02, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

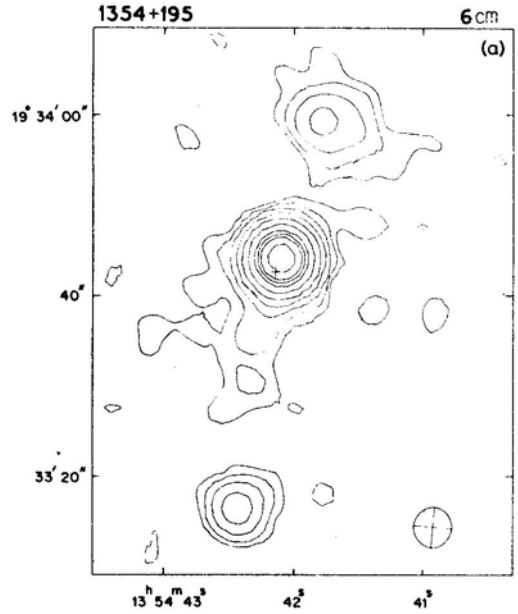


Figure 11. (a) 1354 + 195. Contours: 1229 × (–0.01, – 0.005, 0.005, 0.01, 0.02, 0.04, 0.08, 0.16, 0.30, 0.40, 0.50, 0.75).

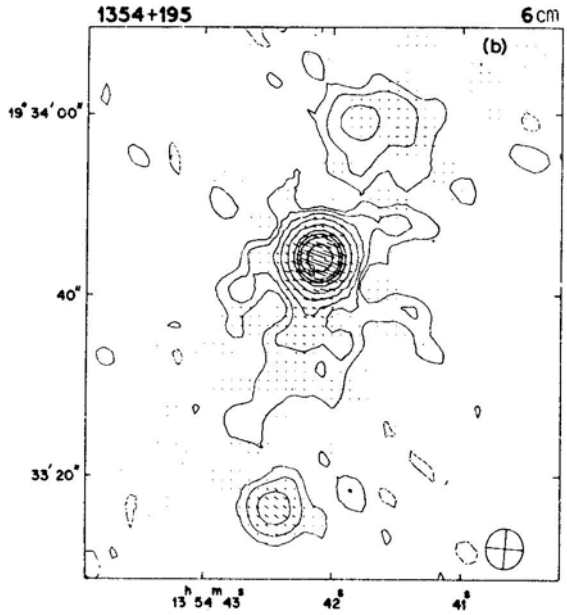


Figure 11. (b) 1354+ 195. Contours: 1202 × (–0.02, –0.01, 0.01, 0.02, 0.04, 0.08, 0.16, 0.30, 0.40, 0.50, 0.75).

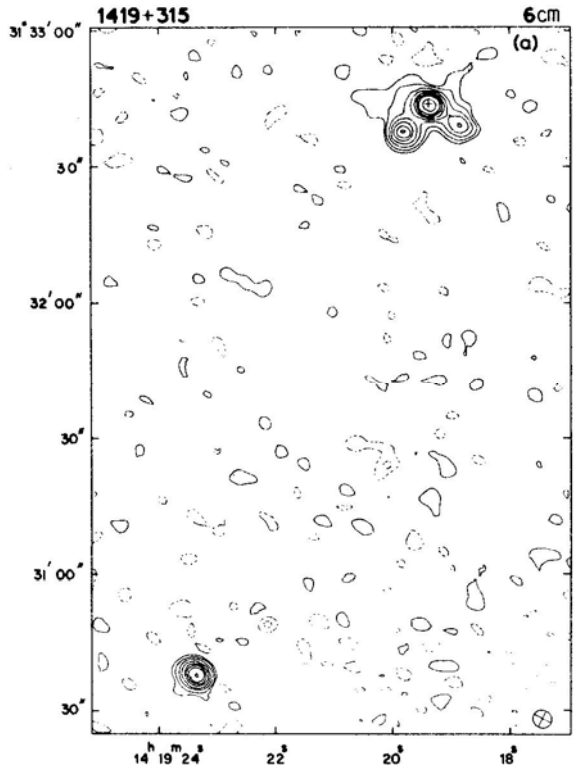


Figure 12. (a) 1419 + 315. Contours: 55 × (– 0.06, – 0.03, 0.03, 0.06, 0.12, 0.20, 0.30, 0.40, 0.50, 0.75).

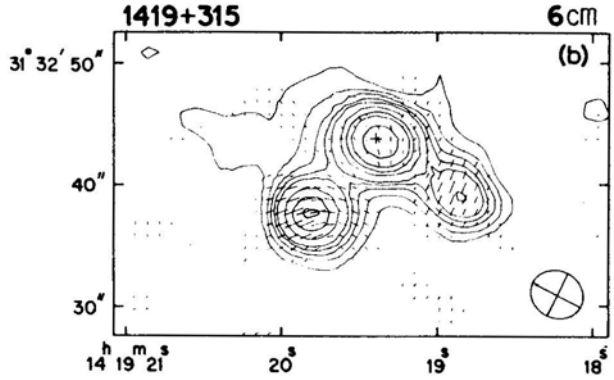


Figure 12. (b) 1419 + 315. Contours: 54 × (– 0.08, – 0.04, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

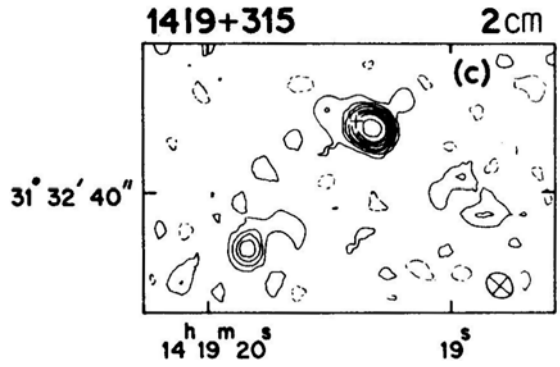


Figure 12. (c) 1419 + 315. Contours: 36 × (– 0.10, – 0.05, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.75).

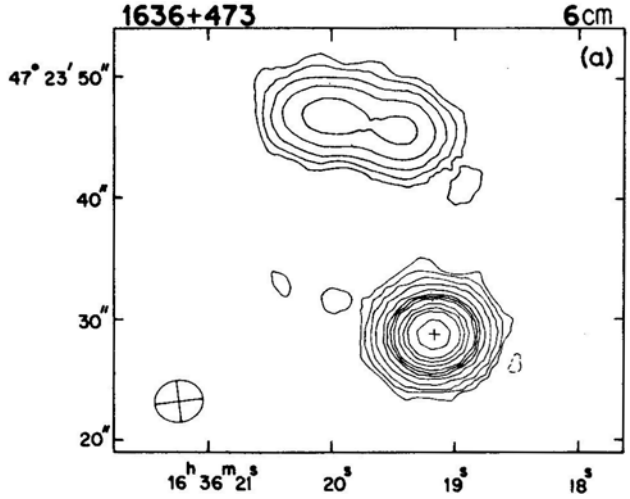


Figure 13. (a) 1636 + 473. Contours: 492 × (–0.01, –0.005, 0.005, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

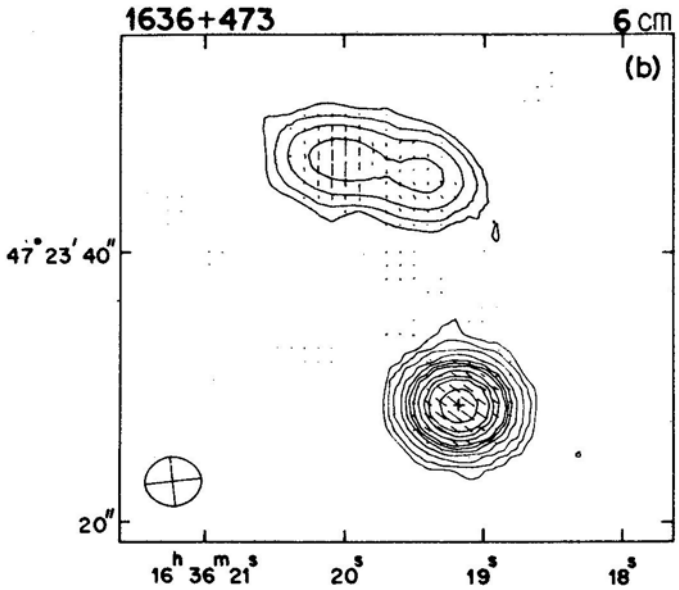


Figure 13. (b) 1636 + 473. Contours: 494 × (– 0.02, – 0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16 0.20, 0.30, 0.40, 0.50, 0.75).

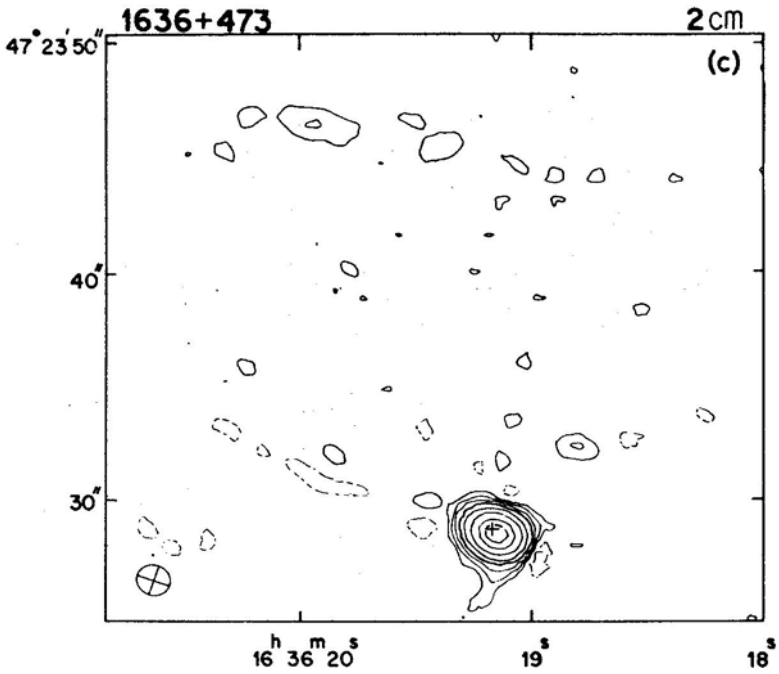


Figure 13. (c) 1636 + 473. Contours: 885 x (– 0.03, – 0.015, –0.0075, 0.0075, 0.015, 0.03, 0.06, 0.12, 0.25, 0.50, 0.75).

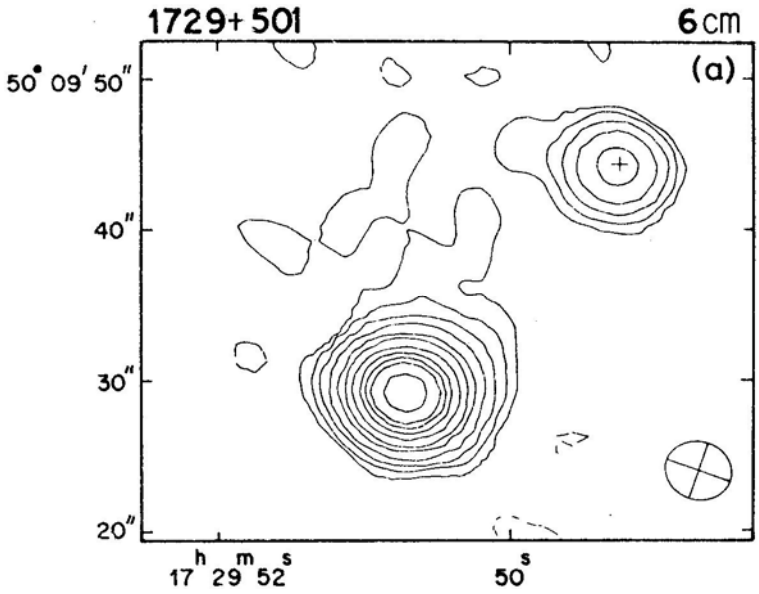


Figure 14. (a) 1729 + 501. Contours: 355 × (– 0.01, –0.005, 0.005, 0.01, 0.02, 0.04, 0.08, 0.12, 0.20, 0.30, 0.40, 0.50, 0.75).

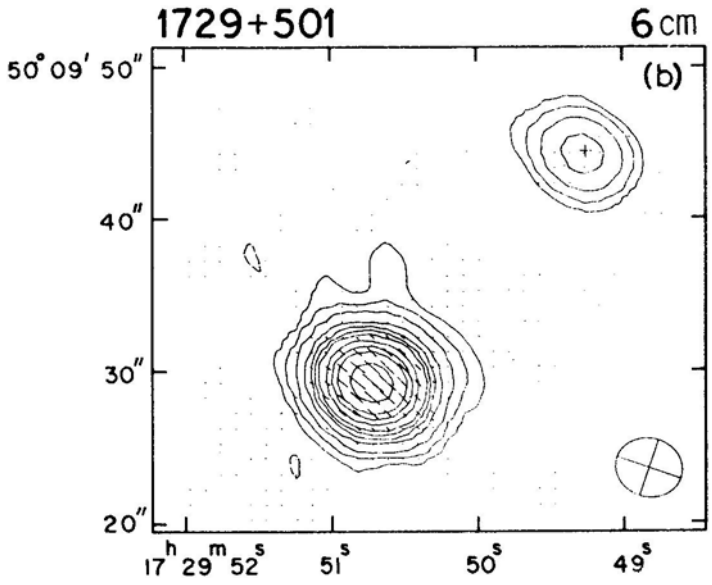


Figure 14. (b) 1729 + 501. Contours: 353 × (–0.02, – 0.01, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.50, 0.75).

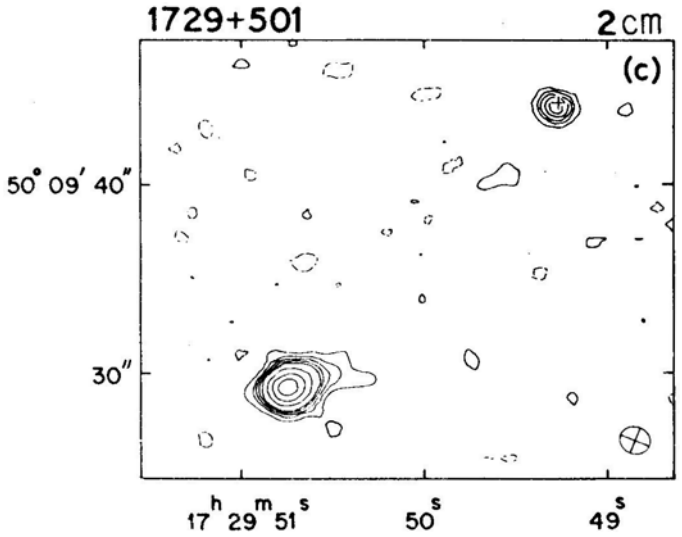


Figure 14. (c) 1729 + 501. Contours: 90 × (– 0.06, – 0.03, 0.03, 0.06, 0.09, 0.12, 0.16, 0.30, 0.50, 0.75).

Columns 3 and 4: The right ascension and declination (epoch 1950) of the sub-component. This gives the position of the best-fit, two-dimensional Gaussian, unless enclosed in parentheses, when it is the position of the pixel of maximum brightness. The pixels are separated by 1 arcsec at λ6 cm and by 0.4 arcsec at λ 2 cm.

Column 5: The deconvolved half-power width and orientation of the best-fit Gaussian to the sub-component. If the component width was less than half the HPBW in all directions it was considered to be unresolved and is marked u.

Column 6: The peak flux density (S_p) and the integrated flux density (S_1) of the sub-component in mJy. S_p is the height of the best-fit, two-dimensional Gaussian, unless enclosed in parentheses, when it is the intensity at the pixel of maximum brightness for weak and/or complex components. For S_1, the integrations were made over rectangular areas containing a complete component, or a combination of components, as indicated. All flux density values have been corrected for attenuation by the primary polar diagram.

Columns 7 to 9: These contain the information on the linear polarization characteristics of the component. In columns 7 and 8 the right ascension and declination of the pixel of maximum polarized intensity are given if significantly shifted from the position of the peak of the total power component. In such cases, the polarization percentage and PA (column 9) are also tabulated for the total power peak.

In Table 5b columns 1 to 6 are identical to that of Table 5a. Since at λ2 cm the pixel of maximum polarized intensity is coincident with the total power peak in all cases, the positions of peak polarization have not been included in the table. The polarization percentage and PA are listed in column 7.

For both λ6 and 2 cm, a correction for the effects of noise on the Stokes parameters Q and U was applied to the measured polarized intensity before computing the polarization percentage to give the most likely value of polarized intensity (Wardle &

Table 5a. λ6 cm data for the individual sources.

Source	Component	Right Ascension			Declination			Component size major	minor	PA	S_p	S_t	Right Ascension (polarization)			Declination (polarization)			Polarization per cent	PA (deg)
		h	m	s	°	′	″	(″)	(″)	(°)	(mJy)	(mJy)	h	m	s	°	′	″		
(1)	(2)		(3)			(4)			(5)		(6)			(7)			(8)		(9)	
0232−042	W	02	32	35.97	−04	15	09.4	u			184±11	184							10.8	30
	C			36.65			09.7	3.6	<2	70	226±17	285							⩽0.13	
	E												02	32	36.84	−04	15	08.7	⩾6.8	71
0309+411	C	03	09	44.82	41	08	48.9	u			337±17	337							2.2	30
0717+170	W	07	17	34.24	17	04	07.3	4.8	4.0	89	99±9	197							17.0	45
	C?			36.87			22.1	8.6	<2.0	65	4±0.5	9							⩽16.6	
	E			40.98			50.1	5.5	3.2	37	15±1	30							10.2	140
													07	17	41.09	17	04	49.3	19.7	132
0740+380	C	07	40	56.80	38	00	30.8	2.0	<2.0	142	213±11	229							0.6	24
	E1	(	41	00.03	37	59	05.4)				(8)	39								
	E2		41	00.39	37	58	59.8	2.6	<2.0	127	28±3								⩽0.8	
0742+376	W1	(07	42	18.69	37	39	14.7)				(44)	126	07	42	18.64	37	39	15.2	10.6	75
	W2	(		18.98		39	11.2)				(78)								3.9	143
	C			22.60		38	33.2	u			60±3	60							4.6	175
0821+394	C	08	21	37.34	39	26	28.1	u			1036±53	1247							1.4	120
0836+195	SW	08	36	14.75	19	32	07.1	6.3	3.9	161	3±0.4	7							⩽5.0	
	C			14.98			24.9	u			74±4	77							3.6	109
	NE			15.31			37.3	2.3	⩽2.2	2	68±4	88							7.7	95
0932+022	W	09	32	40.95	02	17	16.1	11.4	5.0	24	(51)	137	09	32	40.95	02	17	15.3	15.4	159
	C			43.00			41.2	6.3	2.4	53	(82)	143			42.93			40.6	1.1	125
	E	(		43.26			44.6)				(45)				43.26			44.1	5.7	115
1007+417	SW	10	07	25.91	41	47	05.6	3.2	2.3	111	302±16	454							8.8	160
	C			26.10			25.3	u			149±9	189							3.2	108
	NE			26.33			36.6	3.9	2.2	38	21±2	38							4.2	106

1047+096	W	10	47	45.11	09	42	12.3	5.2	<2.0	122	11±0.6	21											3.8	157
													10	47	45.23		09	42	11.7				8.6	16
	NW	10	47	48.59	09	42	1.6	4.9	3.6	1	5±0.5	13											9.0	48
													10	47	48.61		09	41	59.2				25.2	60
	C	(		49.02		41	44.7)				(29)				48.97				44.4				7.3	1
	SE			49.14		41	42.1	6.1	2.3	147	56±4	117			49.22				40.7				4.4	85
1055+201	NW	10	55	37.12	20	08	15.3	2.3	<2.0	94	510±27	690											3.2	125
													10	55	37.23		20	08	15.3				6.8	135
	C			37.55		07	54.9	u			846±43	959			36.95			08	15.3				3.9	90
															37.52			07	57.3				0.1	93
																							1.0	91
1320+299	C	13	20	40.52	29	57	23.8	u			277±14	298											4.7	135
	E1			42.21		57	12.3	2.1	<2.0	74	195±10	229											9.5	126
	E2			43.77		56	54.3	u			69±4	78											9.6	9
1347+539	C	13	47	42.61	53	56	08.5	u			836±43	891											1.8	13
1354+195	NW	13	54	41.73	19	33	59.2	5.3	3.7	65	60±5												3.4	178
													13	54	41.41		19	33	58.6				14.8	96
	C			42.10			44.1	u			1292±65	1761											4.8	70
	SE			42.43			16.6	2.0	≤2.0	119	73±5												14.3	50
1419+315	W	(14	19	18.82	31	32	38.8)				(17)												10.9	157
	C			19.36			43.5	3.5	2.3	86	54±5	140											1.6	13
	E	(		19.84			37.8)				(28)												11.6	108
	SE	14	19	23.37	31	30	38.1	u			54±3	54											≤0.9	
1636+473	C	16	36	19.17	47	23	28.8	u			500±25	500											1.6	53
	NE1	(		19.45			46.0)				48	150											1.5	164
	NE2	(		19.99			47.0)																11.8	179
1729+501	C	17	29	49.27	50	09	44.2	u			39±2	43											5.5	50
	E			50.71			29.3	1.6	<1.6	118	352±18	414											18.2	48

Table 5b. $\lambda 2$ cm data for the individual sources.

