Binary Stars

Steven N. Shore
Indiana University, South Bend

GLOSSARY

Accretion disk Structure formed when the material accreting onto a star has excess angular momentum, forming a circulating disk of material supported by internal pressure and heated by turbulent stresses.

Lagrangian points Stable points in the orbit of a third body in a binary system; the inner Langrangian point, L_1, lies along the line of centers and marks the Roche limit for a tidally distorted star.

Main sequence Phase of hydrogen core burning; first stable stage of nuclear processing and longest epoch in the evolution of a star.

Mass function Method by which the mass of an unseen companion in a spectroscopic binary can be estimated using the radial velocity of the visible star and the orbital period of the binary.

Orbital parameters Inclination, i, of the orbital plane to the line of sight; P, the period of revolution; e, the eccentricity of the orbit; q, the mass ratio.

Red giant Stage of helium core burning; subsequent to the subgiant stage.

Subgiant Stage of hydrogen shell burning, when the deep envelope initiates nuclear processing around an inert helium core produced by main sequence hydrogen core fusion. This is the transition stage in the expansion of the envelope from the main sequence to the red giant phase.

Units Solar mass ($M_\odot$), 2×10^{33} g; solar radius ($R_\odot$), 7×10^{10} cm.

BINARY STARS are gravitationally bound stars, formed simultaneously or by capture, that orbit a common center of mass and evolve at the same time. These stars are formed with similar initial conditions, although often quite different masses. Visual binaries are both sufficiently separated in space and sufficiently near that their angular motion can be directly observed. Spectroscopic binaries are unresolved systems for which motion is detected using

radial velocity variations of spectral lines from the component stars. If the orbital plane is inclined at nearly 90° to the line of sight, the components will display mutual eclipses. Depending on the orbital period and mass, the stars may share a common envelope (contact), be in a state of unidirectional mass transfer between the components (semidetached), or evolve without mass transfer but with mutual gravitational perturbation (detached). In semidetached systems, depending on the rate of mass transfer and the nature of the accreting object, a hot accretion disk will be formed. If the companion star is gravitationally collapsed, being a neutron star or black hole, X-ray emission will result. Accretion onto white dwarf or neutron stars results in flash nuclear reactions that can trigger the nova event.

I. HISTORICAL INTRODUCTION

At the close of the eighteenth century, William Herschel argued that the frequency of close visual pairs was larger in any area of the sky than would be expected by chance. On this basis, it was suggested that binary stars—that is, physically gravitationally bound stellar systems—must exist. Prior to the discovery of Neptune, this was the most dramatic available demonstration of the universality of Newton's theory of gravitation. Herschel, Boole, and others extended this study to the demonstration that clustering is a general phenomenon of gravitational systems. The discovery of the wobble in the proper motion of Sirius led Bessel, in the 1840s, to argue for the presence of a low-mass, then unseen companion; it was discovered about 20 years later by Clarke. Routine observations of visual binary star orbits were being made by the end of the century. For very low mass stars, the method is still employed in the search for planets through proper motion perturbations, much like the method by which Neptune was discovered, although velocity variations have now supplanted this search strategy.

About the same time as Herschel's original work, Goodricke and Piggott observed the photomeric variations in the star β Persei, also known as Algol. The variations, he argued, were due to the system being an unresolved, short-period binary system, with one star considerably brighter than the other but otherwise about the same physical size. They postulated that the light variations were consistent with eclipses, and that were we able to resolve the system, we would see two stars orbiting a common center of mass with about a three-day period. The dramatic confirmation of this picture came with the discovery, by Hartmann and Vogel at the end of the last century, of radial velocity variations in this system consistent with the eclipse phenomenology. By mid-century, in part as a result of the work at Bamberg under Argelander, large-scale searches for variable stars began to produce very large samples of stars with β Persei-like behavior. By the mid-1920s, much of the theory of geometric eclipses had been developed. Russell and Shapley, in particular, included the effects of reflection (scattering) and ellipsoidal (tidal) distortions.

Most theoretical work on binary stars is the product of the past 70 years. Methods for the analysis of eclipses, based on light curve fitting by spheroids, were developed by Russell and Merrill in the first two decades of this century. Atmospheric eclipses were first discussed by Kopal in the 1930s. The study of mass transfer in binary systems was initiated largely by Struve in the mid-1930s, and the applications of orbital dynamics to the study of mass transfer began in the 1940s with Kuiper's study of β Lyrae. Using large-scale computer models, stellar evolution in binary star systems was first studied in detail in the 1960s by Paczynski and collaborators in Warsaw, Plavec and colleagues in Prague. Hydrodynamic modeling has only recently been possible using realistic physics and remains a most interesting area of study. Much recent work on binary star evolution and hydrodynamics has been spurred by the study of binary X-ray sources. Following the discovery of binarity for several classical novas, by Walker in the 1950s, the connection between low-mass X-ray binaries and cataclysmics has been central to the study of evolution of stars undergoing mass exchange.