Source	Component	Right Ascension			Declination			Component size major	minor	PA	S_p	S_I	Polarization per cent	PA
		h	m	s	°	′	″	(″)	(″)	(deg)	(mJy)	(mJy)		(deg)
(1)	(2)		(3)			(4)			(5)		(6)		(7)	
0232−042	W	02	32	35.92	−04	15	09.7	1.2	0.7	166	33±3	75	⩽11.2	
	C			36.59			10.1				120±7	159	⩽ 2.9	
	E	(		36.80			09.4)				(12)		⩽23.1	
0309+411	C	03	09	44.78	41	08	48.8	u			465±39	465	2.1	39
0717+170	W	(07	17	34.24	17	04	- 07.5)				(12)	158	⩽20.9	
0740+380	C	07	40	56.78	38	00	31.0	1.9	0.75	136	30±2	78	⩽ 8.5	
	E2	(	41	00.42	37	58	59.9)				(⩽3.5)			
0742+376	W1	(07	42	18.46	37	39	15.8)				(8)	103	⩽38.2	
	W2	(		19.00		39	11.0)				(10)		⩽27.6	
	C			22.60		38	33.2	u			49±3	52	⩽ 6.4	
0821+394	C	08	21	37.31	39	26	28.2	u			992±54	1270	0.6	116
0836+195	C	08	36	14.97	19	32	25.0	0.7	⩽0.7	37	28±3	38	⩽ 5.7	
	NE			15.30			37.8	1.3	⩽0.7	75	11±1	18	⩽21.9	
0932+022	C	09	32	42.94	02	17	41.1	u			76±4	93	⩽ 3.9	
	E	(		43.30			44.8)				(11)	25	⩽22.2	

1007+417	SW	10	07	26.00	41	47	05.2					(52)	183	⩽ 9.9		
	C			26.11			25.3	u				70±5	111	⩽ 4.8		
	NE			26.28			37.0					7±1	20			
1047+096	C	10	47	48.92	09	41	47.0	1.2	0.7	122		16±2	112	⩽27.5		
	SE			49.15			41.1	3.8	0.8	51		7±2		⩽22.4		
1055+201	NW	10	55	37.16	20	08	15.7	1.7	0.9	87		103±11	277	⩽ 2.8		175
	C			37.56		07	55.3	u				832±53	889	5.4		
1320+299	C	13	20	40.51	29	57	23.6	u				159±9	159	4.3		149
	E1			42.17		57	12.0	0.75	<0.70	70		39±4	49	16.9		133
	E2			43.75		56	54.1	1.6	<0.70	36		14±3	29	⩽24.7		
1347+539	C	13	47	42.58	53	56	08.4	u				611±32	666	4.5		50
1354+195	C	13	54	42.07	19	33	43.9	u				944±48	1048	3.4		55
1419+315	W	(14	19	18.86	31	32	38.8)					(4)				
	C			19.33			43.4	u				38±3	89	⩽ 7.4		
	E			19.83			37.1	1.35	<0.7	155		9±2				
1636+473	C	16	36	19.15	47	23	28.5	u				914±47		1.3		53
	NE1			19.34			45.6					(13)	991			
	NE2			19.89			46.4					(13)				
1729+501	C	17	29	49.28	50	09	44.1	u				18±2	23	⩽10.7		
	E			50.74			29.3	1.2	0.6	120		93±6	142	10.2		43

Kronberg 1974). When upper limits on the polarization percentage are given these represent a lσ polarized intensity.

4. Spectra

The integrated spectra of the 17 sources and, where possible, the spectra of individual components are shown in Fig. 15. All the total flux densities are on the scale of Baars *et al.* (1977) and have been compiled largely from the catalogues listed by Kühr *et al.* (1979, 1981). Estimates of the total flux density from aperture synthesis observations have sometimes been included, usually either to extend the spectrum to higher frequencies or to provide information at frequencies for which single-dish measurements are not available. These flux densities have been estimated either from the correlated flux density at the shortest spacings or from a sum of the component flux densities. Wherever possible, the component flux densities have also been converted to the scale of Baars *et al.* (1977). We have assumed an error of 10 per cent in the flux density, unless the quoted error is larger than this value.

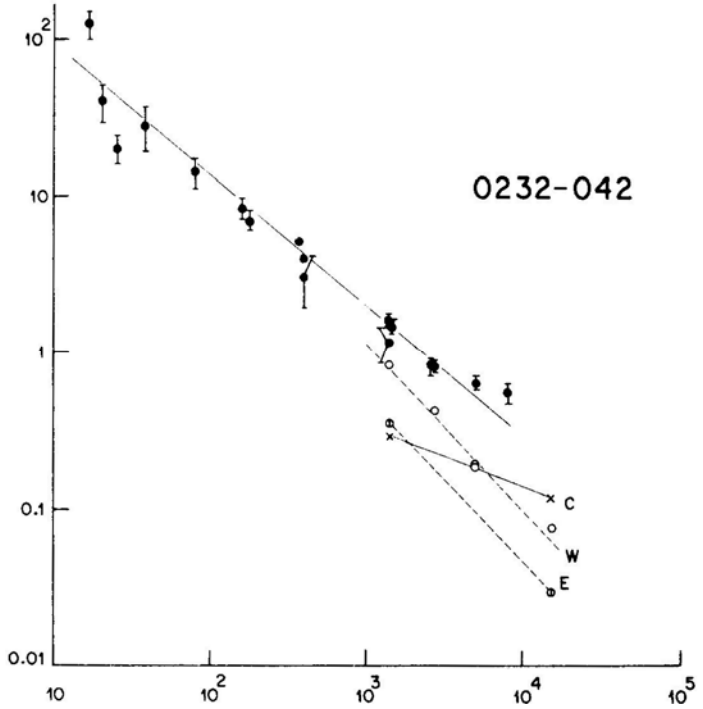
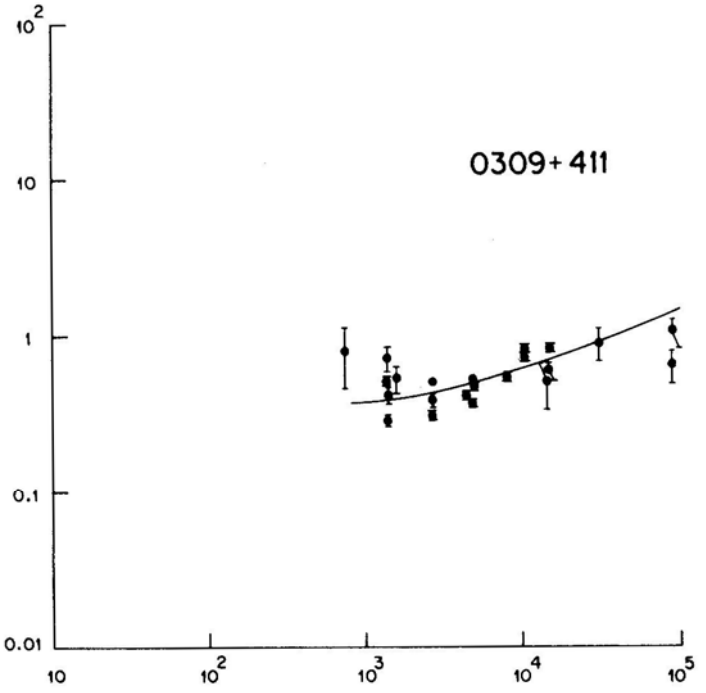


Figure 15. The spectra of the entire source and the individual radio components. The frequency values are in MHz while the flux densities are in Jy. The filled circles represent the total flux density measurements, and the solid curves, the best fits to these points. The spectra of the individual components are shown by broken lines. The different components are represented by the following symbols. × : core (C); O : western (W), south-western (SW) and north-western (NW); ⊕ : eastern (E), north-eastern (NE) and the northern triple source in 1419 + 315; ⊗: the western component of 1047 + 096; ϴ : the triple source of 1047 + 096, the southern component of 1419 + 315 and the far-eastern component (E2) of 1320 + 299.



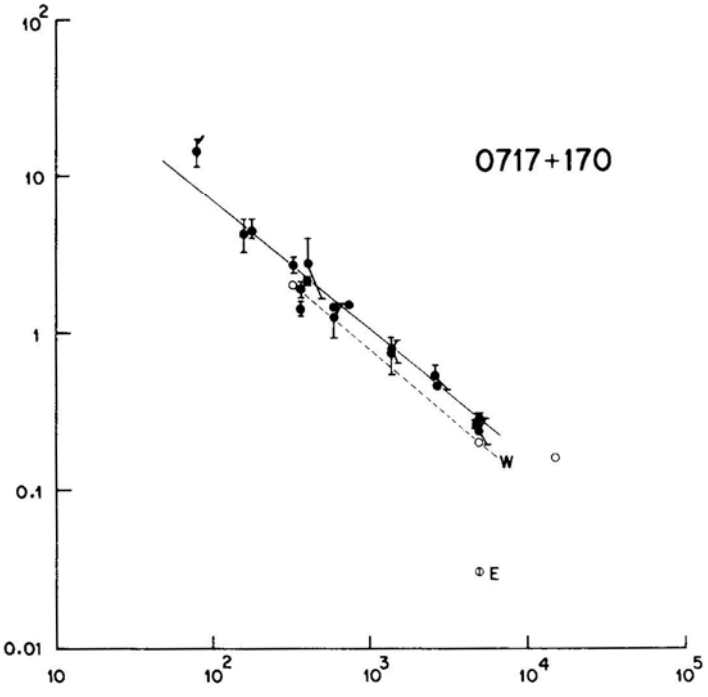


Figure 15. Continued

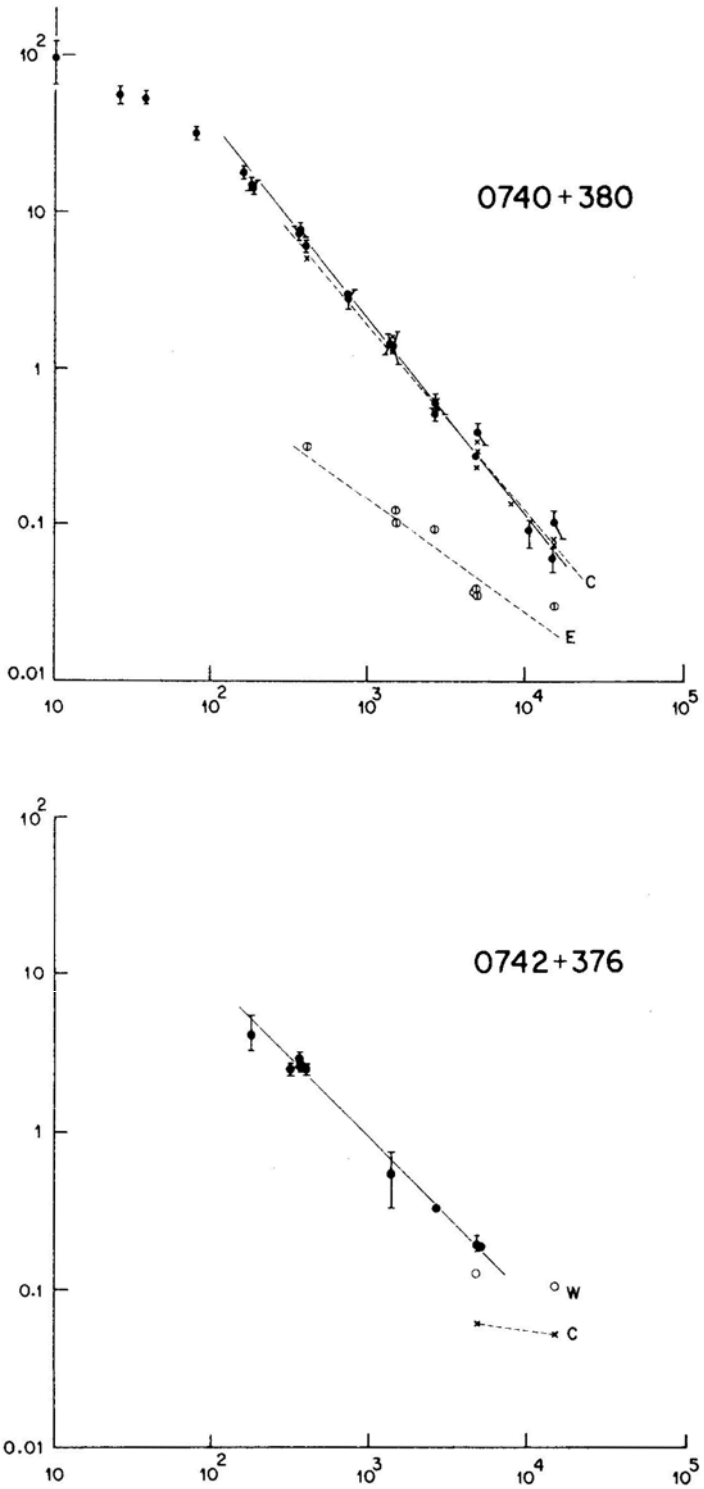
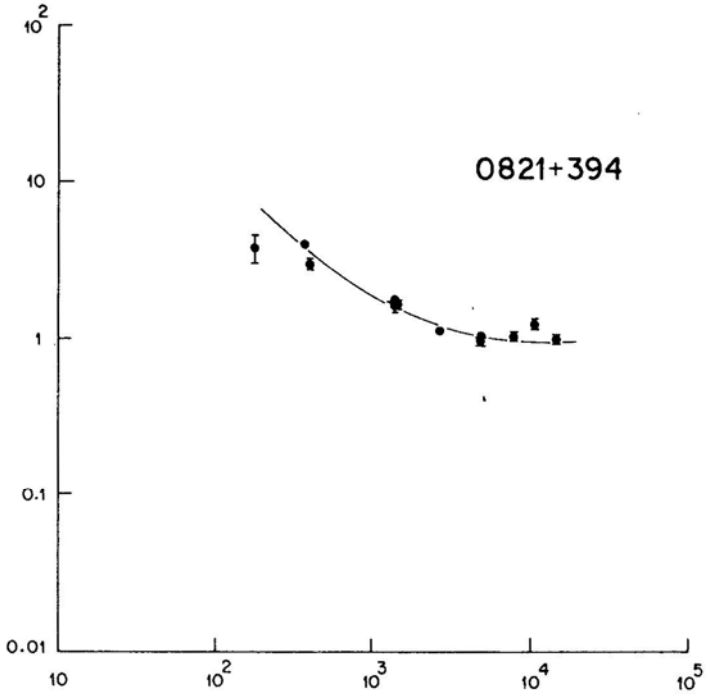


Figure 15. Continued.



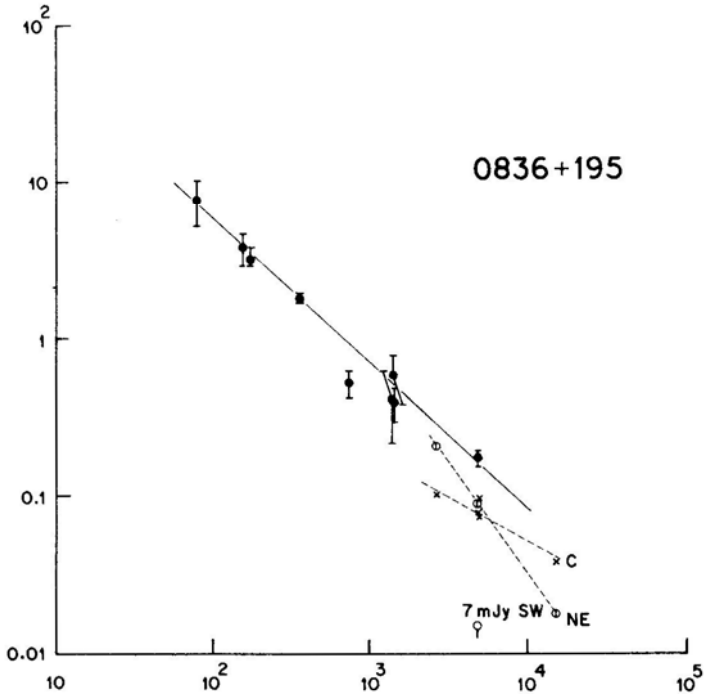


Figure 15. Continued.

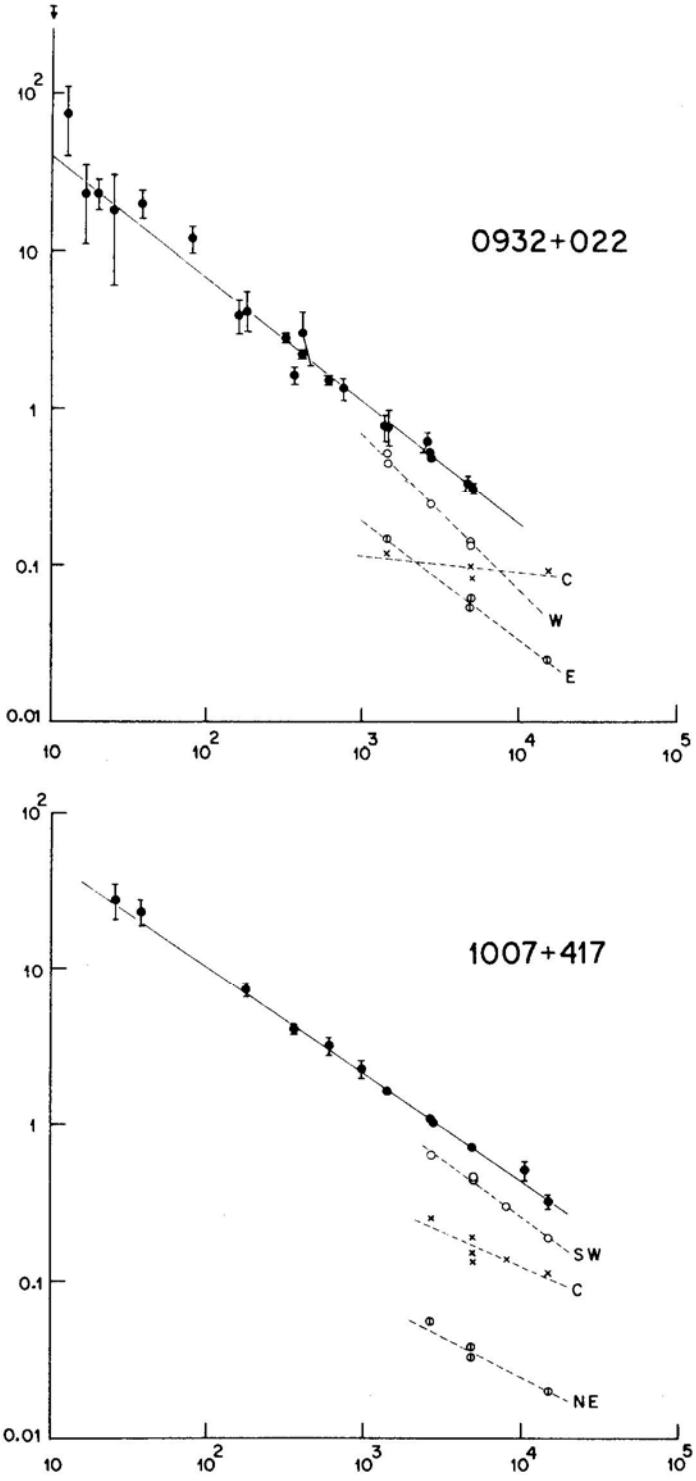
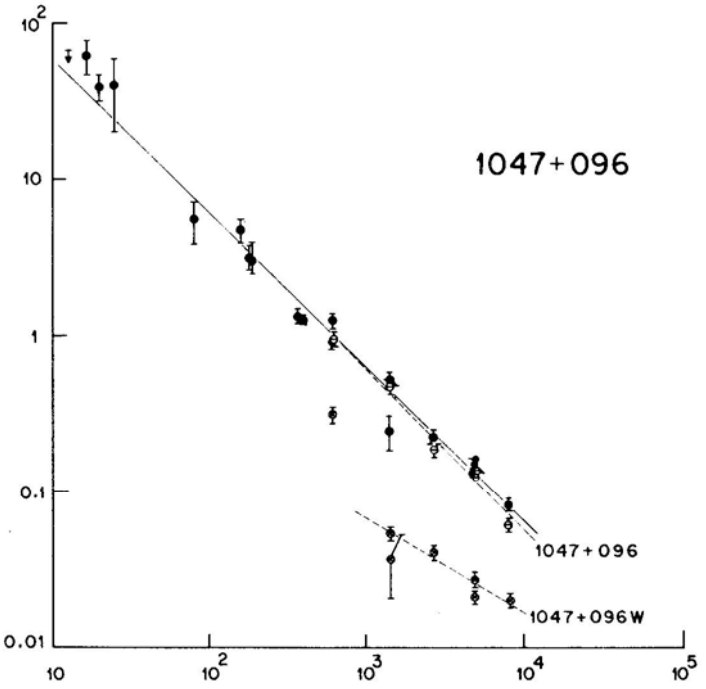


Figure 15. Continued.



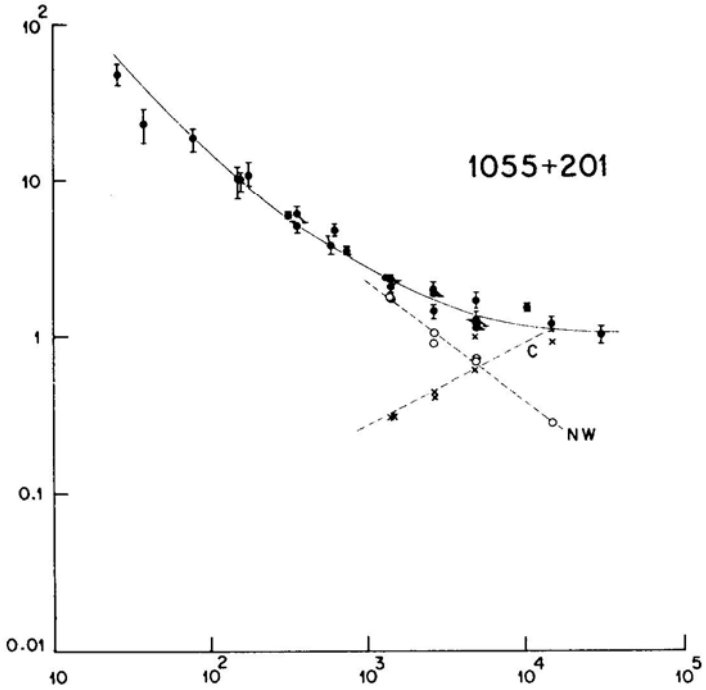


Figure 15. Continued.

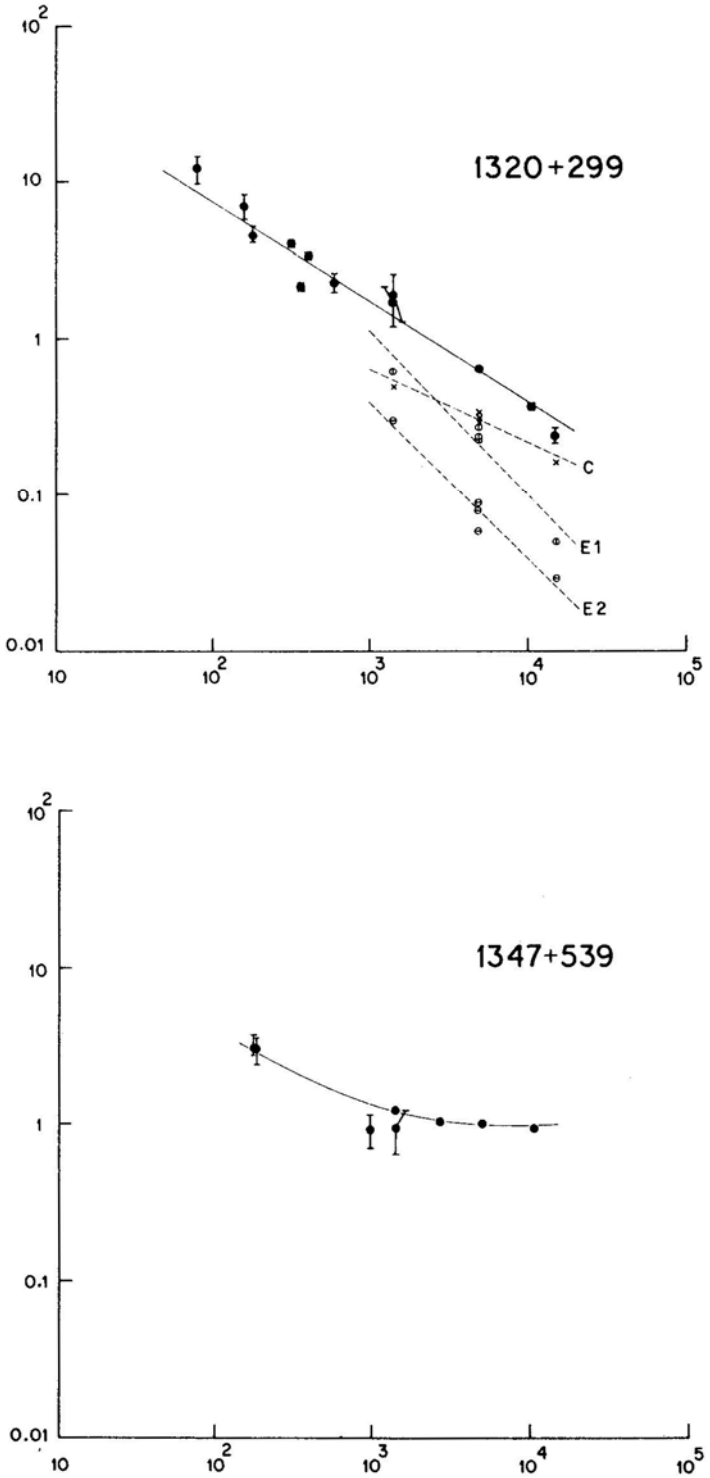
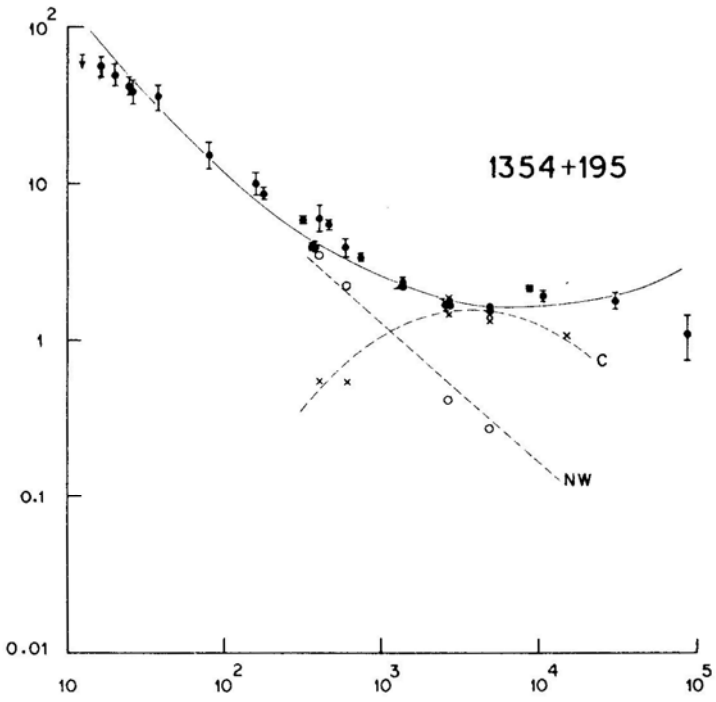


Figure 15. Continued.



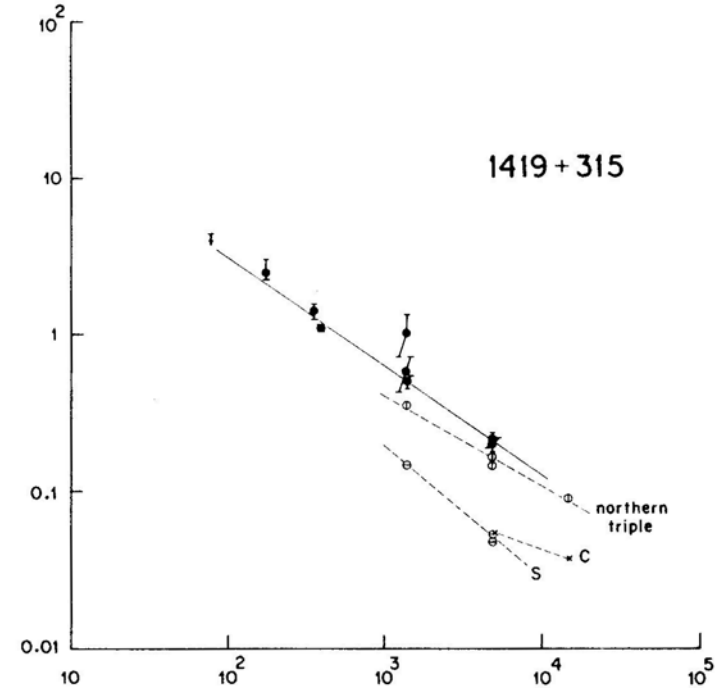
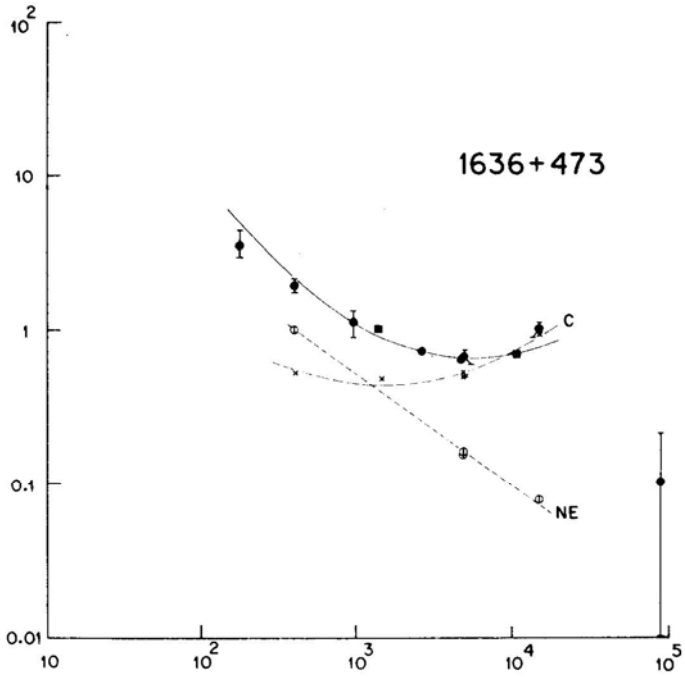


Figure 15. Continued.



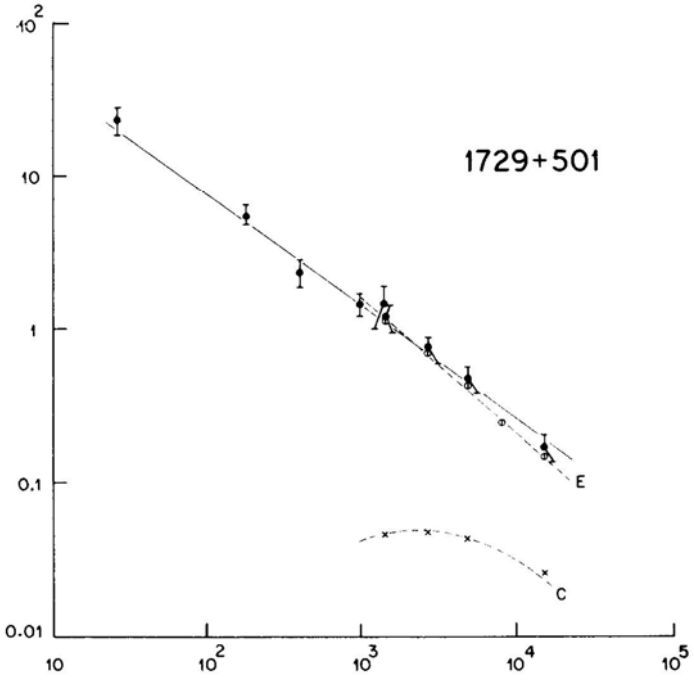


Figure 15. Continued.

A linear fit has been made to the spectrum unless it appears significantly curved, in which case we have fitted a spectrum of the form $\log S = a + b \log \nu + c \log \nu^2$. Both the linear and parabolic least-square fits were made using the NOD2 package (Haslam 1974) at Bangalore.

5. Notes on individual sources

023–2042

This is an asymmetric source where the λ2 cm flux-density ratio of the OCs is ~ 2, and the ratio of their separations from the core is ~ 3. The eastern component is ~ 3 arcsec from the core and hence was not separated from it by MH 78. We find no evidence for the component they suggested 28 arcsec to the west. Hintzen, Ulvestad & Owen (1983, hereinafter HUO) have found a jet/bridge connecting the western component to the core at λ20 cm.

The integrated spectrum of the source is straight below ~ 2 GHz with a spectral index ($S \propto \nu^{-\alpha}$) $\alpha = 0.83 \pm 0.06$, and appears to flatten at higher frequencies. The spectral indices of the individual components are $\alpha_W = 1.05 \pm 0.08$, $\alpha_E \simeq 1.0 \pm 0.1$ and $\alpha_C = 0.38 \pm 0.06$.

The rotation measure (RM) has been reported to be 4.6 ± 1.2 rad m^{-2} with the intrinsic E vector in PA $34 \pm 3°$ by Tabara & Inoue (1980, hereinafter TI80) and 5 ± 1 rad m^{-2} at $32 \pm 3°$ by Simard-Normandin, Kronberg & Button (1981, hereinafter SKB 81). The 5 GHz polarization integrated over the source gives a PA consistent with the above values although the PAs for the eastern and western components differ by ~ 40°. The central component is. unpolarized.

0309 + 411

Kapahi (1979) reported this source to have an unresolved core coincident with an 18-mag galaxy and an extended component ~ 87 arcsec to the north-west with a flux density of ~ 150 mJy at 1415 MHz. Only the core component was detected in our observations. If the extended component, which we estimate from Kapahi's observation to have a size ~ 40 arcsec × $\lesssim$ 40 arcsec, were to possess $\lesssim 3\sigma$ brightness, then its flux density at λ6 cm would be $\lesssim$ 250 mJy. Recently, we have observed this source with the VLA at 1415 MHz with a beam size of about 6 × 4 arcsec but did not detect the outer component implying an upper limit of ~ 160 mJy at 1415 MHz. Although this could be just consistent with Kapahi's observations, we do not find any emission at the peak position of the extended component to a level of ~ 2.5 mJy/beam at 1415 MHz. A VLA D-array λ20 cm map would be useful to investigate the extended structure.

For $\nu \gtrsim 2$ GHz, the continuum spectrum appears to rise slowly. Observations are clearly required to study the spectrum at frequencies $\lesssim 1$ GHz. A large spread is apparent in the measured flux densities at 1.4, 2.7, 5 and 15 GHz implying that the core is strongly variable.

The polarization percentages are similar at λ6 and 2 cm and the measured position angles imply a RM of about -50 ± 25 rad m^{-2}. However, our position angles seem incompatible with the value of $83 \pm 8°$ measured by Simard-Normandin, Kronberg & Button (1982) at 10.5 GHz, although their percentage polarization (1.3 ± 0.4) is in reasonable agreement with our values. Earlier, Simard-Normandin, Kronberg &

Neidhöfer (1981) had reported a polarization percentage of 6.55 ±0.6 at PA= 13 ±2° at 1600 MHz. It seems very likely that the source is also a polarization variable and further measurements would clearly be of interest.

0717+170

This source was not included in the list of one-sided sources of K 81. It was discussed by Joshi (1981), who suggested a possible head-tail structure for the source. From occultation scans at 327 MHz, Joshi concluded that the source consists of a 'head' containing ~ 75 per cent of the total flux density and a 'tail' of length ~ 100 arcsec at PA~ 71° with a weak secondary component at the end of the tail. He found an18-mag red galaxy close to the 'head' of the source and suggested it as a possible identification.

The present observations at l 6cm give a source size of 106 arcsec in PA 66°. However, the morphology of the source looks very similar to a high-luminosity class II source of Fanaroff & Riley (1974), with a possible weak central component which appears extended along the source axis at λ6 cm. The polarization information also suggests that he two outer components represent the lobes of at typical double source. Additional evidence for the western component being a lobe of a double source comes from the λ2 cm map. An extremely relaxed structure is seen with no evidence of any strong compact component coincident with the 18-mag galaxy marked in Fig. 3.

A search near the central component revealed no new candidate optical identification above the limit of the Palomar Observatory Sky Survey (POSS) prints. However, there are several other galaxies in the field (marked on Fig. 3). It seems likely from the above arguments that the proposed identification of Joshi (1981) is a chance superposition. We presently classify this source as unidentified.

The radio source has a straight spectrumbetween80 MHz and 5 GHz with α =0.81 ± 0.04.The two-point spectral index of the western component is 0.85 ± 0.05 between 327 and 4885 MHz. The apparent excess flux density from the western component at 15 GHz may not be significant in view of the low surface brightness of the source at this frequency.

0740 +380.

The source consists of two components separated by ~ 100 arcsec. The stronger is coincident with a 17.6-mag quasar. Whether the two components are physically related is not clear. There is no bridge connecting them but also no optical object is visible on the POSS prints at the position of the SE component.

The core component was found to be a close double of separation 1.8 arcsec at PA 131.5° by Wilkinson, Richards & Bowden (1974). The double structure has been confirmed by Schilizzi, Kapahi & Neff (1982) who report a separation of 1.2 arcsec along PA 137°. The accurate quasar position measured by Clements (1983) is—within the errors—equidistant from the two peaks in the core component.

The integrated radio spectrum appears straight above 100 MHz with α = 1.24 ± 0.02, but appears to flatten below this frequency. The spectral index of the core (~ 1.19 ± 0.04). is similar to the integrated source spectrum. However, the spectrum of the SE component is substantially flatter with a spectral index of 0.71 ± 0.09. VLBI observations of the core at 1417 MHz with a baseline of 1.26 Mλ showed a maximum correlated flux density of ~ 22 per cent of the total flux density at this frequency

(Kapahi & Schilizzi 1979). As noted by these authors, this is unlikely to be due to any flat-spectrum nuclear component.

Using only the measurements > 2σ among those quoted by TI80 and our VLA observations gives a RM ~ 30 rad m^{-2} and an intrinsic PA ~ 15°.

0742 + 376

This source appears to be one-sided with an angular size of ~ 62 arcsec, consistent with the structure reported by Katgert-Merkelijn, Lari & Padrielli (1980). The integrated spectrum is straight between ~ 150 MHz and 5 GHz, with α = 1.00 ± 0.02. The two-point spectral index for the core is about 0.1 ± 0.1 between 5 and 15 GHz. The 15-GHz flux density of the western component is unreliable due to its low brightness. An intriguing aspect of this source is the change in PA of the l6 cm polarization by ~ 70° between components W1 and W2.

0821 + 394

A secondary radio component south-west of the core was reported by Stannard & Neal (1977) and Kapahi (1981b). However, the positions given by these authors differ by ~ 16 arcsec. Stannard & Neal (1977) have also noted that the structure is uncertain due to the presence of a confusing source ~ 6.2 arcmin away along PA 60° (*cf.* Bridle *et al.* 1972). Potash & Wardle (1979) observed this source at 2.7 and 8.1 GHz and did not detect the secondary component, setting an upper limit of 50mJy for it at both frequencies.

We find no evidence for this component in our data and suggest that it is unreal. It appeared barely resolved by Kapahi (1981b) at λ6 cm who found its flux density to be ~ 130 mJy. With our beam of 4.6 ×4.3 arcsec at λ6 cm and a root-mean-square (rms) noise of ~ 6.4 mJy/beam we should have certainly detected this component, if real. It is curious, however, that the spectrum appears to steepen at low frequencies, suggesting that there is extended emission.

The RM is 14 ± 5 rad m^{-2} with an intrinsic PA of 119 ± 7° (SKB 81). Our measured PAs of 116° and 120° at λ*2* cm and λ6 cm respectively are similar to the intrinsic value.

0836 + 195

This is a triple source with a prominent core and a flux-density ratio of ~ 13:1 at λ6 cm for the outer components. The SW component flux density of 7 mJy is consistent with the < 10 mJy limit of MH 78. It is also lower than the noise level of Jenkins, Pooley & Riley (1977). The SW component has possibly been detected by Douglas *et al* (1980) at 365 MHz who deduced the source size to be 28 ± 1 arcsec along PA 9 ± 3°, in broad agreement with our value of ~ 31 arcsec along PA ~ 15°. The integrated spectrum of the source is straight between 80 MHz and 5 GHz with α = 0.92 ± 0.05. The spectral index of the NE component appears to be rather high (~ 1.4) but more accurate flux densities are required to check this. For the core we find α = 0.58 ± 0.07 between 2.7 and 15 GHz.

The RM of the source is 10 ± 1 rad m^{-2} with an intrinsic PA of 98 ± 4° (SKB 81) which is consistent with our observations at λ6 cm.

0932 + 022

The outer components are very asymmetrically located with a separation ratio of ~ 8. The eastern component is at a distance of ~ 5 arcsec from the core along PA ~ 45°, and was not resolved from the core by MH 78. On the western side of the core, there is a weak component at ~ 3.5 arcsec from the core with $\alpha \simeq 1$ and S_{1413} ~ 9 mJy, in addition to the normal edge-brightened lobe at ~ 40 arcsec from the core (HUO; Swarup, Sinha & Hilldrup 1984). This component has not been resolved from the core in our λ6 cm map and is too weak to be detected in the λ2 cm observations.