II. SOME PRELIMINARIES

The broadest separation between types of binary stars is on the basis of observing method. For widely separated systems, which are also close to us, the stars appear physically distinct on the sky. If the orbital periods are short enough (that is, less than millennia), it is possible to determine the plane of the projected orbit by observing directly the motion of the stars. For those systems in which the luminosity ratios (and possibly mass ratios) are large, it is possible to obtain orbital characteristics for the two members by observing periodic wobbling in the proper motion of the visible member. Such methods are frequently employed in the search for planetary-like companions to high proper motion stars (that is, stars with large transverse velocities to the line of sight, such as Barnard's star). For systems of low proper motion and possibly long period, speckle interferometry, intensity, and Michelson interferometry, as well as lunar occultations, can be useful in separating components and at least determining luminosity ratios. Such methods, extended to the near infrared, have been recently employed in the search for brown dwarf stars, objects of Jupiter-sized mass.

When the system is unresolved, even at separations of several milliarcseconds, it is necessary to employ spectroscopic methods to determine the composition and motion of the constituent stars. These are the spectroscopic binaries, by far the largest group so far studied. Two methods of analysis, which are happily sometimes complementary, can be used—observation of radial velocity variations of the components and eclipse phenomena.

A. Spectroscopic Binary Velocity Curves

Consider two stars in a circular orbit about the center of mass. Regardless of the perturbations, we can say that the separation of the two stars of mass M_1 and M_2 is $a = r_1 + r_2$ in terms of separation of the stars from the center of mass. The individual radii are the distances of the components from the center of mass:

$$r_1/r_2 = M_2/M_1 \tag{1}$$

The velocity ratio is given by $V_2/V_1 = M_1/M_2$ for a circular orbit, where V is the orbital velocity. By Kepler's law,

$$GM = \omega^2 a^3 \tag{2}$$

where ω is the orbital frequency, given by $2\pi/P$ where P is the period, and M is the total mass of the system. Now, assume that the inclination of the plane of the orbit to the observer is i, that the maximum *observed* radial velocity for one star is given by K, and that we observe only one of the stars. Then,

$$K^3 P/2\pi G = M_2^3 \sin^3 i \,/ M^2 = f(M) \tag{3}$$

The function $f(M)$ is called the *mass function* and depends only on observable parameters, the maximum radial velocity of one of the stars, and the period of the orbit. If M_1 is the mass of the visible star, $f(M)$ serves to delimit the mass of the unseen companion. If both stars are visible, then,

$$K_1/K_2 = M_2/M_1 \tag{4}$$

independent of the inclination. Thus, if both stars can be observed, both the mass ratio and the individual masses can be specified to within an uncertainty in the orbital inclination using $f(M)$. The mass function permits a direct determination of stellar masses, independent of the distance to the stars. This means that we can obtain one of the most important physical parameters for understanding stellar evolution merely by a kinematic study of the stars.

If the orbit is eccentric, departures from simple sinusoidal radial velocity variations with time are seen. The eccentricity of the orbit also introduces another symmetry-breaking factor into the velocity variations, because the angle between the observer and the line of apsides, the line that marks the major axis of each ellipse, determines the shapes of the velocity variation curve with orbital phase.

B. Eclipsing Binary Light Curves

The orbital plane of eclipsing stars lies perpendicular to the plane of the sky. Depending on the relative sizes of the stars, the orbital inclination over which eclipses can occur is considerable, but in general only a small fraction of the known binary systems will be seen to undergo eclipses. The variation in light serves two purposes. It permits a determination of the relative radii of the stars, since the duration of ingress and the duration of eclipse depend on the difference in the sizes of the stars. That is, if Δt_1 is the total time between first and last contact, and Δt_2 is the duration of totality, assuming that the eclipse is annular or total, then,

$$\frac{\Delta t_1}{\Delta t_2} = \frac{r_g + r_s}{r_g - r_s} \tag{5}$$

where r_s is the smaller and r_g is the greater radius, respectively. The diminution in light from the system depends on the relative surface brightness of the stars, which in turn depends on the surface (effective) temperature, T_{eff}. Eclipses will not be total if the two stars are not precisely in the line of sight, unless they differ considerably in radius, so that the mark of totality is that the light does not vary during the minimum in brightness.

C. Distortions in Photometric and Velocity Curves

Several effects have been noted that distort the light curve and can be used to determine more physical information about the constituent stars.

1. Reflection Effect: External Illumination

Light from one component of a close binary can be scattered from the photosphere and outer atmosphere of the other, producing a single sinusoidal variation in the system brightness outside of eclipse. This *reflection effect* is useful in checking properties of the atmospheres of the stars. If the illuminating star is significantly higher temperature, it can also produce a local heating, which alters the atmospheric conditions of the illuminated star. Such an effect is especially well seen in X-ray sources, particularly HZ Her = Her X-1, which varies from an F-star to an A-star spectrum as one moves from the unilluminated side to the substellar point facing the X-ray source.

2. Photospheric Nonuniformities: Starspots

A similar phenomenon has been noted in the RS CVn stars, where it is caused by the presence of large-scale, active magnetic regions, called *starsports*, on the stellar surfaces. Unlike reflection, these dark regions migrate with time through the light curve as the active regions move with the differential rotation of the stellar envelope, analogously to the motion of sunspots. Chemically peculiar magnetic stars also show departures from simple eclipse profiles, because of locally cooler photospheric patches, but these appear to be stable in placement on the stellar surface.