The spectral index of the entire source is 0.78 ± 0.02 while those of the western and eastern lobes are 0.99 ± 0.05 and 0.76 ± 0.03 respectively. The core spectral index is 0.11 ±0.09 between 1.4 and 15 GHz. From long-baseline interferometric observations at Jodrell Bank, Bentley *et al* (1976) detected a flux density of 150 ± 50 mJy at 1666 MHz from a component with size < 0.1 arcsec. Since both the outer hot-spots appear to be resolved by the VIA with a 0.5 × 0.5 arcsec beam at λ6 cm (Swarup, Sinha & Hilldrup 1984), the compact component seen by Bentley *et al.* (1976) is possibly the core. Within their quoted errors, their 1666-MHz flux density is consistent with the core spectrum.

The RM of this sourceis –11± 1 rad m^{-2} with an intrinsic PA = 157 ± 5° (SKB 81). This is consistent with our values at λ6 cm. The most polarized feature on our map is the SW component which is ~ 15 per cent polarized along PA ~ 159°.

1007 + 417

This source was not in the K 81 list of one-sided sources. It was included in these observations because of the high flux-density ratio for the outer components reported by Owen, Porcas & Neff (1978). The northern component has a λ6 cm flux density of ~ 38 mJy, implying a flux density ratio of ~ 12.

The source has a straight spectrum with α = 0.68 ± 0.01 between 20 MHz and 15 GHz. The spectral indices of the northern and southern lobes are 0.51 ± 0.10 and 0.72 ± 0.04 respectively. The core spectral index is 0.44 ± 0.13. This appears to be steeper than the true spectrum of the core due to the presence of a jet which has been recently mapped by Owen & Puschell (1984) and by Saikia, Kulkarni & Porcas (in preparation).

1047 + 096

This source was suspected to be one-sided from the observations of MH 78. They suggested that the extreme western component, W, ~ 60 arcsec from the central source is related to it and the eastern component, if any, is below the detection limit, implying a flux-density ratio $\gtrsim 6$ at λ6 cm. From the present VLA observations we find the central component of MH78 to be itself a triple with LAS ~ 21 arcsec. This source was also observed at the VLA at 1413 MHz by HUO who did not detect the northern lobe of our triple. We find this northern component to have a total flux density of ~ 13 mJy at λ6 cm and a size of 4.9 × 3.6 arcsec along PA 1°, consistent with its non-detection by HUO to a limit of ~ 4 mJy/beam at 1413 MHz.

MH 78 suggested a possible jet linking the central source to the extreme western component, W. There appears to be no evidence of such a jet in either our data or those of HUO. In the map of HUO, the component W itself seems to be a double source. We

have looked for a possible optical identification of this component but find no object above the limit of the POSS prints. Presently, we suggest that this is an unrelated background source.

The spectrum of the entire complex appears straight between ~ 16 MHz and 8 GHz (α ~ 0.98 ± 0.03). The spectral index of the western, possibly unrelated, source is 0.60 ± 0.10 between 1.4 and 8 GHz, while that of the triple associated with the quasar is 1.02 ± 0.07 between ~ 0.6 and 8 GHz. The 610-MHz flux density of the western source appears to be overestimated and has not been used in evaluating the spectral index.

Reliable measurements of the integrated polarization are not available to determine the rotation measure.

1055 + 201

This source appears to have a one-sided radio structure. HUO have observed it at 1413 MHz and also find it to be one-sided with LAS ~ 21 arcsec. The only suggestion of extended structure on the opposite side of the core is from the observations of Douglas *et al.* (1980) at 365 MHz who find the source to be an asymmetric double with a flux-density ratio of ~ 8.5 and LAS ~ 50 arcsec, but their PA of –78 ± 2° for their model fit is not consistent with our data.

The integrated spectrum of the source is straight below ~ 1 GHz (α = 0.69 ± 0.02) but tends to flatten at higher frequencies. The spectral index of the extended component is 0.76 ± 0.04 between 1.4 and 15 GHz. The core has an inverted spectrum (α = – 0.52 ± 0.12) between 1.4 and 15 GHz. Our measured flux density of the core at 5 GHz is higher than that of MH 78 at the same frequency by ~ 60 per cent suggesting that the core may be variable.

The source has a RM of –22.8 ± 2.2 rad m^{-2} and an intrinsic PA of 123 ± 5° (TI80). Our λ6 cm PAs are consistent with this value.

1320 + 299

This source has an interesting asymmetrical structure with two outer components on the same side of the nucleus and an overall angular extent of ~ 51 arcsec. It has been discussed recently in some detail by Feretti, Giovannini & Parma (1982). Further VLA observations and a more detailed discussion will be presented elsewhere (Cornwell *et al,* in preparation).

The integrated spectral index of the source is 0.64 ± 0.03 between ~ 80 MHz. and 15 GHz. The spectral indices for the core and the components E1 and E2 are 0.46 ± 0.07, 1.04 ± 0.16 and 0.99 ± 0.12 respectively. The outer components appear to have similar spectral indices. However, the spectrum of El appears to steepen above ~ 5 GHz, with $\alpha_{1\,4}^{5}$ = 0.73 ± 0.07 and α_{5}^{15} = 1.43 ± 0.09. Further observations at 15 GHz are needed to verify this steepening in the spectrum.

By combining our measurements with those of Feretti, Giovannini & Parma (1982), we estimate the RM for the nucleus and component E2 to be ~ 2 ± 3 and ~ 3 ± 0.3 rad m^{-2} respectively, the intrinsic PAs being ~ 150° and ~ 10° respectively. The RM of E1 is also small with an intrinsic PA ~ 130°, but has larger uncertainties because of the smaller range in wavelength for which polarization data are available.

1347 + 539

This source was suspected to have one-sided radio structure from the observations of Owen, Porcas & Neff (1978) who detected a marginally resolved component with a flux density of 50 ± 18 mJy at 2695 MHz at a distance of ~ 31 arcsec from the core. We do not detect this component in the present observations with a beam of ~ 4.9 × 3.8 arcsec and an rms noise of ~ 1.8 mJy/beam at λ6cm. Recent VLA observations with the A array at λ6 cm (Perley 1982; Owen & Puschell 1984) show one-sided emission extending upto ~ 5 arcsec to the NW of the core.

1354 + 195

This is a core-dominated triple source with an extent of ~ 44 arcsec, and a possible jet towards the southern component. The outer components are of similar intensity but are asymmetrically located with respect to the core. The separation ratio is 1.76. The SE component was not detected by MH78 and Macdonald & Miley (1971), but has recently been mapped at Cambridge by Peacock & Wall (1982). The radio spectrum is straight below 2 GHz (α = 0.65 ± 0.04) but flattens above this frequency. Assuming that the weaker component detected by Wilkinson, Richards & Bowden (1974) is the core, then the core spectrum turns over around 1–2GHz.The spectral index of the NW component is 1.04 ± 0.05.

Marscher & Broderick (1983) have made VLBI observations of this source and find structure on scales ranging from a fraction of a milliarcsec to ~ 7 milliarcsec. The most compact structure consists of an equal double separated by ~ 0.75 milliarcsec along PA –29°, close to the PA of the entire source.

The RMs suggested for this source are 4.2 ± 0.8 rad m^{-2} with intrinsic PA 72 ±1° (TI80) and 5 ± 1 rad m^{-2} with PA 69 ± 2° (SKB 81). Our λ6 cm PAs are consistent with these values.

1419 + 315

This source was suggested to be one-sided by Fanti *et al.* (1977) and Fanti *et al.* (1979a). While the core component appeared resolved in their observations the outer component located ~ 130 arcsec to the south-east appeared unresolved. In the present observations the outer component still appears unresolved at λ6 cm but the core is resolved into three components misaligned by about 80°. The quasar is co-incident with the middle component which is also the most weakly polarized of the triple and has a flat spectrum (α ~ 0.31 ± 0.11). The magnetic field lines appear to follow the bends in the overall structure.

The spectral index of the northern triple is 0.58 ± 0.07 between 1.4 and 15 GHz while that of the distant southern component is 0.84 ± 0.09 between 1.4 and 5 GHz. The spectrum of the entire complex appears straight between 178 and 5000 MHz with $\alpha \sim 0.69 \pm 0.04$.

The southern component could be an independent source unrelated to the triple structure. An examination of the POSS prints, however, showed no possible identification for this component.

1636 + 473

The extended emission is inclined at ~ 70° to the line joining the core to the nearest peak in the extended structure. The source has been observed with better resolution with MERLIN (Browne *et al.* 1982; Wilkinson 1982).

The extended component appears somewhat similar to a normal double radio source. We have looked for a possible identification for this on the POSS prints but find no likely candidates. It is of interest that from VLBI observations, Porcas (personal communication) finds the core to be extended by ~ 0.5 milliarcsec in the north-south direction.

The integrated spectrum of the source flattens at high frequencies and possibly rises above ~ 10 GHz. Although the 90 GHz measurement has large errors it shows that the spectrum must steepen again between 10 and 90 GHz. While the spectrum of the core appears to be curved, that of the extended emission is straight between 408 MHz and 15 GHz with $\alpha \sim 0.72 \pm 0.03$.

The PAs of polarization vectors in the core are similar at λ6 and 2 cm. Assuming Faraday rotation to be small at λ6 cm, the magnetic field of the extended component appears to be along its axis.

1729+501

This source appears to be one-sided but has a relatively weak core contributing only ~ 10 per cent of the total flux density at λ6 cm. HUO also find the source to be one sided from λ20 cm VLA observations.

The core spectrum appears to flatten below ~ 5 GHz. The integrated spectrum appears straight between 26 MHz and 15 GHz with $\alpha \sim 0.73 \pm 0.03$. The 26-MHz flux density has been estimated by subtracting the extrapolated flux density of the confusing source, 4C49.29, at this frequency. The spectral index of the extended component to the east is 0.87 ± 0.03 between 1.4 and 15 GHz.

6. Conclusions

Of the 17 sources observed, five appear one-sided on the present maps, three are unresolved while seven have radio lobes on both sides of the nucleus. Of the remaining two sources, 0717 + 170 is a double-lobed source whose identification is probably incorrect, while for 0740 + 380, the association of the outer component with the core is uncertain. Higher resolution observations of one of our unresolved sources, namely 1347 + 539, show extended emission on one side of the core (Perley 1982; Owen & Puschell 1984).

Although several sources, previously classified as one-sided, now appear to have lobes on both sides, the outer components either have significantly different surface brightness or are very asymmetrically located with respect to the core. It is also interesting to note that for the 13 sources with extended emission, the average value of the fraction of emission from the core at λ6 cm is ~ 0.5, which is much larger than the corresponding value for a complete sample of sources selected at a low frequency (*cf.* K81).

Acknowledgements

It is a pleasure to thank Chris Salter who made a significant number of the maps presented here. His help and advice during the course of this work has been invaluable. We also thank the NRAO staff for help in the observations. The National Radio Astronomy Observatory is operated by Associated Universities, Inc., under contract with the National Science Foundation.

References

Baars, J. W. M., Genzel, R., Pauliny-Toth, 1.1. K., Witzel, A. 1977, *Astr. Astrophys.*, **61,** 99.
Bentley, M., Haves, P., Spencer, R. E., Stannard, D. 1976, *Mon. Not. R. astr. Soc*, **176,** 275.
Bridle, A. H., Davis, M. M., Fomalont, E. B., Lequeux, J. 1972, *Astr. J.*, **77**, 405.
Bridle, A. H., Fomalont, E. B., Cornwell, T. J. 1981, *Astr. J.*, **86,** 1294.
Browne, I. W. A., Orr, M. J. L, Davis, R. J., Foley, A., Muxlow, T. W. B., Thomasson, P. 1982, *Mon. Not. R. astr, Soc*, **198,** 673.
Clements, E. D. 1983, *Mon. Not. R. astr. Soc,* **203,** 861.
Cohen, A. M., Porcas, R. W., Browne, I. W. A., Daintree, E. J., Walsh, D. 1977, *Mem. R. astr. Soc,* **84**, 1.
Conway, R. G., Davis, R. J., Foley, A. R., Ray, T. P. 1981, *Nature,* **294,** 540.
Douglas, J. N., Bash, F. N., Torrence, G. W., Wolfe, C. 1980, *Univ. Texas Publ. Astr.*, No. 17.
Edwards, T., Kronberg, P. P., Menard, G. 1975, *Astr. J.*, **80,** 1005.
Fanaroff, B. L., Riley, J. M. 1974, *Mon. Not. R. astr. Soc,* **167**, 31P.
Fanti, C, Fanti, R, Formiggini, L., Lari, C, Padrielli, L. 1977, *Astr. Astrophys. Suppl. Ser.*, **28,** 351.
Fanti, R, Feretti, L., Giovannini, G., Padrielli, L. 1979a, *Astr. Astrophys. Suppl. Ser.*, **35,** 169.
Fanti, R, Feretti, L, Giovannini, G., Padrielli, L. 1979b, *Astr. Astrophys.*, **73,** 40.
Feretti, L., Giovannini, G., Parma, P. 1982, *Astr. Astrophys.*, **115,** 423.
Haslam, C. G. T. 1974, *Astr. Astrophys. Suppl Ser.* **15,** 333.
Hewitt, A., Burbidge, G. 1980, *Astrophys. J. Suppl. Ser.*, **43,** 57.
Hintzen, P., Ulvestad, J., Owen, F. 1983, *Astr. J.*, **88,** 709 (HUO).
Jenkins, C. J., Pooley, G. G.. Riley, J. M. 1977, *Mem. R. astr. Soc,* **84,** 61.
Joshi, M. N. 1981, *Mon. Not. R. astr. Soc,* **197,** 7.
Kapahi, V. K. 1979, *Astr. Astrophys.*, **74,** L11.
Kapahi, V. K. 1981a, *J. Astrophys. Astr.*, **2,** 43 (K81).
Kapahi, V.K.1981b, *Astr. Astrophys. Suppl. Ser.*, **43,** 381.
Kapahi, V. K., Schilizzi, R. T. 1979, *Nature,* **277,** 610.
Katgert-Merkelijn, J., Lari, G, Padrielli, L. 1980, *Astr. Astrophys. Suppl. Ser.*, **40,** 91.
Kühr, H., Nauber, U, Pauliny-Toth, I. I. K., Witzel, A. 1979, *Max-Planck-Institut für Radioastronomie Preprint,* No.55.
Kühr, H., Witzel, A., Pauliny-Toth,I.I. K., Nauber, U.1981, *Astr. Astrophys. Suppl. Ser.*, **45,**367.
Macdonald, G. H., Miley, G. K. 1971, *Astrophys. J.*, **164,** 237.
Marscher, A. P., Broderick, J. J. 1983, *Astr. J.*, **88,** 759.
Miley, G. K. 1971, *Mon. Not. R. astr. Soc,* **152,** 477.
Miley G. K., Hartsuijker, A. P. 1978, *Astr. Astrophys. Suppl. Ser.*, **34,** 129 (MH78).
Owen, F. N., Porcas, R. W., Neff, S. G., 1978, *Astr. J.*, **83,** 1009.
Owen, F. N., Puschell, J. J. 1984, *Astr. J.*, in press.
Peacock, J. A., Wall, J. V. 1982, Mon. *Not. R. astr. Soc,* **198,** 843.
Perley, R. A. 1981, in *Optical Jets in Galaxies: Proc. 2nd ESO/ESA Workshop,* ESA SP162, p. 77.
Perley, R. A. 1982, *Astr. J.*, **87,** 859.
Potash, R. L, Wardle, J. F. C. 1979, *Astr. J.*, **84,** 707.
Schilizzi, R. T., Kapahi, V. K., Neff, S. G. 1982, *J. Astrophys. Astr.*, **3,** 173.
Schwab, F. R.1980, in *Proc Int. Optical Computing Conf*., Ed. W. T. Rhodes: *Proc. Soc. Photo-opt. Instrum. Eng*., **231,** 18.

Simard-Normandin, M., Kronberg, P. P., Button, S. 1981, *Astrophys. J. Suppl. Ser.*, **46,** 239 (SKB81).
Simard-Normandin, M., Kronberg, P. P., Button, S.1982, *Astr. Astrophys. Suppl. Ser.*, **48,**137.
Simard-Normandin, M., Kronberg, P.P., Neidhöfer, J.1981,*Astr. Astrophys. Suppl Ser.*, **43,**19.
Stannard, D., Neal, D. S. 1977, *Mon. Not. R. astr. Soc*, **179,** 719.
Swarup, G., Sinha, R. P., Hilldrup, K. 1984, *Mon. Not. R. astr. Soc*, **208,** 813.
Tabara, H., Inoue, M. 1980, *Astr. Astrophys. Suppl. Ser.*, **39,** 379 (TI80).
Thompson, A. R., Clark, B. G., Wade, C. M., Napier, P. J. 1980, *Astrophys. J. Suppl. Ser.*, **44,**151.
Wardle, J. F. C, Kronberg, P. P. 1974, *Astrophys. J.*, **194,** 249.
Wilkinson, P. N. 1982, in *IAU Symp.* 97: *Extragalactic Radio Sources*, Eds D. S. Heeschen & C. M. Wade, D. Reidel, Dordrecht, p. 149.
Wilkinson, P. N., Richards, P. J., Bowden, T. N. 1974, *Mon. Not. R. astr. Soc*, **168,** 515.

J. Astrophys. Astr. (1984) **5**, 475–482

Radio Observations of the BL Lae Object 1400+ 162

D. J. Saikia *Radio Astronomy Centre, Tata Institute of Fundamental Research, Post Box 1234, Bangalore 560012*

G. Swarup *Radio Astronomy Centre, Tata Institute of Fundamental Research, Post Box, 8 Udhagamandalam 643001*

R. P. Sinha *Systems and Applied Science Corporation, 5809 Annapolis Road, Hyattsville, MD 20784, USA*

Received 1984 May 7; accepted 1984 August 23

Abstract. We present total-intensity and linear-polarization observations made with the VLA at 5 GHz of 1400+ 162, a BL Lac object in a group of galaxies. It has a misaligned triple structure with a prominent radio jet towards the east. There is evidence of a weak counter-jet towards the western component, which also has the more prominent warm-spot.

We discuss possible explanations for some of the observed features of this source. Although interaction with the cluster medium is possibly partly responsible for the observed distortion, we suggest that the large observed misalignment could also be due to amplification of a smaller misalignment by projection effects. In the relativistic beaming model, where BL Lac objects arise when the relativistic jets are seen end-on, we suggest that 1400+ 162 is more oblique to the line of sight than most members of this class.

Key words: BL Lac objects—radio structure—linear polarization—relativistic beaming—variability

1. Introduction

The BL Lac objects, characterized by a steep, optical, non-thermal continuum with weak or no emission lines, high optical polarization, strong variability and core-dominated radio structure, have been studied extensively in recent years. Many of their properties are similar to those of highly polarized quasars, which has led to the suggestion that they belong to a single family of active objects called ‘blazars’ (*cf.* Angel & Stockman 1980). It has also been widely suggested that blazars may not be a different class of objects but are merely quasars or giant elliptical galaxies with relativistic jets which are pointed towards us.

In this scenario, the flux density of the core component of a double-lobed source would be enhanced by relativistic beaming, while the outer lobes would appear to merge and give rise to diffuse extended emission around the core. Over the last few years, a large number of BL Lac objects have been mapped with high angular resolution and sensitivity, largely with either the VLA or MERLIN (*e*.g. Kapahi 1979; Weiler & Johnston 1980; Hintzen & Owen 1981; Stannard & McIlwrath 1982; Ulvestad,

Johnston & Weiler 1983; Wardle, Moore & Angel 1984; Antonucci & Ulvestad 1984). These observations do indeed show that most BL Lac objects have prominent cores with diffuse extended emission, often in the form of a halo. The morphology is broadly consistent with the predictions of the relativistic beaming hypothesis.

In this paper we present high-angular-resolution radio observations of 1400 + 162, a BL Lac object which might be more oblique to the line of sight than most members of its class. This source has been observed earlier in total intensity by Baldwin *et al.* (1977), Wills (1979), Weiler & Johnston (1980), Hintzen & Owen (1981), Stannard & McIlwrath (1982), Wardle, Moore & Angel (1984) and Gower (1984). Unlike many BL Lac objects, it has reasonably well-defined radio lobes on opposite sides of the optical object. The lobes are, however, very misaligned, the supplement of the angle formed at the core by the outer warm-spots being ~ 52° (*cf.* Hintzen & Owen 1981).

The source was classified as a BL Lac object by Baldwin *et al.* (1977) on the basis of its steep, polarized non-thermal optical spectrum with no strong emission lines. The high optical polarization has been confirmed by Angel & Stockman (1980) who report a strength of ~ 12 per cent along PA 96°. They also note that the polarization is reasonably steady in PA and perhaps in strength, if one of the measurements quoted in Baldwin *et al* (1977) is assumed to be in error. O'Dell *et al.* (1978) have, however, found that the infrared–visual continuum increased by ~ 50 per cent between mid-1976 and mid-1977. It is a weak, soft X-ray source but 'no variations above 50 per cent were detected between 1979 June and July' (Maccagni & Tarenghi 1981).

The source appears superposed on a group of galaxies (Hazard & Murdoch 1977; Hutchings *et al.* 1984). The redshift of the apparently nearest galaxy is, within the errors, the same as that of the BL Lac object (z = 0.245), suggesting that the source might indeed be associated with the group (Baldwin *et al.* 1977; Miller, French & Hawley 1978). Weistrop *et al.* (1983) have recently presented CCD photometry of some of the members of this group and have suggested that most of them are ellipticals. Hutchings *et al.* (1984) have found that the BL Lac object has a nebulosity which is extended towards the direction of its nearest companion galaxy. Faint optical emission along the direction of the radio jet has also been possibly observed.

In the following sections we first describe our observations and the radio structure of 1400 + 162. This is followed by a discussion on possible explanations for some of the observed features of this source.

2. Observations

The source was observed with 22 antennas of the VLA in the A configuration on 1981 March 17 at 4873 MHz using a bandwidth of 25 MHz. There were 3 short observations of ~ 5 min each at hour angles -3^h, -1.5^h and 0^h, and one at $+4^h$ lasting ~ 9 min. The amplitudes and phases were calibrated using the compact radio source 1345 + 125. The primary flux-density calibrator was 1328 + 307 (3C 286) with S (4873) = 7.41 Jy on the scale of Baars *et al.* (1977). The polarization calibration was done using both 1328 + 307 (3C 286) and 0518 + 165 (3C 138). The total-intensity maps were made using the AIPS package at Charlottesville, while the one with the polarization information was made using the GIPSY package at Ooty. While the total-intensity maps have been self-calibrated this was not possible for the polarization map due to non-availability of software.

3. Results

Fig. 1 shows the total-intensity map of 1400+ 162 with a resolution of 0.43 × 0.41 arcsec^2 along position angle (PA) 75°.9. In addition to the prominent jet towards the east, there also appears to be a weak counter-jet. The radio jet is, within the uncertainties, initially unresolved perpendicular to its major axis, but its most intense feature, E2, appears resolved and is perhaps a region where a significant fraction of the energy carried by the beam is dissipated.

To detect any diffuse, low surface-brightness features which may not have been seen in Fig. 1, we made maps of coarser resolution by tapering the visibility data. In Fig. 2 we show the total-intensity map made with a resolution of 1.40 × 1.35 arcsec^2 along PA 33^o.7, the *u*–*v* taper being 105 kilolambdas. A new component E4 is seen towards the extreme end of the jet. The largest angular size of the source estimated from this map is ~ 20 arcsec which corresponds to a linear size of ~ 97 kpc in an Einstein-de Sitter universe with H_0 = 50 km s^{-1} Mpc^{-1}.

The total-intensity map with the polarized-intensity vectors superposed on the contours is shown in Fig. 3. This map has been made with a resolution of 1.7 × 1.6 arcsec^2 along PA 0° using the GIPSY package at Ooty. Assuming Faraday rotation to be small at this frequency, the magnetic field lines in the inner portion of the jet appear to be largely along its axis. A more sensitive map would be useful to verify this, and also to determine the field orientation in the outer regions. The core appears to be ~ 4 per cent polarized along PA 23°.

Integrated polarization measurements are not available to get a reliable estimate of the rotation measure (RM) of the source. Since it lies at a galactic latitude of ~ 70°, RM

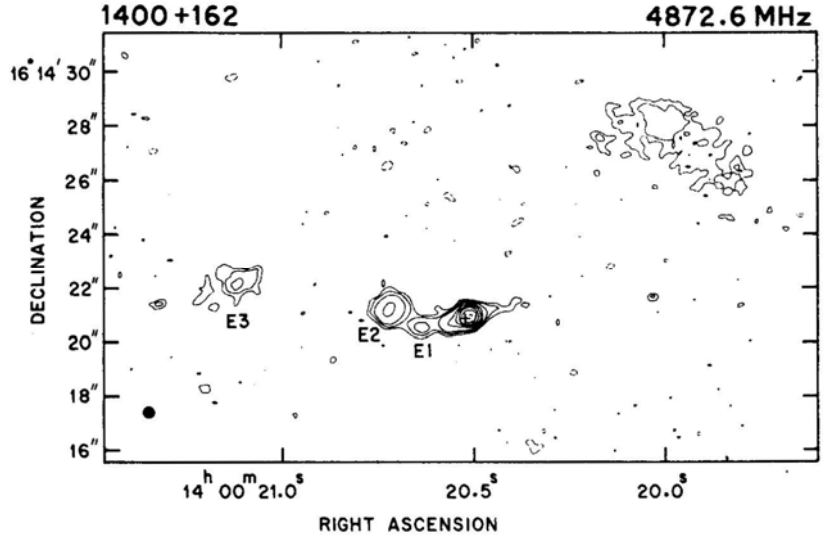


Figure 1. The distribution of total intensity of 1400 + 162 at λ6 cm. The half-power width of the beam, indicated by the shaded ellipse, is 0.43 × 0.41 arcsec^2 along PA 75°.9. The contours are –0.6, 0.6, 1, 2, 8, 16, 32, 64 and 128 mJy/beam. The peak flux density is ~ 145 mJy/beam. In Figs 1 to 4, the position of the optical quasar ($14^h\ 00^m\ 20^s.52$; 16° 14′ 20″.9), marked by the cross, is from Murdoch & Sanitt (1979).

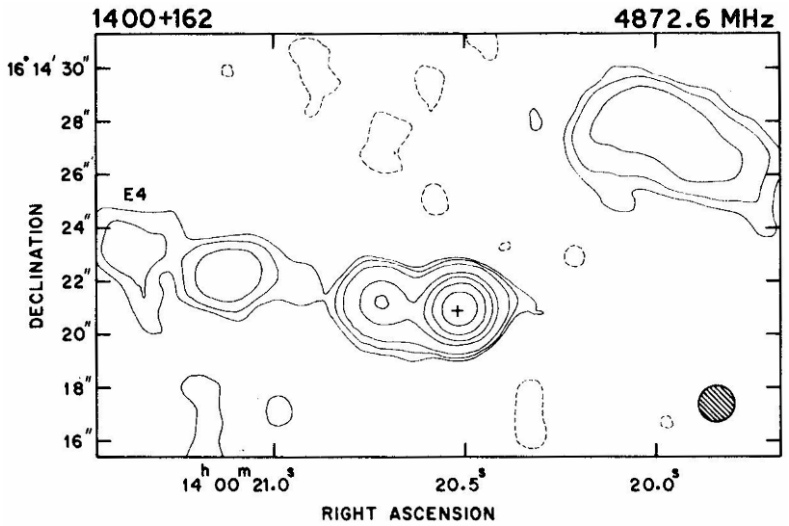


Figure 2. The total-intensity map of 1400+162 at λ6 cm with a resolution of 1.40 × 1.35 arcsec2 along PA 33°.7. The contour levels are –0.9, 0.9, 1.5, 3, 12, 24, 48 and 96 mJy/beam. The peak flux density is ~ 166 mJy/beam.

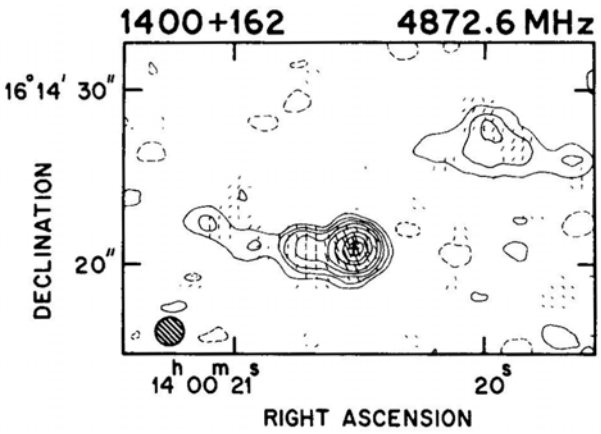


Figure 3. The distribution of total intensity of 1400 +162 at λ6 cm with the polarized-intensity vectors superposed on the contours. One arcsec of the polarization vector corresponds to an intensity of ~ 5.9 mJy/beam. The half-power width of the beam is 1.7 × 1.6 arcsec^2 along PA 0°. The contours are –8,– 4, 4, 8, 12, 20, 32, 64, 96 and 128 mJy/beam. The peak flux density is ~ 165 mJy/beam.

contribution due to the galaxy is small (Simard-Normandin & Kronberg 1980). Polarization measurements at 4.8 and 2.7 GHz by Aller, Aller & Hodge (1982) and Gardner, Whiteoak & Morris (1975) respectively are consistent with a small RM.

The integrated spectrum of the source (Gower, Scott & Wills 1967; Douglas *et al.* 1973; Sutton *et al.* 1974; Slee & Higgins 1975; Murdoch 1976; Slee 1977; Baldwin *et al.*

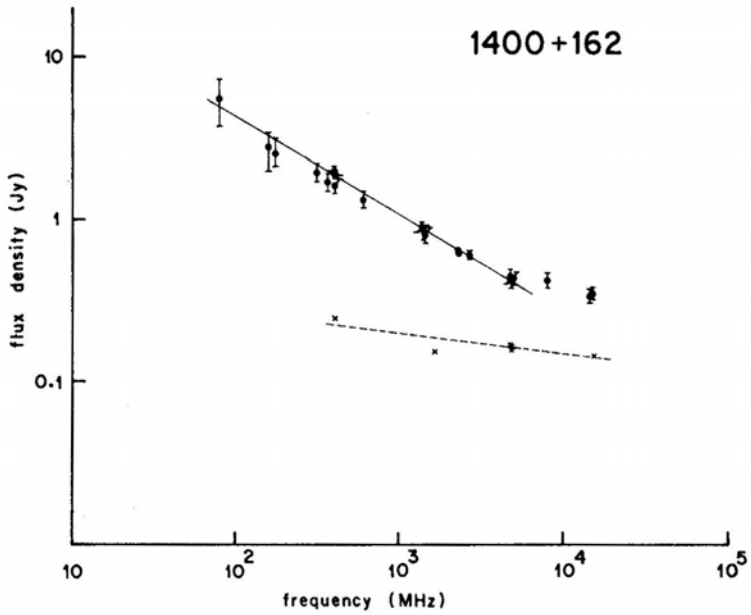


Figure 4. The spectra of the entire source and the radio core. The filled circles represent the total flux density measurements, and the solid curve the best fit to all the points at frequencies $\lesssim$ 5 GHz. The crosses represent the core flux density and the broken line the best fit to these points. We have arbitrarily assumed an error of 10 per cent for those measurements whose errors have not been quoted in the literature.

1977; Weiler & Johnston 1980; Hintzen & Owen 1981; Large *et al.* 1981; Aller, Aller & Hodge 1982; Altschuler 1982, 1983) appears straight below ~ 5 GHz with a spectral index of 0.60 ± 0.01 ($S \propto \nu^{-\alpha}$, but appears to flatten above this frequency. The core has a flat spectrum with $\alpha_c = 0.13 \pm 0.04$ (Fig. 4).

4. Discussion

One of the interesting aspects of the radio morphology of 1400 + 162 is the presence of radio jets. The more prominent jet which is towards the east has no hot-spot at its end, suggesting that a significant fraction of the energy carried by the beam is dissipated along its path. The locations of the components El, E2, E3 and E4 (Figs 1 and 2) indicate that this jet might have an oscillating structure. It is also interesting to note that the most intense feature in the jet, E2, is resolved (Fig. 1), and is perhaps a region where dissipation of energy is high. There also appears to be a weak counter-jet which is on the side of the outer lobe with a more prominent warm-spot. Although a more sensitive, high-resolution map would be useful to investigate the dynamics of the jet in detail, some of its features such as the possible oscillations and also the absence of a hot-spot at its end, are broadly reminiscent of the jet in the well-known quasar 3C280.1 (Swarup, Sinha & Saikia 1982). Swarup *et al,* suggested that the observed morphological features of the jet in 3C280.1 are perhaps largely due to the development of hydrodynamic instabilities.

Another interesting feature of this source is the large misalignment exhibited by the two outer components. Since this source is known to occur in a cluster of galaxies, Hintzen & Owen (1981) have suggested that the observed distortion might be due to interaction with the intracluster medium. In this context it is relevant to note that the luminosity of this source at 178 MHz is $\sim 5 \times 10^{25}$ W Hz^{-1} sr^{-1} which places it in class II of the Fanaroff & Riley (1974) classification scheme. In this class very distorted structures are relatively uncommon.

However, although interaction with the medium could indeed be partly responsible for the observed distortion, an important factor which could also contribute is amplification of a smaller intrinsic misalignment by projection effects when the source is inclined close to the line of sight. Readhead *et al.* (1978) suggested that the observed misalignments of VLBI structures on different scales in core-dominated sources might be due to such an effect. This result was first extended to a large sample of double-lobed quasars by Saikia & Kapahi (1982) who found a correlation between f_c, the fraction of emission from the core at an emitted frequency of 8 GHz, which is used as a statistical measure of the orientation of the source, and Δ, the supplement of the angle formed at the core by the outer hot-spots. The value of f_c for 1400+ 162 is ~ 0.5, and is broadly consistent with the f_c–Δ correlation (Kapahi & Saikia 1982; Saikia 1984).

Saikia & Shastri (1984) have recently examined the relative orientation of the core polarization vector at λ6 cm and the over all radio axes of quasars. The core polarization presumably arises in nuclear VLBI-scale structures. They find that the *E*-vector tends to be perpendicular to the source axis in weak-cored quasars which are presumably at large angles to the line of sight, while the extremely core-dominated sources exhibit no such trend. This result appears to be broadly consistent with the relativistic beaming hypothesis. Since 1400 + 162 is a very misaligned source, it is probably more meaningful to compare the angle, ψ, between the core-polarization *E*-vector and the radio jet. With $\psi \sim 60°$ and $f_c \sim 0.5$, 1400 + 162 fits in the above correlation.

The optical polarization, however, is in PA 96° and is only about 15° off the PA defined by the core and the radio jet. This is consistent with the finding of Moore, Wardle & Angel (1982) that the optical polarization in a small sample of BL Lac objects which have a 'preferred' characteristic PA, usually lies within ~ 15° of the radio structure.

As discussed earlier, it has often been argued that the observed properties of BL Lac objects are due to relativistic jets being seen almost end-on (*cf.* Blandford & Rees 1978; Stannard & McIlwrath 1982; Browne 1983 and references therein). Here, we briefly summarize some of the evidence which suggests that, in this scenario, 1400 + 162 is possibly more oblique to the line of sight than most members of its class.

(i) It has a double-lobed structure. For a source seen end-on, the lobes would merge to form diffuse extended emission around the Doppler-boosted core, as is seen in most BL Lac objects. Stannard & McIlwrath (1982) have noted that jet-like structures are often much less conspicuous in BL Lac objects. They have suggested that this could be due to a combination of geometrical and beaming effects in sources seen end-on. The detection of radio jets in 1400 + 162 is consistent with a slightly larger angle of inclination of the ejection axes to the line of sight.

(ii) The value of f_c for 1400 + 162, which is ~ 0.5, is much smaller than the median of ~ 0.9 for a sample of 38 well-observed BL Lac objects. Since *fc* appears to be a reasonably good statistical measure of the orientation of a source *(cf.* Orr & Browne 1982; Kapahi & Saikia 1982; Moore *et al.* 1981), this is also consistent with a more

oblique angle to the line of sight for 1400+ 162 than most BL Lac objects.

(iii) The absence of strong and rapid variability is also consistent with a larger angle of inclination. As noted earlier, the optical polarization is steady in PA and perhaps also in strength. Altschuler (1982) monitored the total intensity of this source at 2380 MHz over a period of ~ 6 months but found no evidence for variability. He later extended his baseline to ~ 2.5 yr but again found that the source did not vary (Altschuler 1983). A comparison of our measurement of the core flux density at λ6 cm with those of Baldwin *et al.* (1977) and Weiler & Johnston (1980), suggests that, within the uncertainties, the core flux density has remained the same over a period of at least ~ 3 yr.

Acknowledgements

We thank Dilip Banhatti and S. Krishnamohan for a careful reading of the manuscript, an anonymous referee for several valuable comments, the NRAO staff for help in the observations and Ms N. N. Shantha for typing the manuscript. The National Radio Astronomy Observatory is operated by Associated Universities, Inc, under contract with the National Science Foundation.

References

Aller, M. F., Aller, H. D., Hodge, P. E. 1982, in *IAU Symp.* 97: *Extragalactic Radio Sources,* Eds D. S. Heeschen, & C. M. Wade, D. Reidel, Dordrecht, Holland, p. 335.

Altschuler, D. R. 1982, *Astr. J.,* **87,** 387.

Altschuler, D. R. 1983, *Astr. J.,* **88,** 16.

Angel, J. R. P., Stockman, H. S. 1980, *A. Rev. Astr. Astrophys.*, **18,** 321.

Antonucci, R. R. J., Ulvestad, J. S. 1984, *Nature,* **308,** 617.

Baars, J. W. M., Genzel, R., Pauliny-Toth , 1.1. K., Witzel, A. 1977, *Astr. Astrophys.*, **61,** 99.

Baldwin, J. A., Wampler, E. J., Burbidge, E. M.5 O'Dell, S. L., Smith, H. E, Hazard, C, Nordsieck, K. H., Pooley, G., Stein, W. A. 1977, *Astrophys.* J., **215**, 408.

Blandford, R. D., Rees, M. J. 1978, in *Pittsburgh Conf. BL Lac Objects,* Ed. A. M. Wolfe, Univ. Pittsburgh, p. 328.

Browne, I. W. A. 1983, *Mon. Not. R. astr. Soc,* **204,** 23P.

Douglas, J. N., Bash, F. N., Ghigo, F. D., Moseley, G. F., Torrence, G. W. 1973, *Astr. J.,* **78,** 1.

Fanaroff, B. L., Riley, J. M. 1974, *Mon. Not. R. astr. Soc,* **167,** 3IP.

Gardner, F. F., Whiteoak, J. B., Morris, D. 1975, *Aust. J. Phys., Astrophys. Suppl. No.* 35.

Gower, A. C. 1984, in preparation.

Gower, J. F. R., Scott, P. F., Wills, D. 1967, *Mem. R. astr. Soc,* **71,** 49.

Hazard, C, Murdoch, H. S. 1977, *Aust. J. Phys., Astrophys. Suppl. No.* 42.

Hintzen, P., Owen, F. 1981, *Astr. J.,* **86**, 1577.

Hutchings, J. B., Crampton, D., Campbell, B., Duncan, D., Glendenning, B. 1984, *Astrophys. J. Suppl. Ser.,* **55,** 319.

Kapahi, V. K. 1979, *Astr. Astrophys.,* **74,** L11.

Kapahi, V. K., Saikia, D. J. 1982, *J. Astrophys. Astr.,* **3**, 465.

Large, M. I., Mills, B. Y., Little, A. G., Crawford, D. F., Sutton, J. M. 1981, *Mon. Not. R. astr. Soc.,* **194,** 693.

Maccagni, D., Tarenghi, M. 1981, *Astrophys, J.,* **243,** 42.

Miller, J. S., French, H. B., Hawley, S. A. 1978, in *Pittsburgh Conf. BL Lac Objects,* Ed. A. M. Wolfe, Univ. Pittsburgh, p. 176.

Moore, R. L., Wardle, J. F. C, Angel, J. R. P. 1982, *Bull. Am. astr. Soc,* **14,** 934.

Moore, P. K., Browne, I. W. A., Daintree, E. J., Noble, R. G., Walsh, D. 1981, *Mon. Not. R. astr. Soc,* **197,** 325.

Murdoch, H. S. 1976, *Mon. Not. R. astr. Soc*, **177,** 441.
Murdoch, H. S., Sanitt, N. 1979, *Aust. J. Phys.*, **32**, 511.
O'Dell, S. L, Puschell, J. J., Stein, W. A., Warner, J. W. 1978, *Astrophys. J. Suppl. Ser.*, **38**, 267.
Orr, M. J. L., Browne, I. W. A. 1982, *Mon. Not. R. astr. Soc*, **200**, 1067.
Readhead, A. C. S., Cohen, M. H., Pearson, T. J., Wilkinson, P. N. 1978, *Nature*, **276**, 768.
Saikia, D. J.1984, in *Proc. Bangalore Winter School on Extragalactic Energetic Sources*, Ed. V. K. Kapahi, Indian Academy of Sciences, Bangalore (in press).
Saikia, D. J., Kapahi, V. K. 1982, *Bull. astr. Soc. India*, **10**, 42.
Saikia, D. J., Shastri, P. 1984, *Mon. Not. R. astr. Soc*, **211** (in press).
Simard-Normandin, M., Kronberg, P. P. 1980, *Astrophys. J.*, **242**, 74.
Slee, O. B. 1977, *Aust. J. Phys., Astrophys. Suppl. No.* 43.
Slee, O. B., Higgins, C. S. 1975, *Aust. J. Phys., Astrophys. Suppl. No.* 36.
Stannard, D., McIlwrath, B. K. 1982, *Nature*, **298**, 140.
Sutton, J. M., Davies, I. M., Little, A. G., Murdoch, H. S. 1974, *Aust. J. Phys., Astrophys. Suppl. No.* 33.
Swarup, G., Sinha, R. P., Saikia, D. J. 1982, *Mon. Not, R. astr. Soc*, **201**, 393.
Ulvestad, J. S., Johnston, K. J., Weiler, K. W. 1983, *Astrophys. J.*, **266**, 18.
Wardle, J. F. C, Moore, R. L, Angel, J. R. P. 1984, *Astrophys. J.*, **279**, 93.
Weiler, K. W., Johnston, K. J. 1980, *Mon. Not. R. astr. Soc*, **190**, 269.
Weistrop, D., Shaffer, D. B., Reitsema, H. J., Smith, B. A. 1983, *Astrophys. J.*, **271**, 471.
Wills, D. 1979, *Astrophys. J. Suppl. Ser.*, **39**,291.