3. Circumstellar Material

The presence of disks or other circumstellar matter also distorts the light curves and can alter the radial velocity variations as well. In Algol systems, this is especially important. The timing of eclipses indicates a circular orbit, while the radial velocity variations are more like that of a highly eccentric one. The explanation lies in the fact that here is considerable optical depth in the matter in the orbit, which results in the atomic lines producing a distortion in the radial velocity variations. Many of the W Serpentis stars show this effect. It is most noticeable in eclipsing systems because these present the largest path length through material in the orbital plane. In some cases, atmospheric eclipses can also distort the lines because of stellar winds and convection cells intercepted by the line of sight. These motions, however, are generally small compared with the radial velocity and so alter the photometry (light-curve instabilities during eclipse are well marked in the ζ Aur stars) but do not seriously affect the radial velocity determinations.

4. Ellipsoidal Distortions: Tidal Interaction

If the stars are close enough together, their mutual gravitational influences raise tides in the envelopes, distorting the photospheres and producing a double sinusoidal continuous light variation outside of eclipse. Many of these systems also suffer from reflection-effect distortion, so there are many equivalent periods in these systems, depending on whether or not they eclipse.

Departures from symmetric minima should accompany expansion of the stars within their tidal surfaces. As the photosphere comes closer to the tidal-limiting radius, the *Roche limit*, the star becomes progressively more distorted from a symmetric ellipsoid and the photometric variations become more like sinc curves. An additional feature is that as the stars become larger relative to the Roche limit they subtend a greater solid angle and display increasing reflection effect from the companion.

5. Limb Darkening

Stellar surfaces are not solid, and they have a continuous variation in surface brightness as one nears the limb. This effect, called *limb darkening*, is produced by the temperature gradient of the outer stellar atmosphere compared with the photospheres. The effect of limb darkening on light curves is to produce a departure from the behavior of simple, uniform spheres most notable in the softening of the points of contact during eclipse. It is one of the best ways available for measuring the temperature gradients of stellar atmospheres.

6. Apsidal Motion: Orbital Precession

The additional effect of the tidal distortion is that the stars are no longer simple point sources, but produce a perturbation on the mutual gravitational attraction. The departure of the gravitational potential from that of two point masses produces *apsidal motion*, the slow precession of the line connecting the two stars. This rate depends on the degree of distortion of the two stars, which in turn provides a measure of the degree of central concentration of the stars. Such information is an important input for stellar evolutionary models. One system that has been especially well studied is α Virginis (Spica). An additional source of apsidal motion is the emission of angular momentum from the system, and the presence of a third body.

7. Third Light

Either because of circumstellar material in the orbital plane, which is not eclipsed but which scatters light from the binary components, or because of the presence of a faint third body in the system that is unresolved, some additional light may be present at a constant level in the eclipsing binary light curve. Frequently, high-resolution spectroscopy is able to reveal the lines of the companion star, as in Algol, but often it remains a problem to figure out the source for the nonphotospheric contributions to the light curve. This is simply added as an offset in the determinations of eclipse properties in most methods of light-curve analysis. Such emission may also arise from shocks in accretion disks and from intrinsic disk self-luminosity.

III. CLASSIFICATION OF CLOSE BINARY SYSTEMS

There are several distinctive classes of binary stars, distinguished nominally by their prototypes, usually the first observed or best known example of the phenomenology. In several cases, however, overlaps in the properties of the

various systems make the prototypical separation confusing and less useful. Two main features distinguish classes of stars: the masses of the components and the separations. Alternatively, the period of the binary and the evolutionary status of the components are useful, and we shall use these alternately as needed.

The broad distinction among various binaries is whether the stars are physically separated by sufficient distance that the photospheres are distinct, in which case they are called *detached*, or have been significantly tidally distorted and may be in the process of mass transfer in some form. This latter class divides into those that have only one star transferring mass, the *semidetached*, and those with both stars immersed in a common envelope of gas that is mutually exchanged, the *contact* systems. This classification, first developed by Kuiper, has proven to be a most general taxonomic tool for distinguishing the various physical processes that occur at different stages in the evolution of close binaries. It is most important to note that a binary, depending on its initial period, mass, and mass ratio may pass through any or all of these stages at some time in its life. This is due to the expansion of stellar envelopes as stars evolve.

A. The Roche Surface

The key element in binary star evolution is the role of the Roche limit, which was first introduced in the three-body problem. It is known in the celestial mechanics literature as a *zero-velocity surface*, but we will treat it as the bounding equipotential for a self-gravitating star:

$$\Phi(r) = -\frac{GM_1}{r_1} - \frac{GM_2}{r_2} - \frac{1}{2}r^2\omega^2 \qquad (6)$$

where $r_i = |\mathbf{r} - \mathbf{x}_i|$ for masses located at positions $x_1 = M_2 a/M$ and $x_2 = M_1 a/M$ for a circular orbit with the stellar separation a. This is the potential in the corotating frame with frequency ω. There are five equilibrium points where the gradient of Φ vanishes, three of which are along the line of centers. Two are peripheral to the masses and lie as the critical points along equipotentials that envelope both stars. These are saddle points for which particle trajectories are unconditionally unstable. Two other points, L_4 and L_5, lie diametrically opposite each other perpendicular to the line of centers. These are quasi-equilibrium points for which local orbits are possible because of the Coriolis acceleration. The *Roche lobe* is the equipotential that passes through L_1, called the *inner* Lagrangian point, that lies between the masses along the line of centers. Mass loss is inevitable for the star that is contacting its Roche lobe. As a star's radius approaches the Roche surface, the body becomes more distorted and eventually, when it contacts L_1, the inner Lagrangian point, it loses mass to the companion and possibly from the system. This actually occurs before the photosphere reaches R_{RL} in the absence of magnetic fields or other constraints on the flow. Several analytic approximations have been derived for the radius of the equivalent sphere whose volume equals that of the appropriate lobe. In general, the Roche radius depends on the mass of the components and q through $R_{RL} = f(q)a$, where $f(q)$ is provided by functional approximate fits to the exact calculations. Two compact, though restricted, approximations are

$$\begin{aligned} f(q) &= 0.2 + 0.38 \log q \qquad (0.5 \le q \le 20) \\ f(q) &= 0.462\left(\frac{q}{1+q}\right)^{1/3} \qquad (0 \le q \le 0.5) \end{aligned} \qquad (7)$$

Binary systems are theoretically distinguished by the sizes of the components relative to their respective Roche lobes, although a system in its lifetime may pass through several or none of these, depending on the separation of the stars and their masses. Stars that are in mutual contact with, or overflowing, their critical equipotential surfaces are called *common envelope* or *contact systems*. When only one star's radius equals its Roche radius, the system is called *semidetached*. Finally, if both stars are separate, however much they may be distorted, the binary is *detached*. Geometrical methods for treating light curves have been developed based on this idea, each treating the shape of the star and the distribution of temperature over the surface more or less phenomenologically (by scaling model atmospheres to the local conditions on tidal ellipsoids or Roche surfaces and piecing together the surface). The tidal distortion is also important for the internal structure of the stars and must be handled in a more detailed way than the axisymmetric case, but similar problems arise nonetheless.

The Roche surface is the star's response to the tide raised by its companion. There are two sorts of tides. One is the *equilibrium tidal surface* that is instantaneously in hydrostatic equilibrium everywhere in the star. The baroclinicity of the surfaces, as in the rotationally distorted case, induces slow circulation that ultimately redistributes angular momentum as well as energy. This produces circularization by loss of orbital angular momentum through viscosity and is most efficient for stars with deep convective envelopes because the turbulence acts like an effective viscosity. The other is a dynamical tide that acts like a nonradial pulsation of the envelope and produces faster internal flows, internal mixing, and redistribution of angular momentum. In the Earth–Moon system, the Moon rotates with the same period as its orbit—that is, synchronously—while the Earth rotates more rapidly. Any point on Earth must then experience a periodic tidal acceleration since, in the

rotating frame, any locale on the planet is carried through the alternating extrema of the perturbing force. This slowly alters the lunar orbit. The solid Earth is not distorted sufficiently to dissipate its rotational energy, hence the rotation only slowly approaches the lunar orbital period through tidal friction. The physical mechanism must be something like this. Stars are fluid and fluids yield to shear. Hence, the induced tidal distortion produces flows because the gravitational potential develops along equipotential surfaces. These flows transport momentum in the rotating frame, leading toward solid body rotation and synchronism depending on the internal viscosity, the precise nature of which is currently not known.

B. Evolutionary Stages of Stars in Close Binaries

On reaching the Roche surface, the stellar envelope is presumed to become unbound, and mass loss or transfer is initiated on a hydrodynamic timescale. The star is generically referred to as the *loser* or *donor*, terms usually applied in the case of mass transfer between the binary components. The companion is called the *gainer*, which implies some amount of accretion. The alternative—to call one star the *primary* and the other the *secondary*—is based on the relative contributions to the combined spectroscopic and/or photometric properties and does not capture the physical nature of the interaction. For detached systems on the main sequence, these terms also describe the respective masses but that correspondence breaks down once the stars begin their ascent of the H–R diagram.

The taxonomic distinctions for the different evolutionary cases are based on the stage at which the nomenclature applies. *Case A* occurs before the terminal main sequence, when the loser is still undergoing core hydrogen burning. This is a slow nuclear stage, and not very sensitive to the stellar mass. For mass transfer to occur requires very small orbital separation because of the small stellar radii and the time scale for stellar expansion is very slow. *Case B* occurs after hydrogen core exhaustion and during a relatively rapid stage of radial expansion, either in the traverse across the Hertzsprung gap (hydrogen shell burning) or on first ascent of the giant branch but before helium core ignition. *Case C* is a late stage, when the star has developed a helium core and is on or near the giant branch or asymptotic giant branch (helium shell burning). Two mass-transfer cases are distinguished, as well. In *conservative* mass transfer, the process can be studied in a straightforward way because we neglect any net mass or angular momentum losses. The Roche surface would recede into the loser were it not for the increase in the separation between the components that results from the change in the mass ratio. The loser maintains contact with R_{RL} until contact is broken by expansion of the system and thereafter the more evolved star continues as if it were a single star. In the event of mass loss from the gainer, the process of mass transfer may be reinitiated at some later stage, but this is unlikely. Dropping these assumptions of constant binary mass and total angular momentum makes the scenario more realistic but leads to a dizzying range of phenomenological models, each marked by the adoption of specific mechanisms for breaking constancy of one or both of these quantities.