J. Astrophys. Astr. (1984) **5**, 483–493

Short-Term Intrinsic-Intensity Variations of Pulsars

S. Krishnamohan *Radio Astronomy Centre, Tata Institute of Fundamental Research, P.O.Box 1234, Bangalore560012*

V. Balasubramanian *Radio Astronomy Centre, Tata Institute of Fundamental Research, P.O. Box 8, Ootacamund 643001*

Received 1984 May 30; accepted 1984 August 23

Abstract. Pulsars show intensity variations over timescales ranging from a few microseconds to a few years. Short-term intensity variations, *i.e.* those having timescales of a few minutes to a few hours had been difficult to study as their timescales are similar to those due to interstellar scintillations.

We present here a method to separate the autocorrelation function of the short-term broadband intensity variations from that of the interstellar scintillations and thus overcome the above difficulty. The method assumes that the intrinsic variations are correlated over a bandwidth much larger than the decorrelation bandwidth for scintillations. Hence the ratio of the power in the variations due to the two causes depends on the bandwidth used. By applying the method to the intensity variations of 24 pulsars, we show that the presence of short-term intrinsic variations is very common in the radiation of pulsars. Quasi-periodicities were detected in the intensity variations of many pulsars, but their origin is not clear.

Key words: pulsars—intensity variations

1. Introduction

In contrast to the high stability of their periods, pulsars show erratic variation of the intensity of their radio emission. The intensity variations are produced either by causes intrinsic to the pulsars, or by propagation effects in the interstellar medium. The intrinsic intensity variations (IIV) occur over a wide span of timescales ranging from a few microseconds (Hankins 1971; Bartel & Hankins 1982) to more than a year (Helfand, Fowler & Kuhlman 1977). The timescales for the intensity variations are roughly as follows: (a) a few to several tens of pulse periods for the pulse-to-pulse variations, (b) a few minutes to several hours for the short-term variations and (c) more than a day for the long-term variations. However, Sieber (1982) has shown that long-term variations are likely to be due to propagation effects.

Pulse-to-pulse as well as the long-term variations have been studied very extensively (*e.g.* Backer 1970; Cole, Hesse & Page 1970; Hesse 1972; Huguenin, Taylor & Helfand 1973; Ritchings 1976; Helfand, Fowler & Kuhlman 1977). Short-term variations did not receive much attention mainly because their study is difficult due to the deep fading caused by interstellar scintillations (ISS) at metre wavelengths. All the known pulsars

scintillate in the strong scattering regime at metre wavelengths and hence have modulation indices close to one. In principle the modulation index due to ISS can be brought down to any desired level by averaging the intensity over many decorrelation bandwidths, but it is difficult in practice since the fall in the modulation index with increasing bandwidth is slow (Lee 1976). As such this is not a practical method to suppress fluctuations due to ISS, especially for pulsars of low dispersion measure.

Study of short-term IIV is highly desirable for the following reasons:

(a) The presence of short-term IIV, whatever may be its origin, complicates the interpretation of ISS measurements. It may cause over-estimation of velocities from two-station observations (Slee *et al.* 1974) and considerable error in the measurement of scintillation bandwidths (Backer 1975). Separation of IIV from ISS makes the interpretation of ISS parameters less prone to errors.

(b) Free precession of neutron stars may produce broadband intensity variations with timescales ranging from a few minutes to a few hours (Pines & Shaham 1974).

In this paper, we describe a method for separating the autocorrelation function (ACF) of IIV from that of ISS.

2. The method

The proposed method is based on the fact that intensity variations produced by ISS have small decorrelation frequencies, whereas those due to IIV have large decorrelation frequencies compared to typical observational bandwidths.

The intensity of a pulsar in the channel i of a multichannel receiver can be written as

$$I_i'(t) = a_i\{1 + s_i(t)\}\{1 + p_i(t)\}$$

where the steady intensity of the pulsar, α_i, is modulated by the time-dependent scintillation and intrinsic modulations s_i and p_i respectively. a_i may be a function of the channel number, since different channels may have different overall gains. Since intrinsic intensity variations are frequency independent,

$$p_1(t) = p_2(t) = \ldots = p_n(t) \equiv p(t),$$

where n is the total number of channels. Mean subtracted intensity,

$$I_i(t) = a_i\{s_i(t) + p(t) + s_i(t)p(t)\}.$$

Autocorrelation function (ACF) of the intensity $I_i(t)$,

$$\begin{aligned}\langle I_i(t)I_i(t+\tau)\rangle = a_i^2\{&\langle s_i(t)s_i(t+\tau)\rangle + \langle p(t)p(t+\tau)\rangle \\ &+ \langle s_i(t)s_i(t+\tau)\rangle\langle p(t)p(t+\tau)\rangle\},\end{aligned}$$

where the angular brackets stand for time averaging. Here, we used the fact that $s_i(t)$ and $p(t)$ are statistically independent, and hence (Papoulis 1965),

$$\langle s_i(t)p(t+\tau)\rangle = \langle s_i(t)\rangle\langle p(t)\rangle = 0$$

and

$$\langle s_i(t)p(t)s_i(t+\tau)p(t+\tau)\rangle = \langle s_i(t)s_i(t+\tau)\rangle\langle p(t)p(t+\tau)\rangle.$$

Though they are not equal to zero for finite length of data, the terms $\langle s_i(t)\rangle$ and $\langle p(t)\rangle$ tend to zero as the length of the data increases. We set them to zero in this

presentation of the method as otherwise the expressions become cumbersome. A full analysis in which these terms are not set to zero shows that for most of the results presented in this paper, the error introduced by setting $\langle s_i(t) \rangle = \langle p(t) \rangle = 0$ is a few per cent on $\langle p(t)p(t+\tau) \rangle$ and $\langle s(t)s(t+\tau) \rangle$. For brevity, from now onwards we drop out the arguments of $p(t)$ and $s_i(t)$; *i.e.*, $\langle p(t)p(t+\tau) \rangle \langle pp \rangle$ and so on. The summed up ACF,

$$\sum_{i=1}^{n} \langle I_i I_i \rangle = \{ \langle ss \rangle + \langle pp \rangle + \langle ss \rangle \langle pp \rangle \} \sum_{i=1}^{n} a_i^2 .$$

In arriving at the above equation, we have assumed that,

$$\langle s_i s_i \rangle = \langle s_j s_j \rangle \equiv \langle ss \rangle .$$

The above assumption is justified due to the following reason. Though the width of the ACF of intensity variations due to ISS depends linearly on the radio frequency (f), the widths of $\langle S_i S_i \rangle$ and $\langle S_j S_j \rangle$ would be practically the same since the difference in the central frequency of different channels is negligible compared to their central frequencies. If by broadband intensity (I_B) we denote,

$$I_B(t) = \sum_{i=1}^{n} I_i(t),$$

then

$$\langle I_B I_B \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle I_i I_j \rangle$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \{ \langle s_i s_j \rangle + \langle pp \rangle + \langle s_i s_j \rangle \langle pp \rangle \}.$$

To progress further, we make the following two simplifying assumptions:

(i) For ISS, the cross-correlation function of intensities at two close by frequencies and the ACF at one of the frequencies differ only by a constant factor. The error introduced by this assumption is small as can be inferred from the numerically computed curves presented in Figs 5, 6 and 7 of Lee (1976).

(ii) The multiplicative factor depends only on the frequency separation and not on the actual frequencies. This assumption is justified since,

$$BW/f \ll 1$$

where BW is the total bandwidth used for the observations.

Under these assumptions,

$$\langle s_i s_j \rangle = \langle ss \rangle C_{|i-j|}$$

where $C_{|i-j|}$ is the correlation coefficient between the intensity fluctuations due to ISS observed in channels i and j for zero time lag. Using the above relation, we get,

$$\langle I_B I_B \rangle = \langle pp \rangle \sum_i \sum_j a_i a_j + \langle ss \rangle (1 + \langle pp \rangle) \sum_i \sum_j a_i a_j C_{|i-j|}.$$

On writing

$$K1 \equiv \sum_i a_i^2, \quad K2 \equiv \sum_i \sum_j a_i a_j, \quad \text{and} \quad K3 \equiv \sum_i \sum_j a_i a_j C_{|i-j|},$$

we obtain

$$CN \equiv \sum_i \langle I_i I_i \rangle = K1 \langle pp \rangle + K1 \langle ss \rangle (1 + \langle pp \rangle)$$

and

$$CB \equiv \langle I_B I_B \rangle = K2 \langle pp \rangle + K3 \langle ss \rangle (1 + \langle pp \rangle),$$

where *CN* stands for the summed-up temporal auto-correlation coefficients for the narrow bandwidths and *CB* stands for the temporal auto-correlation coefficients for the broad bandwidth. From the above two equations we obtain,

$$\langle pp \rangle = \frac{CB/K3 - CN/K1}{K2/K3 - 1} \quad \text{and} \quad \langle ss \rangle = \frac{CN/K1 - \langle pp \rangle}{1 + \langle pp \rangle}.$$

One should remember that *CB, CN,* ⟨ pp ⟩ and ⟨ *ss* ⟩ are all functions of the lag τ which had been dropped out for notational convenience.

This method gives, in addition to the ACF of the intrinsic intensity variations (⟨ pp ⟩), the ACF of ISS (⟨ ss ⟩) after correcting it for the intrinsic variations which may be present in the data. Hence it is useful in studying ISS also. This aspect is discussed further in a separate paper (Balasubramanian & Krishnamohan 1984; hereinafter BK).

This method is applied on the data for 24 pulsars as discussed below. To apply the method, intensity fluctuations should be observed by using several narrowband channels. For each pulsar there is a range of bandwidths for which the method gives reliable estimates of ⟨ *pp*⟩ as well as ⟨ *ss* ⟩. If the bandwidth of the individual channels is much larger than the ISS decorrelation bandwidth, then most of the fluctuations due to ISS are washed out even in the intensity variations from individual channels making estimation of ⟨ *ss* ⟩ unreliable, On the other hand, if decorrelation bandwidth due to ISS is larger than the total bandwidth covered by all the narrow-band channels, then the method cannot give any information even on ⟨ *pp* ⟩ as there is no way of distinguishing fluctuations due to IIV from those due to ISS. This can also be seen directly from the equation for ⟨ *pp* ⟩. If $C_{|i-j|}$ are practically equal to 1 for all *i* and; *j*, then *K*3 = *K2* and hence the expression for ⟨ *pp* ⟩ blows up. Among the 24 pulsars listed in Table 1, there is no pulsar for which large decorrelation bandwidth due to ISS is a limitation in applying the method. For all the pulsars, except PSR 1857 – 26, for which no ISS timescales are listed in Table 1 the dispersion measure is greater than 73 cm^{-3} pc. For such pulsars the expected decorrelation frequency is much smaller than 50 KHz and for some of them the expected decorrelation time is less than the averaging time (see for *e.g.* BK). They are included in the sample as IIV with timescales larger than the averaging time can still be detected, if present. Though the dispersion measure is only 38.1 cm^{-3} pc for PSR 1857 – 26, no ISS was detected.

To apply the method one should measure *CN* and *CB* from the multichannel data. Summing up of temporal autocorrelation functions of intensities from individual channels gives *CN*. Temporal autocorrelation function of summed-up intensities from individual channels gives *CB.* In addition to *CN* and *CB* one should measure *K*1, *K*2 and *K*3. *K*1 and *K2* are obtained directly from the gains of the individual channels. $C_{|i-j|}$ needed to determine *K3* are obtained by using a two-step process. In the first step, a decorrelation frequency is determined from the multichannel data by assuming Kolmogorov density spectrum for the irregularities in the interstellar medium as discussed in BK. The determination takes into account (a) presence of IIV, (b) finite signal-to-noise ratios and (c) finite bandwidth and overlap of the filters used. In the

next step the expected $C_{|i-j|}$, which are the correlation coefficients for ISS expected in the absence of noise and IIV are computed back again assuming Kolmogorov spectrum. The results obtained by using Gaussian spectra are not significantly different.

3. Observations and analysis

The observations were made during 1976–78 using the Ooty radio telescope which operates at 326.5 MHz. A 12-channel receiver with a selectable channel bandwidth of either 300 or 50 KHz was used for the observations. The separation between the central frequencies of the adjacent channels was also either 300 or 50 KHz, with the exception that the frequency separation between the last three channels for the 50 KHz filters was 500 KHz. From the data acquired using a general purpose programme, two data arrays per channel are formed. The first array $L_i(m)$ contains averaged ONPULSE values for channel *i*. The averaging is done, after subtracting base levels, over a prescribed number of pulses. The base levels are determined by averaging samples over a duration comparable to ONPULSE duration midway between adjacent pulses. The second array I_i^0 (m) contains the averaged OFFPULSE values which are formed in a similar manner from the samples on either side of the pulse.

For each pulsar the decorrelation frequency is determined. The determination takes into account (a) presence of IIV, (b) finite signal-to-noise ratios, and (c) finite bandwidth and overlap of the filters used. The data arrays I_i(m) and I_i^0 *(m)* for each pulsar are subjected to the method described in the previous section using appropriate correlation coefficients.

Let us illustrate the method by applying it to the data on PSR 1919 + 21 obtained on 1976 June 24. In this paper we deal with modified autocorrelation function (MACF). By MACF we mean the autocovariance values, normalized with the square of the mean intensity instead of the variance. With this convention the MACF for the zero lag gives directly the square of the modulation index. Fig. 1 gives the MACFs for (a) the

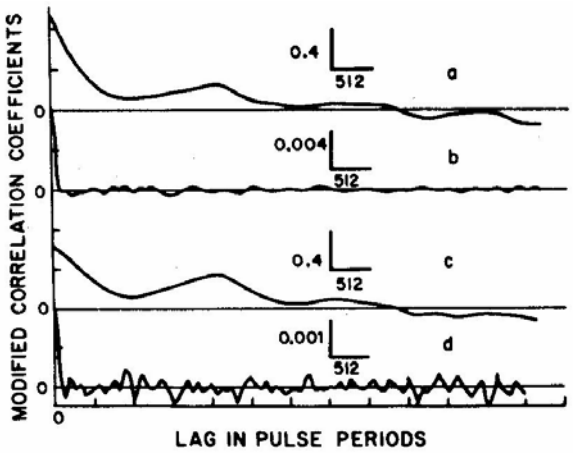


Figure 1. Modified autocorrelation functions for (a) the ONPULSE, and (b) the OFFPULSE narrowband intensity variations, and (c) the ONPULSE and (d) the OFFPULSE broadband intensity variations. The inserts give the scales for the correlation coefficients and the lag in pulse periods.

ONPULSE and (b) the OFFPULSE narrowband intensity variations and (c) the ONPULSE and (d) the OFFPULSE broadband intensity variations. The MACFs of the narrowband channels are averaged over all the 12 channels. For normalizing the OFFPULSE MACFs, we used the corresponding ONPULSE mean intensities since the mean intensities are equal to zero for the OFFPULSE array, and the OFFPULSE MACFs are computed only to check the stability of the receiver system. The vertical and the horizontal bars above each curve give the scale used for the correlation values and the lag respectively.

The MACF for the narrowband intensity, by itself, suggests that quasiperiodic intensity modulations are present since secondary maxima are seen in it. A comparison between the MACFs for the narrowband and the broadband intensities suggests that the quasiperiodicities are due to broadband modulations as the correlation value of the

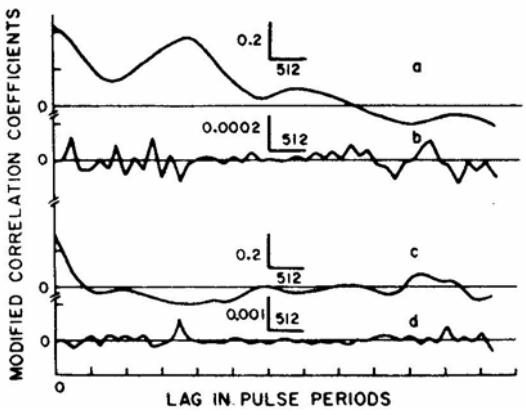


Figure 2. Results of applying the method to the correlation functions shown in Fig. 1. MACFs due to broadband (a) ONPULSE and (b) OFFPULSE intensity variations and narrowband (c) ONPULSE and (d) OFFPULSE intensity variations.

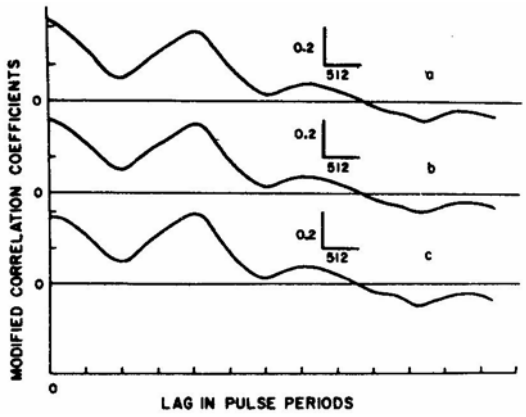


Figure 3. Results of analyzing the data with three different decorrelation frequencies: (a) 300 kHz, (b) 375 kHz, and (c) 450 kHz. The channel width used was 300 kHz.

secondary maxima of the broadband MACF remain as high as that for narrowband MACF. In contrast, there is an appreciable decrease in the value at the zero lag indicating the presence of a nonperiodic narrowband component. The result of applying the method to the MACFs is shown in Fig. 2. The MACF of the IIV separated out by the method clearly shows the presence of quasi-periodicities. There is no evidence for such quasi-periodicities in the MACF of the scintillations.

The MACFs for the noise are obtained from the MACFs of the OFFPULSE intensity. While comparing them with those for the ONPULSE intensities, it should be noted that the noise MACFs are magnified vertically a few hundred times compared to the ONPULSE MACFs.

The results of analysing the data with three different decorrelation frequencies are shown in Fig. 3. The decorrelation frequency for this pulsar is 330 ± 160 KHz. The differences in curves a, b, c in Fig. 3 are negligible, showing that the method is not very sensitive to the assumed decorrelation frequency, if the decorrelation frequency is within the error limits obtained from the data.

4. Results and discussion

The observational details and the results of the application of the method to the 24 pulsars are listed in Table 1. If a pulsar is observed on more than a day, the entries for that pulsar are made in chronological order. Column 5 refers to the signal-to-noise ratio of the ONPULSE intensity obtained by averaging the number of pulses listed in Column 4.

Entries in Column 10 show that IIV with timescales of several tens of minutes are present for many pulsars. These timescales are similar to those for ISS listed in Column 9. A comparison of Columns 7 and 8 shows that IIV can affect appreciably the measurement of ISS parameters since the modulation indices due to the two causes are similar. This aspect is further discussed by BK. If a quasi-periodicity is seen in the MACF of the IIV, the timescale listed in Column 10 is followed by the symbol *'p'*. Quasi-periodic modulations are evident in the MACFs of the pulsars PSRs 0301 + 19,0611+22,0740–28,0818–13,0823 + 26,1237 + 25,1818 –04,1919 + 21 and 2020 + 28 as can be seen in Fig. 4.

Each of the pulsars PSRs 0301 +19,0611 + 22 and 0818–13 were observed on two different days and PSR 1919 + 21 was observed on four different days. The actual dates of observations are marked on the MACF plots. PSRs 0301 + 19,0611 + 22 and 0818 –13 showed quasi-periodicities during only one of the trials. In the case of PSR 1919 + 21 nonrandom features were present in two of the four trials. The pulsar was strong on all the four occasions and the modulation indices due to IIV were all greater than 0.25. Notably, the MACFs obtained from the data acquired on two adjacent days showed dissimilar behaviour. For this pulsar, the presence of wobble was put forward as a possible explanation for the quasi-periodic modulations of the P_3 (Wolszczan 1978). To check whether the IIV is also caused by wobble of the pulsar, we obtained the average pulse shape at 8 different phases of the suspected wobble period. No variation of the pulse shape with the phase of the periodicity was found. Nor did we find any systematic residuals in the times of arrival. If the IIV are due to wobble, from the modulation index one expects systematic residuals ~ 15 ms as well as noticeable changes in the pulse shape if the emission beam of the pulsar could be represented by a

Table 1. Observational details and the results of application of the method to 24 pulsars.