1. Common Envelope Evolution

Since the Roche surface represents the limit of a set of bounding equipotentials, it is a surface along which flows can occur but that a star can maintain as an equilibrium shape. A particularly important state is encountered by the binary if the radius of the outer layers of both stars exceeds R_{RL}, since matter does not have to catastrophically flow toward either one from the other component. Instead, if both stars are in contact with L_1, a low-speed circulation can, in principle, be established because of the pressure reaction from the companion's outer layers. The resulting optically thick common envelope should, for contact systems, behave as if the layer sits on top of a very strange equipotential surface, one where the surface gravity depends on both colatitude and colongitude. The outer bounding surface through which some mass is certainly lost from the binary passes through either L_2 and L_3, but this is small compared with the flow that must pass through L_1 to maintain thermal balance. This is the observed situation in the W UMa stars. Marked by continuous light variations, these stars appear to be surrounded by a common atmosphere, but they have otherwise stable cores that place them on the main sequence. Observationally, although the stars have different masses ($q \approx 0.5$ and luminosities up to a factor of 10), their envelopes have nearly uniform temperature requiring very efficient heat transport between the components while leaving the stellar interior otherwise unaffected. The coolest W UMa stars must have common *convective* envelopes, where the temperature gradient adjusts to the large variation in the surface gravity due to the angular gradients in the equipotential. How this happens has been a controversial question for decades and remains an important unsolved problem in stellar hydrodynamics. The current consensus is that the envelopes are never precisely in thermal equilibrium and that the mass transfer fluctuates between the components. One way to picture this is that the mass-transfer rate exceeds the thermal time-scale for the entire envelope which drives thermal oscillations that periodically overfill the Roche lobe of the gainer, or rather the star that is the accreter at that moment.

Main sequence contact systems are only one example of evolutionary stages where common envelopes occur. Any circumstellar matter that completely engulfs the companion and is optically thick is, in effect, a common envelope. Cataclysmic binaries are extremely close systems with periods of less than 1 day and at least one degenerate component. Their origin is linked in current hypotheses to a relatively late stage in the evolution of one of the more massive components, since there are no main sequence progenitors with these characteristics. There are two obvious ways to form a white dwarf. One is to invoke magic and drive the envelope off during planetary nebula formation in the post-AGB phase. The other occurs in a common envelope. If one star engulfs the other as it evolves, and this can happen for virtually any system with orbital periods of less than a few weeks on the main sequence, the lower mass companion will find itself orbiting within a dense circumstellar environment produced by the more massive star. Differential motion leads to heating, which scales as $v_{orb}^3 a^{-1} \sim a^{-5/2}$, *and* transfer of angular momentum between the lower mass component and the engulfing envelope. Consequently, if the heating is sufficient, the more massive star is stripped of its envelope with the resulting loss of binding energy, and the companion spirals inward. The result is a white dwarf with a much less evolved companion in a very short period system.

C. Some Prototype Subclasses of Binary Systems

In this section, we summarize some general properties of important subclasses of close binary stars. Several of these have been discussed previously as well, in order to place them in a more physical context.

1. W Ursa Majoris Systems

These are main sequence contact binaries. They are typically low mass, of order $1 - 2M_\odot$, with orbital periods from about 2 hr to 1 day. The envelopes are distinguished as being in either radiative or convective equilibrium. The chief observational characteristics are that they show continuously variable light curves and line profile variations indicative of uniform temperature and surface brightness over both stars, although the mass ratio ranges from 0.1 to 1. Surface temperatures range from about 5500 to 8000 K. The lower mass systems are called *W type*, the higher mass systems are *A type*; the W-type envelopes are convective. Typical of this class are W UMa, TX Cnc, and DK Cyg. There may be massive analogs of this class, although the light curves are more difficult to interpret.

2. RS Canes Venaticorum Stars and Active Binaries

Close binaries with periods less than 2 weeks induce synchronous rotation via tidal coupling on time scales of order 10^8—10^9 yrs. For main sequence stars, this generally results in slow rotation compared with that observed in single stars; for evolved stars, the opposite holds. The RS CVn and related stars show rapid rotation of a cool evolved star that displays enhanced dynamo activity. These stars are marked by exceptionally strong chromospheres and coronas, sometimes having ultraviolet (UV) and X-ray fluxes greater than 10^3 times that observed in normal G–K giants (cool giants). The photometric behavior of these systems is marked by the appearance of a dark wave (large active regions) which migrates through the light curve toward a decreasing phase, suggestive of differential rotation. The active stars have deep convective envelopes. Several subgiant systems, notably V471 Tau, have white dwarf companions, although most systems consist of detached subgiant or giant primaries and main sequence secondaries. With the exception of HR 5110, most of these systems are detached. Other representative members of this class are AR Lac, Z Her, and WW Dra. Typically, the mass ratios are very near unity, although this may be a selection effect.

An analog class, the FK Com stars, shows many of the RS CVn characteristics, especially enhanced chromospheric and coronal activity and rapid rotation, but it appears to be a class of single stars. The FK Com stars are argued to have resulted from the common envelope phase of an evolved system leading to accretion of the companion.

Both of these subclasses are especially notable as radio sources, often displaying long-time-duration (days to weeks) flares with energy releases of some 10^7 times that of the largest solar flares. The dMe stars are the low-mass analogs of these systems, although not all of these are binaries. It appears that the binarity is most critical in producing more rapid than normal rotation, which is responsible for the enhanced dynamo activity.