Serial No.	PSR	Channel bandwidth (kHz)	No. of pulses averaged	Signal-to-noise ratio	Data length (hr)	Modulation index		Time scales (min)	
						ISS†	IIV	ISS	IIV*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	0301+19	300	32	1.5	3.0	0.4	0.2	15.5	90.3
2	0301+19	300	32	4.6	6.7	0.5	0.1	12.6	11.8*p*
3	0611+22	50	32	2.2	1.8	<0.1	0.2	—	15.7*p*
4	0611+22	300	64	2.7	6.1	< 0.1	0.1	—	132
5	0628−28	300	16	6.5	3.6	0.5	0.3	9.6	61
6	0740−28	50	32	1.5	0.9	< 0.1	0.1	—	9.6*p*
7	0818−13	300	16	5.0	1.6	0.1	0.1	2	31
8	0818−13	50	32	3.8	2.1	0.4	0.1	2	17.2*p*
9	0823+26	300	32	7.3	2.5	0.5	0.2	4.6	17.6*p*
10	0834+06	300	32	14.4	6.1	0.7	0.3	9.2	27.9
11	1237+25	300	32	12.9	5.8	0.6	0.8	12.5	19.2*p*
12	1604−00	300	64	7.4	2.1	0.5	0.3	18.9	126
13	1642−03	300	32	23.0	1.8	0.2	< 0.05	2.3	—
14	1718−32	50	32	1.4	1.2	< 0.2	0.1	—	37.7
15	1818−04	50	32	3.5	2.3	< 0.1	0.1	—	11.1*p*
16	1831−03	50	32	1.2	1.7	< 0.1	< 0.1	—	—
17	1831−04	50	16	0.7	0.4	< 0.2	< 0.1	—	—
18	1845−01	50	16	0.7	0.7	< 0.2	< 0.2	—	—
19	1845−04	50	32	1.2	0.7	< 0.1	< 0.2	—	—
20	1857−26	50	64	3.8	3.0	< 0.1	< 0.1	—	—
21	1900+01	50	16	0.9	0.3	< 0.3	< 0.2	—	—
22	1907+02	50	32	1.0	2.2	< 0.2	< 0.05	—	—
23	1917+00	50	32	0.9	3.1	< 0.1	< 0.2	—	—
24	1919+21	300	64	14.0	5.9	0.5	0.7	11.1	45.6*p*
25	1919+21	300	32	9.0	3.5	0.6	0.3	9.3	25.7
26	1919+21	300	32	15.5	6.0	0.8	0.4	9.8	14.3*p*
27	1919+21	300	32	12.9	4.1	0.6	0.3	8.6	10.3
28	2016+28	300	32	14.0	2.0	0.2	0.1	23.8	51
29	2020+28	50	64	3.5	3.1	0.6	0.3	7.6	26 *p*
30	2303+30	50	16	0.8	1.0	0.5	< 0.1	4.2	—

† Not corrected for finite bandwidth of the filters.
* The timescale is followed by the symbol *p*, if a quasi-periodicity is seen.

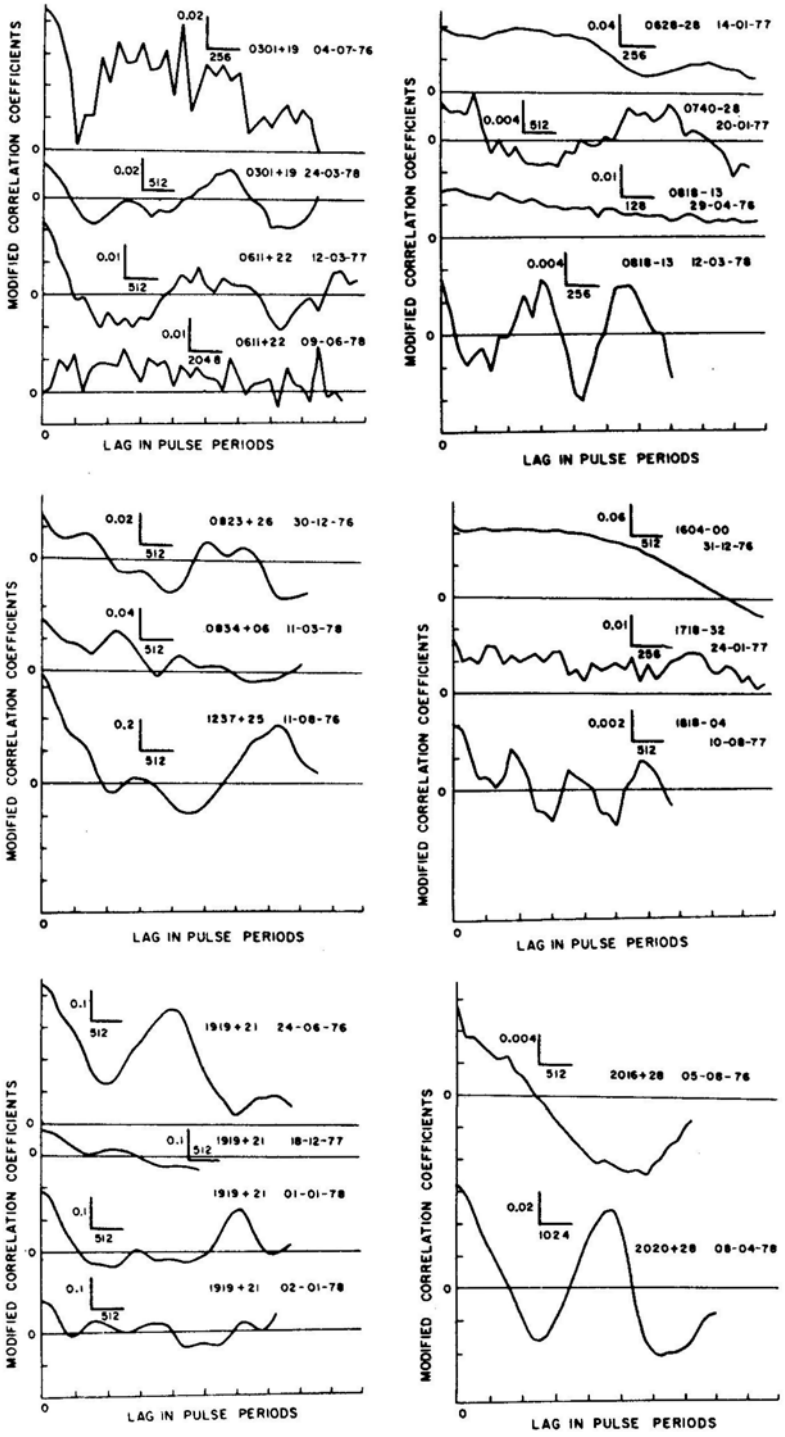


Figure 4. Modified correlation functions of the intrinsic intensity variations, as separated by the method given in Section 2. The data used were collected using 12 channels of either 300 or 50 KHz channel bandwidth. The observational details are given in Table 1.

two-dimensional Gaussian. The expected residuals for the other pulsars are less than ~ 3 ms, since their modulation indices are much less. The above estimates of the residuals are very uncertain since the emission beams of pulsars may not, even crudely, resemble two-dimensional Gaussians.

It should be emphasized that our method ascribes all broadband intensity variations, irrespective of their origin, to IIV. Thus, for example, any intensity variations caused by gain variations common to all the channels would be ascribed to IIV. But, it is very unlikely that the observed IIV are in fact caused by gain variations since the base levels in between pulses were stable over the periods of observation. An instrumental effect that affects only the. ONPULSE MACF but not the OFFPULSE MACF can arise as the feed of the antenna used is linearly polarized. Such an instrument can in principle introduce fully correlated intensity variations in all the channels since pulsar emission is in general highly linearly polarized. In our sample, we have several pulsars whose fractional average linear polarization is negligible compared to the modulation index of the IIV, implying that all the observed IIV cannot be explained away as due to instrumental effects.

One may suspect that the drifting features observed in the dynamic spectra of several pulsars (Roberts & Abies 1982, Ewing *et al.* 1970) would produce quasi-periodic $\langle pp \rangle$. We did not observe significant quasi-periodic variation, as a function of frequency, of the zero-lag cross-correlation coefficient between the intensities from different frequency channels for any of the pulsars. This implies that either the sloping bands were absent in our data, or though present the intensity variations produced by them are not correlated over the bandwidths used for the observations. In either case one has to look for other explanations for the behaviour of $\langle$ pp $\rangle$. Even if strictly periodic features were present in the dynamic spectra, the intensity variations produced by them would not be ascribed by the method to IIV, since the intensities at different frequencies do not peak at the same time. However, the broadband intensity variations denoted by IIV in this paper themselves could be due to some non-conventional form of scintillations.

Acknowledgements

It is a pleasure to thank the staff of the Radio Astronomy Centre, Ootacamund for the support they gave during the observations, Mr. D. K. Mohanty for several comments, and an anonymous referee whose suggestions added to the clarity of the paper.

References

Backer, D. C. 1970, *Nature,* **228,** 42.
Backer, D.C. 1975, *Astr. Astrophys.*, **43**, 395.
Balasubramanian,V., Krishnamohan, S. 1984, *J. Astrophys .Astr.* (submitted) (**BK**).
Bartel, N., Hankins, T. H. 1982, *Astrophys. J.,* **254,** L35.
Cole, T. W., Hesse, H. K., Page, C. G. 1970, *Nature,* **225,** 712.
Ewing, M. S., Batchelor, R. A., Friefeld,R. D., Price, R. M., Staelin, D. H.1970, *Astrophys. J.,* **162,** L169.
Hankins, T. H. 1971, *Astrophys. J.,* **169,** 487.
Helfand, D. J., Fowler, L. A., Kuhlman, J. V. 1977, *Astr. J.,* **82,** 701.
Hesse, K. H. 1972, *Nature Phys. Sci.,* **235,** 27.
Huguenin, G. R., Taylor, J. H., Helfand, D. J. 1973, *Astrophys. J.,* **181,** L139.
Lee, L. C.1976, *Astrophys. J.,* **206,** 744.

Papoulis, A. 1965, *Probability, Random variables and Stochastic Processes*, McGraw-Hill, New York, p.337.
Pines, D., Shaham, J. 1974, *Nature*, **248**, 483.
Ritchings, R. T. 1976, *Mon. Not. R. astr. Soc.*, **176,** 249.
Roberts, J. A., Abies, J. G. 1982, *Mon. Not. R. astr. Soc.*, **201,** 1119.
Sieber, W. 1982, *Astr. Astrophys.*, **113,** 311.
Slee, O. B., Abies, J. G., Batchelor, R. A., Krishna-Mohan, S., Venugopal, V. R., Swarup, G. 1974, *Mon. Not. R. astr. Soc.*, **167,** 31.
Wolszczan, A. 1978, *Astr. Astrophys.*, **63,** 425.

J. Astrophys. Astr. (1984) **5**, 495–506

Faster-than-Light Motion in Quasars

J. V. Narlikar & S. M. Chitre *Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Bombay 400005*

Received 1984 July 27; accepted 1984 September 20

Abstract. Over the past fifteen years, observations of some quasars with the techniques of very-long-baseline interferometry have shown that the angular separation between pairs of radio-emitting regions in their cores is increasing year after year. If the quasars are indeed as far away as implied by Hubble's law, then these angular motions translate into linear speeds several times the speed of light. Several theoretical scenarios have been proposed to show that the observed motions are illusory. The leading contender in this field—the relativistic beam model—and an alternative offered by the concept of a gravitational screen are described and compared in the light of recent observational data.

Key words: quasars—superluminal motion—relativistic beaming—gravitational screen

1. Introduction

The first hint of apparent faster-than-light motion in quasars came from a series of transpacific observations made between 1967 and 1969, of the sizes of variable components in quasars 3C 273 and 3C 279 (Gubbay *et al.* 1969; Moffet *et al.* 1972). More direct evidence for such motions came in 1971 which indicated a double structure, with the two components having separated with a linear velocity 5 to 10 times the speed of light over a period of a few months (Knight *et al.* 1971; Cohen *et al.* 1971; Whitney *et al.* 1971). Although the early data could be partially discounted on the grounds of ambiguity, imperfect resolution and single baseline observations, later studies extending over the past decade or so with increasingly more sophisticated techniques of very-long-baseline interferometry (VLBI) have confirmed a *prima facie* case for superluminal separation of pairs of radio emitting regions within quasars(Cohen & Unwin 1984, and several other papers in Fanti, Kellermann & Setti 1984). The observational results are summarized in Table 1.

While the astronomical scenarios subject the fundamental laws of physics to far more stringent tests than ever possible in the terrestrial laboratory, theoretical astronomers by and large tend to be conventional in outlook, relying as far as possible on the laws of physics *known* at the time of observation. This attitude can be justified by invoking Occam's razor, although one cannot help wondering why the universe should choose to reveal at any given time only that much of its storehouse of mysteries as is understandable in the framework of the physics known at that time.

Following the above attitude we rule out the conclusion that the observed motions are both superluminal and real; for such a conclusion would hit at the very basic tenets

Table 1. Superluminal sources (H_0 = 50 km s^{-1} Mpc^{-1}).

Source*		Redshift	V/c
3C 120	A	0.033	4.2
	B		8.2
	C		7.4
	D		5.2
BL Lac		0.069	3
3C 273	C1	0.158	10.6
	C2		14.1
	C3		10.6
3C 279		0.538	6.8
3C 345	O	0.595	18
	I		13
3C 179		0.846	8.5
NRAO 140		1.258	10.8

* A, B, C . . . *etc.* refer to the separating components; it is observed that some components fade away to be replaced by others (kellermann & pauliny-Toth 1981)

of special relativity. The observed features of quasar radiation do not suggest anything so extraordinary about the radiating particles that one has to appeal to tachyons.

There is another way of resolving the difficulty of superluminal motions. Since all the observations relate to measurements of angular separations, their conversion into linear motions requires the knowledge of the distances to the objects. In deriving the separation velocities of Table 1 it has been assumed that the quasars are at cosmological distances implied by their redshifts. There are some astronomers who question the validity of this assumption and argue that quasars are in fact considerably closer than implied by Hubble's law (Burbidge 1978; Arp 1983). If this turns out to be, the case then the observed motions are *not* superluminal after all. Indeed, one wonders if this phenomenon had been found in the early days of the discovery of quasars, whether it might have swayed the majority opinion towards the view that quasars are local objects!

Nevertheless the majority of astronomers today would like to assume that quasars are as distant as implied by Hubble's law. In that case there is only one recourse left: to argue that the observed motions are illusory and do not correspond to actual physical motions. In the remaining part of this article we take a stock of theoretical attempts to explain the superluminal motions as illusions, although our main emphasis will be on two scenarios.

2. Relativistic beaming

Even before the discovery of the phenomenon of superluminal motion, some quasars had given indications of fast bulk motions, through the rapid time variability of their luminosities both in radio and optical wavelengths (Burbidge & Burbidge 1967). If τ is the characteristic timescale of variability then special relativity imposes an upper limit

$c\tau$ over the linear dimensions of the object. To get out of this stringent upper limit Rees (1966) proposed an ingenious model. If the quasar is expanding relativistically with speed $v \simeq c$, then a remote observer will see its projected boundary expand at a rate $\gamma v \simeq \gamma c$ where

$$\gamma = \frac{1}{(1 - v^2/c^2)^{1/2}} \,. \tag{1}$$

This effect arises because the light rays reaching the observer at any given time do not all start at the same time from the boundary of the expanding object.

Fig. 1 illustrates this concept when adapted to the superluminal separation of the VLBI components of a quasar. Here, component A is fixed in the rest frame of the observer O, while component B is beamed with speed $v \simeq c$ *almost* along the line of sight OA, towards the observer. In Fig. 1 the angle OAB = θ is supposed to be small so that $|\sin\theta| \ll 1$. Then the projected separation perpendicular to OA will be seen to grow at a rate

$$V = \frac{v \sin\theta}{1 - \frac{v}{c}\cos\theta} \,. \tag{2}$$

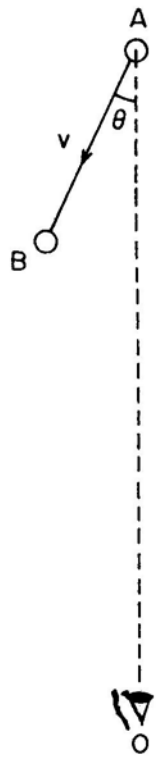


Figure 1. In the beam model schematically shown here A is the stationary component and B is the component beamed at the observer O. For apparent superluminal motion to manifest, the angle θ has to be very small.

This expression reaches a maximum value V_{max} when $\theta = \theta_{max}$, where,

$$\sin\theta_{max} = \gamma^{-1}, \quad V_{max} = \gamma v \simeq \gamma c. \tag{3}$$

Though this conclusion is similar to that for the original Rees model, there is one important difference. In the Rees model the quasar was a spherical object and therefore θ spanned the entire range from 0 to π, with the maximum expansion occurring for its value given by (3). In Fig. 1 we have a linear system and to achieve the large value of γ, θ ($\simeq \sin\theta$) has to be *chosen* to be finely tuned to the value γ^{-1}. We will return to the question of how probable this is, later.

The beam model also predicts that the intensity of blob B is enhanced by a factor

$$f = \left[\gamma\left(1 - \frac{v}{c}\cos\theta\right)\right]^{-(3+\alpha)} \tag{4}$$

due to the Doppler effect, where α is the spectral index of the radiating source (Cohen *et al.* 1971). Thus in cases of interest $\theta \simeq 0$, $v \simeq c$ and $\alpha \simeq 1$, f is expected to be ~ 10.

The beam model starts with the advantage that the underlying idea existed in literature before the phenomenon was discovered. It makes a clever use of the kinematic effect due to narrow, relativistic beaming at the observer. The model received further theoretical support in terms of the twin exhaust model of Blandford & Rees (1974) which provided a scenario for highly collimated beams issuing in opposite direction along the axis of rotation of a massive system such as a galaxy. These jets are supposed to impinge on intergalactic clouds to produce the observed hot spots. It was natural to think of the core beams producing the superluminal motion on the VLBI scale, as part of the largescale phenomenon of the extended jets which produce the hot spots.

The VLA data do show radio jets on extended scales of several tens of kiloparsecs and these findings have led to a general belief in the existence of jets in radio sources, whether on the small scale of a few parsecs seen in VLBI or on the larger scale of the extended radio sources. Against this background the hypothesis of relativistic beaming appears (at first sight) to provide a natural expalnation of the observed superluminal motions in quasars.

With all these attractive features; however, the beam model is not without its difficulties when confronted with certain observational details. For instance, take the formula (4). If both blobs A and B were comparable in luminosity in their respective rest frames, then B should appear brighter (by a factor ~ 10) in the observer's rest frame. Had this turned out to be the case, it would have been a striking demonstration of the beaming effect. In reality the brightness of B is not significantly different from that of A as seen by O. To explain this result one has to assume that B is intrinsically fainter than A. Indeed, the stronger the beaming effect ($\cos\theta \simeq 1$) the larger is this difference in luminosity. By making a comparison with the radio sources which are not beamed at us, Browne *et al.* (1982) placed a lower limit on the beaming angle θ for the various superluminal quasars. They found that these values are not small enough to explain the superluminal effect unless the Hubble constant is increased from 50 km s^{-1} Mpc^{-1} to 100 km s^{-1} Mpc^{-1}. (An increased value of the Hubble constant brings quasars closer and reduces the transverse speed of separation of A and B. This way out of the difficulty is thus closer in spirit to the resolution offered by the local hypothesis of quasars.)

One of the early expectations of the twin exhaust model was that the smallscale (VLBI) structures of radio sources would be aligned with the extended lobes. We will

refer to this as the 'alignment property' in all future discussions here. While initial observations lent some support to the alignment property, discrepancies began to appear in later studies. For example, the VLBI jets in a number of super-luminal sources are found to be inclined at angles of a few tens of degrees to the largescale jets (Readhead *et al.* 1978). More recently, Schilizzi & de Bruyn (1983) have looked at these objects from another angle. If the VLBI and largescale structures were aligned, then, using the angle θ given by Knight *et al.* (1971), one can estimate the linear sizes of the extended structures. Schilizzi & de Bruyn find that these linear sizes are significantly larger than for sources which are not beamed at us, implying that their inclination to the line of sight is considerably larger than θ. Clearly the two structures then cannot be aligned.

To get out of this difficulty it is argued that obstructions due to inhomogeneities cause knots in the relativistic jets (Fomalont 1983). The kinks so caused are magnified by the kinematic effects of $v \simeq c$ and hence the position angles of smallscale jets are found to be different from those of extended jets.

We may mention an added difficulty arising from the breakdown of the alignment property. Most quasars so far seen show only one jet instead of two. In their paper on twin exhaust, Blandford & Rees (1974) had argued that the issue of central plasma as a single jet is basically unstable and that a counterjet must develop. Why do we not see two jets in reality? The explanation offered by the Doppler hypothesis of Narlikar & Subramanian (1983), that ram-pressure of the intergalactic medium stops the exhaust in the forward direction cannot be invoked in the cosmological hypothesis of quasars. So it was argued that there are two jets but we see only the Doppler boosted one beamed at us. However, with a misalignment of the extended jet and the VLBI jets it is not possible to argue that both are being beamed at us.