3. z Aurigae Stars and Atmospheric Eclipses

These systems consist of hot, main sequence stars, typically spectral type B, and highly evolved giants or supergiants with low surface temperatures (G or K giants). For several eclipsing systems, notably ζ Aur, 31 Cyg, and 32 Cyg, eclipses of the hot star can be observed through the giant atmosphere. The B star thus acts like a probe through the atmosphere of the giant during eclipse, almost like a CAT scan. Observations of photometric and spectroscopic fluctuations during eclipse provide a unique opportunity for studying turbulence in the envelopes of evolved stars. The systems are long period, although for several, notably

22 Vul, there is evidence for some interaction between the stars due to accretion of the giant wind by the main sequence star in the form of an accretion wake. The most extreme example of this subclass is ε Aur, a 27-yr-period binary with an unseen supergiant or hypergiant evolved cool star accompanied by an early-type companion.

4. Algol Binaries

These are the classic semidetached systems. They are marked by evidence for gas streams, distortions of the eclipse profiles due to instabilities in the circumstellar material on the time scale of several orbits, and sometimes enhanced radio and X-ray flaring activity of the more evolved star. For several systems, notably U Cep, the stream ejected from the giant hits the outer envelope of the accreting star and spins it up to very high velocity. For others, the stream circulates to form an accretion disk about the companion, which is heated both by turbulent viscosity and the impact of the stream in its periphery. These systems typically show inverted mass ratios, in that the more evolved star has the lower mass. They have masses ranging from about $1M_{\odot}$ each to greater than $5M_{\odot}$ for the constituent stars. Among the best examples of these systems are SX Cas, W Ser, and β Lyr.

5. Symbiotic Stars

These systems are so named because of the observation of strong emission lines from highly ionized species and cool absorption lines of neutral atoms and molecules in the same spectrum. They consist of a highly evolved cool giant or supergiant and either a main sequence, subdwarf, or collapsed companion. Several of the systems, notably R Aqr, show pulsating primary stars with periods of hundreds of days. The orbital periods are typically quite long, of order one year; the mass ratios have not generally inverted except in those systems where a white dwarf is established as a member. The ionizing source appears to be an accretion disk about the companion star, fed by a stellar wind and perhaps gas streams in the system. Radio and optical jets have been observed emanating from several systems, especially CH Cyg and R Aqr. Many of the phenomena observed in these systems are similar to those observed in recurrent novae like RSOph TCrB and V3890 Sgr, which have red giant companions. These systems also display unstable light curves, presumably attributable to instabilities in the accretion disks. The most extreme examples of this class, having the longest periods and the most evolved red components, are the VV Cep stars.

6. Cataclysmic Variables

These systems typically consist of low-mass main sequence stars of less than $1M_{\odot}$, and either white dwarf or neutron star companions. They display outbursts of the nova type when flash nuclear reactions release sufficient energy to expel the outer accreted layers off the surface of the collapsed star, or show unstable disks that appear to account for the dwarf novas. These systems will be discussed later at greater length. Among the best examples of this class are U Gem, SS Cyg, and OY Car for the dwarf novae; GK Per and DQ Her for the classical novae; and AM Her and CW 1103 + 254 for the magnetic white dwarf accreters. The low-mass X-ray binaries share many of the same characteristics without the extreme photometric variability (for example, Cyg X-2 and Sco X-1).

IV. EVOLUTIONARY PROCESSES FOR STARS IN BINARY SYSTEMS

Normally, stars evolve from the main sequence, during which time their energy generation is via core hydrogen fusion, to red giants, when the star is burning helium in its core, at roughly constant mass. While stellar winds carry off some material during the main sequence stages of massive stars, most stars do not undergo serious alteration of their masses until the postgiant stages of their lives. This is only achieved when the escape velocity has been reduced to such an extent that the star can impart sufficient momentum to the outer atmosphere for a flow to be initiated. Envelope expansion reduces the surface gravity of the star, so that the radiative acceleration due to the high luminosity of stars in the postgiant stage, or the extreme heating affected by envelope convection in the outer stellar atmosphere, provides the critical momentum input. In a binary system, however, the star is no longer necessarily free to expand to any arbitrary radius. The presence of a companion fixes the maximum radius at which matter can remain bound to a star.

Should the primary (that is, more massive) star in a binary have a radius that exceeds this value, it will develop a cusp along the line of centers at the *inner Lagrangian point*, L_1. Nuclear processes continue in the stellar interior, driving the increase in the envelope radius, so that even though the mass of the star is decreasing, the center of mass of the system shifts toward the companion star and the continued expansion of the primary causes the mass transfer to be maintained. Inexorably, the mass ratio will continue to increase until the star is so sufficiently stripped of matter that it becomes smaller than the instantaneous Roche lobe. At this point, the mass transfer stops.

The evolution of the system is determined by the fraction of the lost mass that is accreted by the second star and the fraction of both mass and angular momentum of the system that is lost through the outer contact surface. The loss of mass from one of the stars alters its surface

composition as successive layers are peeled off. It is generally assumed that the star will appear as nitrogen enhanced, because the outer layers of the CNO-burning shell will be exposed to view if enough of the envelope is removed. The OBN and WN stars are assumed to be the result of such processing. In addition, the alteration of the mass of the star will produce a change in the behavior of turbulent convection in the envelope, although the details are presently very uncertain. The enhancement of turbulent mixing should be responsible for exposing the effects of nuclear processing of even deeper layers to view, but this has yet to be fully explored.

The behavior of the mass loser with time is significantly sensitive to whether the envelope is convective or radiative, that is, to whether the primary mode of energy transport is by mass motions or photons. In turn, these are sensitive to the temperature gradient. If the envelope is convective, the reduction in mass causes the envelope to expand. If radiative, the envelope will contract on mass being removed. Consequently, the mass transfer is unstable if the envelope is convective, and the star will continue to dump mass onto the companion until it becomes so reduced in mass that its envelope turns radiative. The instability, first described by Bath, may be responsible for the extreme mass-transfer events seen in symbiotic stars and may also be implicated in some nova phenomena.

V. MASS TRANSFER AND MASS LOSS IN BINARIES

Mass transfer between components in a binary system takes place in two ways, by the formation of a stream or a wind. Either can give rise to an accretion disk, depending on the angular momentum of the accreting material. In the ζ Aur systems and in most WR binaries, the accretion is windlike. This also occurs in some high-mass X-rays binaries (HMXRB), notably Cir X-1. In these, the accretion radius is given by the gravitational capture radius for the wind, which varies as:

$$R \sim M \big/ v_{\text{wind}}^2 \tag{8}$$

where M is the mass of the accreting star and v_{wind} is the wind velocity at the gainer. The formation of an accretion wake has been observed in several systems, notably the ζ Aur systems. The wake is accompanied by a shock. Should the star have a wind of its own, however, the material from the primary loser will be accelerated out of the system along an accretion cone, with little actually falling onto the lower mass component. Such wind–wind interaction is observed in Wolf–Rayet systems, notably V444 Cyg, and is responsible for strong X-ray emission.

The formation of a stream is assured if the mass loss rate is low and the star losing mass is in contact with the Roche surface. In this case, the L_1 point in the binary acts to funnel the mass into a narrow cone, which then transports both angular momentum and mass from the loser to the gainer.

The atmosphere of the mass losing star has a finite pressure, even though at the L_1 point the gravitational acceleration vanishes; thus, the mass loss becomes supersonic interior to the throat formed by the equipotentials, and a stream of matter is created between the stars. The fact that the center of mass is not the same as the center of force (that is, L_1) means that the stream has an excess angular momentum when it is in the vicinity of the secondary or mass gaining star. It thus forms an accretion disk around the companion. However, if the ejection velocity is great enough, the size of the companion large enough compared with the separation of the stars, and the mass ratio small enough, the stream will impact the photosphere of the gainer. Instead of the formation of a stable disk, the stream is deflected by the stellar surface, after producing an impacting shock, with the consequence that the outer layers of the gainer are sped up to nearly the local orbital speed, also called the *Keplerian velocity*:

$$v_{\text{K}} = (G M_2 / R_2)^{1/2} \tag{9}$$

Some mass (the fraction is not well known) will also be lost through the outer Lagrangian point, L_3, on the rear side of the loser from the gainer along the line of centers. The mass of the system as a whole is therefore reduced. This means that the matter can also carry away angular momentum from the system. If the reduced mass of the system is given by

$$\mu = M_1 M_2 / M \tag{10}$$

where M is the total mass, then the angular momentum of the binary is

$$J = \mu a^2 \omega = G^{2/3} M_1 M_2 M^{-1/3} \omega^{-4/3} \tag{11}$$

where

$$\frac{1}{\mu} = \frac{1}{M_1} + \frac{1}{M_2}$$

is the reduced mass. The change in the angular momentum of the system then produces a period change

$$F_J = \frac{\delta M_1}{M_1}\left(1 - \frac{1}{3}\frac{M_1}{M}\right) + \frac{\delta M_2}{M_2}\left(1 - \frac{1}{3}\frac{M_2}{M}\right) + \frac{4}{3}\frac{\delta P}{P} \tag{12}$$

for a fraction F_J of angular momentum lost from the system and an amount of δM of mass lost. Notice that the period evolution is very sensitive to both the amount of

angular momentum lost and to the fraction of the mass lost from M_1, which is lost from the binary system.

The loss of angular momentum from the system is one of the currently unknown physical properties of various models. It is the most critical problem currently facing those studying the long-term behavior of the mass transfer in binaries, since it is the controlling factor in the orbital evolution. Two classes of models have been proposed, those in which much of the angular momentum of the stream is stored in the accretion disk or in the spun-up stellar envelope of the gainer and those in which the J is carried out of the system entirely. Magnetic fields can also act to transport angular momentum, and the degree of spin–orbit coupling between the components also affects the overall system evolution. As a result, much of the detailed behavior of mass exchanging or semidetached systems is not yet fully understood.

VI. X-RAY SOURCES AND CATACLYSMICS

The presence of a collapsed component in the system alters much of the observable behavior of binaries. In particular, the signature of mass accretion onto a white dwarf, neutron star, or black hole is X-ray emission. With the discovery of binary X-ray sources in the late 1960s, following the launch of UHURU, the first all-sky X-ray survey satellite, the field has rapidly grown. Early observations were interpreted as accretion onto white dwarf stars, but the physical details of the accretion processes onto specific compact objects have been refined so that it is now possible to distinguish many of the marks of a specific gainer by the observable behavior.