In the Blandford–Rees model the magnetic field in the jet was supposed to be oriented largely perpendicular to the outflowing relativistic plasma. Recent observations of the optical emission from the extended radio source Coma A by Miley *et al.* (1981) suggests the alignment of the magnetic field along the direction of the jet.

Optical spectroscopy of the beam fluid indicates that the motion is nowhere near relativistic speeds. In fact there is a certain amount of evidence (Heckman *et al.* 1982) that the jets associated with low-luminosity radio galaxies such as 3C 305 have velocities characterizing the bright regions of the emission line gas more like ~ 300 km s^{-1}. By examining the correlation between the optical and radio emission in radio sources in general, Strittmatter (1984) has argued that relativistic beaming is unlikely to be taking place in the extended sources. Scheuer (1983) also has given arguments to show that the jet velocities in both the VLBI and extended sources must have the same characteristic of being either relativistic or nonrelativistic.

Finally, it should be remarked that the existence of intervening beams carrying material from the central source to the outer lobes was utilized in the earlier studies to emphasize the formation of outer lobes as a result of plasma impinging upon the intergalactic medium. In the light of the resolution and dynamic range available at the time the relativistic beams themselves were not expected to be observed which is probably why their observable features were not particularly stressed. The beams were later revealed from the radio maps on the VLBI and the extended scales.

These arguments suggest that although the relativistic beaming hypothesis is the best sell theory today, it is not manifestly the best buy theory. For it to have the latter property it is necessary to examine what alternative hypotheses exist in literature today.

3. Alternative scenarios

There have been a number of ingenious proposals advanced to explain the phenomenon of apparent superluminal motion.

(a) *Christmas-tree model* (Dent 1972) involves independent flares erupting at random in various locations of the source and these could mimic a regular superlight motion. However, it was soon realized (Cohen *et al.* 1977) that the observed motion was highly systematic and only superlight expansions were generally observed.

(b) *Light echo model* (Lynden-Bell 1977) attributes the superluminal motion to an outward-propagating signal like a relativistic blast wave which causes a progressive brightening of the source region of increasingly large size. Such a signal directed in opposite directions along an axis making a small angle with the line of sight can result in a superluminal expansion. Clearly, the model is not compatible with the observed core–jet structure of these sources.

There are a number of other suggestions: dipole field model, synchrotron opacity model, kinematic illusions caused by the finite time of propagating signals. We shall not discuss them here, but refer to the reviews by Marscher & Scott (1980), and Kellermann & Pauliny-Toth (1981).

4. The gravitational screen model

The gravitational screen model (Chitre & Narlikar 1979; Chitre & Narlikar 1980) was proposed by us as an explanation of superluminal separation a few months *before* the discovery of the twin quasar 0957 ± 561 A, B and the consequent popularity of gravitational lens in quasar astronomy. Fig. 2 illustrates schematically how this model operates. A, B are two radio blobs in a source S which is 'screened' by an intervening massive object D which defects the light rays from A and B *en route* to the observer O. As a result of deflection, O sees the virtual images A′, B′ of A, B. While A and B separate from each other at subluminal speed, is it possible for O to see A′, B′ separate superluminally? This is the question we set out to investigate, and the outcome is summarized below. It is worth pointing out at the outset that this model is different from the earlier models of Barnothy (1965; see also Barnothy & Barnothy 1971) or the work of Gott & Gunn (1974) all of which invoke gravitational bending in one form or another.

First we note that a typical gravitational screen like a galaxy causes differential gravitational bending as the impact parameter a of the light ray increases. In a spherical galaxy, $a = 0$ for a ray passing through the centre and the bending angle $\Delta(a) = 0$. As a increases, $\Delta(a)$ rises sharply and then falls slowly outwards. For a galaxy of mass M and radius R, $\Delta(a)$ equals the Einstein value

$$\Delta(r) = \frac{4GM}{c^2 r} \quad \text{for } r \geqslant R. \tag{5}$$

In the central regions of the galaxy $\Delta(a)$ could be significantly higher than $\Delta(R)$. [Both $\Delta(a)$ and $\Delta(R)$ in weak-field general relativity turn out to be twice the respective values in Newtonian gravity.]

Denote by x_D, x_{DS} and x_s the distances between the observer and the deflector, the

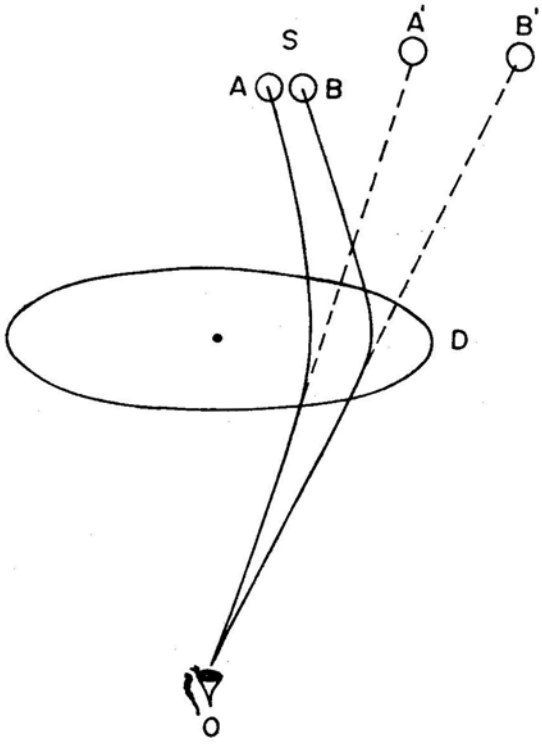


Figure 2. Source S has two components A and B which are separating subluminally. The images A′ and B′ formed by the gravitational deflector D under certain circumstances appear to the observer O to separate at superluminal speeds.

source and the deflector and the observer and the source, respectively. Let $v_\perp$ denote the transverse speed of separation between A and B relative to the direction OS. Then the apparent separation velocity as seen by O is

$$V = \frac{v_\perp}{1 - x_D \frac{x_{DS}}{x_S} \Delta'(a)} . \tag{6}$$

From this we see that a large magnification of velocity is possible if $\Delta'(a) > 0$ and the denominator of (6) almost vanishes, *i.e.,*

$$\Delta'(a) = \frac{1}{x_D} \cdot \frac{x_S}{x_{DS}} . \tag{7}$$

For this condition to be satisfied, the source and the observer should be at conjugate positions with respect to the deflector. How probable is such a situation? Again, we defer the computation of probabilities to the end.

In essence, therefore, apparent superluminal speeds will be seen provided a suitable gravitating mass intervenes between the source and the observer at a suitable intermediate point. Intervening galaxies have been invoked to account for absorption-line redshifts in quasars, and more recently, for explaining very similar closely spaced multiple quasars as gravitationally lensed images. Thus the supposition of an

intervening deflector for producing superluminally separating images is no more or no less implausible against the current backdrop of theoretical ideas. The best 'proof' of the correctness of the screen model lies of course in the detection of an actual screen. We will consider evidence of this kind shortly. First we outline some observable effects in the screen model and compare its performance with the beam model.

(i) A significant feature of lensing is that the amplification produced is not uniform in all directions perpendicular to the line of sight. Thus the image of a straight line or a straight path may appear distorted. These effects will be confined to the VLBI features which are influenced by the magnification formula (6), but will be negligible in the extended features. It is therefore expected that the VLBI jets will be misaligned with respect to the extended features, especially in quasars which show superluminal motion. (It is worth recording that when the screen model was first proposed in 1978, a referee of our paper pointed this property as a *drawback* of the model since it was then believed that the alignment property holds!)

(ii) Since gravitational deflection is independent of wavelength, the superluminal separation is expected to be the same at all wavelengths of observation.

(iii) The apparent velocity of separation in 3C 345 seems to show an increase from 7.5 c to 12.2 c (H_0 = 50 km s^{-1} Mpc^{-1}). This superluminal acceleration appears to be genuine (Moore, Readhead & Baath 1983), but is hard to understand in the simple beaming model. In the screen model, accelerations (and decelerations) of this kind are artifacts of changes in the magnification as the light rays encounter varying density regions with changing Δ ($\mathcal{A}$). This effect is expected to be large when the source and the observer are near conjugate points. Besides, smallscale inhomogenities in the deflector are expected to produce short-term changes in V (of duration ~ 1 yr), thus making the plot of angular separation against time a jagged curve with a linear trend.

(iv) If the optical object coincides with the core where superluminal separation is being observed, the lensing phenomenon will lead to an amplification of the apparent optical luminosity of the quasar. We will discuss an explicit example of this circumstance in the next section. A similar effect is not expected in the beam model.

(v) A testable prediction of the screen model is the likely existence of super-luminal separation in those quasars which are believed to be lensed by intervening galaxies or clusters. For example, the twin quasars 0957 + 561 A, B where the lens system has been detected, should show a magnification of velocity by a factor ~ 2–3. Hence, provided the source components are separating at speed $v \lesssim c$ we should see apparent separation speed $V \lesssim 3c$.

(vi) It should be recognized that in the screen model the probability measure of two images of a single source can be larger than for a single bright image. The double imaging would therefore be expected to occur in many of the super-luminal sources, although none have been detected so far. This may be due to a selection effect which favours a single bright image over a couple of less bright images. It would be worthwhile undertaking high-resolution studies with a view to look for multiple imaging in superluminal sources.

5. The quasar 3C 273

3C 273, the first quasar to be detected has the redshift z = 0.158 and an apparent magnitude 12.8. It is not only abnormally bright optically, but is in fact the brightest

quasar in the sky at optical, X-ray and γ-ray wavelengths. Further, the strength of the emission line region of this quasar also makes it a unique object (Rees 1984). It is perhaps significant for the screen model (but not for the beam model) that the radio component 3C 273B which coincides with the optical object contains the core with superluminally separating components. The extended optical jet in the source is misaligned by about 20° with the direction of the VLBI jet.

Let us consider 3C 273 in both scenarios, that of relativistic beaming and of screening by an intervening galaxy. For H_0 = 50 km s^{-1} Mpc^{-1} we have $V \simeq 10c$. The best, case of Equation (3) requires $\gamma = 10$. However, to avoid the criticism of choosing the 'best' as 'typical' we assume that the beaming in a direction $\theta \neq \theta_{\max}$. A straightforward calculation shows that for $\gamma > 10$, a small range of values of θ around $\theta_{\max}$ can generate $V \geqslant 10\ c$. The probability of beaming in this range is given by

$$P(\gamma) = \frac{1}{1 + V^2/c^2}\left(\frac{\gamma^2 - 1 - V^2/c^2}{\gamma^2 - 1}\right)^{1/2}. \tag{8}$$

For the limit $\gamma = \infty$, $P = \theta^2_{\max} \simeq 10^{-2}$. However, $\gamma = \infty$ gives clearly an upper limit on P. For γ = 10.1, the value is ~ 10^{-3}. However, we should multiply this value by a further rarity factor ~ 10^{-2} to include the abnormal brightness of 3C 273 in the optical and X-rays as well its strong emission line region (Rees 1984). Thus the probability comes down to $P \sim 10^{-5}$. [The brightness of the emission-line region could not be explained as a beaming effect since there is no evidence for a relativistic motion of the emission-line gas (Kellermann & Pauliny-Toth 1981; Heckman *et al.* 1982).]

We shall now demonstrate that the exceptional features of 3C 273 can be accounted for in a natural manner by postulating a gravitational screen at an intermediate distance along the line of sight to the quasar.

First we note that studies of the nebulosity around 3C 273 by Tyson, Baum & Kreidl (1982) furnish possible support for the screen hypothesis in the following way. These authors have argued that the isophotal distribution of the nebulosity resembles that of the brightest elliptical galaxy in a cluster. Earlier, Wycoff *et al.* (1980) had found that the redshift of the galaxy is very nearly the same as the redshift of the quasar, thus implying that the quasar is being hosted in the nuclear region of the galaxy. There are two anomalous features, however, which are found from the work of Tyson, Baum & Kreidl (1982). At apparent magnitude of 16 the supposed galaxy is considerably brighter than other galaxies similarly hosting quasars in their nuclei (W. A. Baum, personal communication). Further, the quasar is not exactly at the centre of the galaxy but is off-set by about 1 arcsec to the east of it. Both these anomalies could be explained if there were an intervening galaxy which is fainter than the nebulosity by ~ 2–3 magnitudes. We show that such a galaxy can also explain the abnormal brightness of the optical (and X-ray) object in 3C 273 as well as the superluminal motion ~ 10c.

A typical solution with a spherical lens galaxy located at a redshift of 0.07 and core-radius r_c = l kpc has a mass ~10^{11} $M_\odot$, line-of-sight velocity dispersion σ ~ 187 km s^{-1} and yields a linear magnification ~ 10 and an intensity magnification ~ 22. The details of various screen models are given elsewhere (Chitre *et al.* 1984). We may add that these solutions are also able to explain the misalignment ~ 20° between the extended jet and the VLBI jet which alone is subjected to the dominant gravitational influence of the screen. Furthermore, the models are by no means unique, but typical ones. It is possible to generate other sets of values for the screen at different

intermediate redshifts with a scaling that holds the parameter

$$\mu = \frac{4GM x_D x_{DS}}{c^2 r_c^2 x_S} \tag{9}$$

fixed. (r_c = radius of the core producing the bending.)

Finally we compare the probability of such a scenario with that computed earlier for the beam model. For this purpose we adopt the luminosity function $\phi(L)$ for galaxies given by Schechter (1976) and assume the Faber–Jackson (1976) relation between the luminosity L and the line of sight velocity dispersion σ. The probability of finding intervening galaxies with a velocity dispersion in the range (σ, $\sigma' + d\sigma$) and permitted redshift range (z_{min}, z_{max}) in the $q_0 = 0$ Friedmann universe is given by

$$dP = \frac{1}{2}\left(\frac{c}{H_0}\right)\left[(1+z_{max})^2 - (1+z_{min})^2\right] A^{3/2} f(\Delta\theta)\Sigma(\sigma)\phi(L/\sigma)d\sigma. \tag{10}$$

Here $\phi(L/\sigma)$ is the Schechter function convolved with the Faber–Jackson relation, $\Sigma(\sigma)$ is the cross section for bending required to produce the large superluminal speed, $F(\Delta\theta)$ is the probability that the position angle of the major axis of the lens galaxy yields the desired configuration and A is the flux amplification factor. The net probability so computed comes out $\gtrsim 6.3 \times 10^{-5}$. The $>$ sign allows for deflectors which are not in galactic forms. If the existence of unseen mass in the universe is confirmed, then it is not unreasonable to suppose either that the bending masses used in models are underestimates or that faint objects with large mass-to-light ratio exist in the universe over and above the galaxies used in $\phi(L)$.

We mention in passing that the optical object associated with 3C120 (another VLBI superluminal case) also presents another likely case of superposition of the line-of-sight screen galaxy with the galaxy hosting the radio source. The optical studies of Arp (1975) and Baldwin *et al.* (1980) suggest a two-component system, one with the axis of symmetric gas motion aligned with the radio axis and the other made of star distributions whose isophotal ellipses have major and minor axes pointing in altogether different directions. Could the latter system be a screen for the former?

6. Conclusion

As the VLBI techniques continue to improve in the future we may very well encounter further cases of superluminal motions in quasars. While the beaming model has several attractive features it has difficulties also, and it may be worthwhile having more than one theoretical basket to store our eggs of speculation. We believe that the screen model which offers yet another striking application of gravitational lensing phenomenon can be one of these additional baskets.

Acknowledgements

We have enjoyed discussing this subject with Martin Rees, Richard Porcas, Bill Saslaw and Geoffrey Burbidge. We thank Kandaswamy Subramanian and D. Narasimha for

help in computing the gravitational lens models and for numerous discussions. We would also like to express our thanks to R. Nityananda for his critical comments.

References

Arp, H. C. 1975, *Publ. astr .Soc. Pacific,* **87,** 545.
Arp, H. C. 1983, in *Liege Coll.* 24: *Quasars and Gravitational Lenses,* Ed. J.-P. Swings, Univ. Liege.
Baldwin, J. A., Carswell, R. F., Wampler, E. J., Smith, H. E., Burbidge, E. M., Boksenberg, A. 1980, *Astrophys. J.,* **236,** 388
Barnothy,J. M. 1965, *Astr. J.,* **70,** 666.
Barnothy, J. M., Barnothy, M. F. 1971, *Bull. Am. astr. Soc.,* **3,** 472.
Blandford, R. D., Rees, M. J. 1974, *Mon. Not. R. astr. Soc.,* **169,** 395.
Browne, I. W. A., Clark, R. R., Moore, P. K., Muxlow, T. W. B., Wilkinson, P. N., Cohen, M. H., Porcas, R. W. 1982, *Nature,* **299,** 788.
Burbidge, G. 1978, *Phys. Scripta,* **17,** 281.
Burbidge, G., Burbidge, E. (Eds) 1967, *Quasistellar Objects,* Freeman, San Francisco.
Chitre, S. M., Narlikar, J. V. 1979, *Mon. Not. R. astr. Soc.,* **187,** 655.
Chitre, S. M., Narlikar, J. V. 1980, *Astrophys. J.,* **235,** 335.
Chitre, S. M., Narlikar, J. V., Narasimha, D., Subramanian, **K.** 1984, *Astr. Astrophys.* (in press).
Cohen, M. H., Unwin, S. C. 1984, in *IAU Symp.* 110*: VLBl and Compact Radio Sources,* Eds R. Fanti, K. Kellermann & G. Setti, D. Reidel, Dordrecht, p. 95.
Cohen, M. H., Cannon, W., Purcell, G. H., Shaffer, D. B., Broderick, J. J., Kellermann, K. I., Jauncey, D. L. 1971, *Astrophys. J.,* **170,** 207.
Cohen, M. H. *et al.* 1977, *Nature,* **268,** 405.
Dent, W. A. 1972, *Science,* **175,** 1105.
Faber, S. M., Jackson, R. E. 1976, *Astrophys. J.,* **204,** 668.
Fanti, R., Kellermann, K., Setti, G. (Eds) 1984, *IAU Symp.* 110*: VLBI and Compact Radio Sources,* D. Reidel, Dordrecht.
Fomalont, E. B. 1983, in *Astrophysical Jets,* Eds A. Ferrari & A. G. Pacholczyk, D. Reidel, Dordrecht, p. 37.
Gott, J. R., Gunn, J. E. 1974, *Astrophys. J.,* **190,** L105.
Gubbay, J., Legg, A. J., Robertson, D. S., Moffet, A. T., Ekers, R. D., Seidel, B. 1969, *Nature,* **224,** 1094.
Heckmann, T. M., Miley, G. K., Balick, B., van Breugel, W. J. M., Butcher, H. R. 1982, *Astrophys. J.,* **262,** 529.
Kellermann, K. I., Pauliny-Toth, I. I. K. 1981, *A. Rev. Astr. Astrophys.,* **19,** 373.
Knight, C. A. *et al.* 1971, *Science,* **172,** 52.
Lynden-Bell, D. 1977, *Nature,* **270,** 396.
Marscher, A. P., Scott, J. S. 1980, *Publ .astr. Soc. Pacific,* **92,** 127.
Miley, G. K., Heckman, T. M., Butcher, H. R., van Breugel, W. J. M. 1981, *Astrophys. J.,* **247, L5.**
Moffet, A. T., Gubbay, J., Robertson, D. S., Legg, A. J. 1972, in *IAU Symp.* 44*: External Galaxies and Quasi-stellar Objects,* Ed. D. S. Evans, D. Reidel, Dordrecht, p. 228.
Moore, R. L., Readhead, A. C. S., Baath, L. 1983, *Nature,* **306,** 44.
Narlikar, J. V., Subramanian, K. S. 1983, *Astrophys. J.,* **273,** 44.
Readhead, A. C. S., Cohen, M. H., Pearson, T. J., Wilkinson, P. N. 1978, *Nature,* **276,** 768.
Rees, M. J. 1966, *Nature,* **211,** 468.
Rees, M. J. 1984, in *Extragalactic Energetic Sources,* Ed. V. K. Kapahi, Indian Academy of Sciences, Bangalore (in press).
Schechter, P. 1976, *Astrophys. J.,* **203,** 297.
Scheuer, P.A. G.1983, in *Liege Coll.* 24: *Quasars and Gravitational Lenses,* Ed. J. P. Swings, Univ. Liege.
Schilizzi, R. T., de Bruyn, A. G. 1983, *Nature,* **303,** 26.
Strittmatter, P. 1984, in *Extragalactic Energetic Sources,* Ed. V. K. Kapahi, Indian Academy of Sciences, Bangalore (in press).

Tyson, J. A., Baum, W. A., Kreidl, T. 1982, *Astrophys. J.*, **257,** L1.
Whitney, A. R., Shapiro, I. I, Rogers, A. E. E., Robertson, D. S., Knight, C. A., Clark, T. A., Goldstein, R. M., Marrandino, G. E., Vandenberg, N. R. 1971, *Science*, **173,** 225.
Wyckoff, S., Wehinger, P. A., Gehren, T., Morton, D. G, Boksenberg, A., Albrecht, **R.** 1980, *Astrophys. J.*, **242,** L59.