X-ray emission results from accretion onto a collapsed star because of the depth of its gravitational potential well. As the infalling matter traverses the accretion disk, it heats up because of collisions with rapidly revolving matter and radiates most of its energy away. If the disk is optically thick, this radiation will appear at the surface of the accretion disk as a local blackbody emitter at a temperature characteristic of the local heating rate for the matter in the disk. Since the material is slowly drifting inward, because of loss of angular momentum through viscosity-like interactions within the disk, the heating can be likened to that resulting from a turbulent medium that is capable of radiating away kinetic energy gained from infall. The mass distribution is not radially uniform, and the vertical extent of the disk is determined by pressure equilibrium in the z direction, so that the surface area and temperature vary as functions of distance from the central star. As a result, the flux merging from the surface and seen by a distant observer is an integrated one, summing up different regions of the disk which have different temperatures. The emergent spectrum of the material is not that of a blackbody, nor even very similar to a star. In general, it will appear to be a power law distribution with frequency, looking nonthermal but in fact reflecting the run of temperature and pressure in the disk.

A. Accretion Disks Processes

If only one star is in contact with R_{RL}, then secular mass transfer happens. That will now occupy our attention. Mass loss must occur whenever a star comes into contact with its Roche surface as a stream directed toward the companion. The net gravitational acceleration vanishes at L_1. Gas in the envelope of a star whose radius equals R_{RL} therefore generates a pressure-driven acceleration that at L_1 reaches the sound speed. The result is a highly supersonic flow that launches toward the companion with the sound speed from the L_1 point and is deflected by the Coriolis acceleration. The stream orbits the companion and collides with itself, again supersonically, forming an oblique shock that eventually deflects it into an orbit. Ultimately, a disk is formed, the structure of which we now treat.

Mass lost through the L_1 point generally carries angular momentum, so it so cannot fall directly onto the gainer. Even if it were coming from the precise center of mass, the Coriolis acceleration in the corotating frame causes the stream to deviate from radial infall. The condition for disk formation is that the angular momentum is sufficient to send material into orbit around the gainer. If the angular momentum is too small, the stream may directly impact the gainer. The result is local shock heating and a deflection of the stream by the gainer's atmosphere. If a disk does form, this interaction point is moved out, but it is still present. The reason is that the stream, which is flowing supersonically, cannot adjust its structure on the slower sonic time scale. As a result, it slams into the circulating material and forms a standing shock. Since it is an oblique impact, this region refracts the shock and produces a contact, slip, discontinuity at the boundary.

Assuming hydrostatic equilibrium, the disk's vertical structure is governed by the tidal component of the acceleration:

$$\frac{1}{p}\frac{dP}{dz} = -\frac{GM}{r^2}\frac{z}{r} = -\frac{v_K^2}{r^2}z \tag{13}$$

The simplest estimate of the disk thickness comes from assuming that it is vertically isothermal, so $P = pc_s^2$, with c_s being the sound speed. The density therefore displays a gaussian vertical profile with a scale height:

$$z_0 = \frac{c_s}{v_K}r \tag{14}$$

This is a thin disk since generally $c_s/v_K \ll 1$, although because of the radial dependence of the angular velocity, the thickness increases with increasing distance from the central body.

For accretion to occur, the disk gas must lose angular momentum and drift radially inward. If the material remains dissipationless in the disk, and cannot somehow lose angular momentum, it will continue to circulate and accumulate, storing angular momentum that it gained from infall by the stream. This may actually happen in some systems, and it is important in the stability of the planetary ring systems observed in the solar system, but it will not explain accretion. This is why much of the effort in accretion disk theory goes into understanding the mechanisms that generate viscosity, which is parameterized by the quantity η. How does this work? The shear in an axisymmetric flow has the explicit form

$$\sigma_{r\phi} = \frac{\partial v_\phi}{\partial r} - \frac{v_\phi}{r} = r\,\frac{d\omega}{dr} \tag{15}$$

and, with Σ representing the surface density, the viscous torque is

$$\tau = \frac{1}{r}\frac{\partial}{\partial r}\left[\eta\Sigma r^2\,\frac{d\omega}{dr}\right] \tag{16}$$

For any fluid, the volumetric energy dissipation rate is

$$\frac{dE}{dt} \sim \eta \int \sigma_{r\phi} T_{r\phi}\, dV \tag{17}$$

where V is the volume and the stress, $T_{r\phi}$, is assumed to be proportional to the pressure through $T_{r\phi} - = \alpha P$, where α is a free parameter that is usually assumed to be a constant in space, although not necessarily a constant in time. The viscosity is given by $\eta \approx \alpha\Sigma \sim c_s z_0$. This is the so-called α-disk. With this parameterization, you see that η depends on the pressure which in turn connects the mass transfer through the continuity equation to the vertical structure of the disk. For constant α, the heating also depends on the pressure.

These disks are thermally unstable if the mass accretion rate is changed over time and can react much like a pulsating star. The local energy generation depends on α, and any local heating produces a decrease in Σ because of the increase of z_0. This, in turn, leads to smaller optical depths and subsequent cooling. The collapse of the disk vertically means the plane is "mass loaded" and this matter must be advected inward by the torque. Therefore, in the cooling state, or if α is sufficiently small, the mass transfer can be large toward the gaining body. If, on the other hand, the efficiency of radiation is reduced or the cooling in the vertical direction is large, the disk can remain hot and the mass will not effectively transfer to the inner region. If the disk is optically thin, it can be treated as a circulating gas that immediately radiates any energy gained by its drift onto the gainer, and the emission should result from an accretion shock that may develop at the surface of the mass-accepting star. However, for many systems, especially cataclysmic variables, the depth of the gravitational potential and the opacity of the disk conspire to produce complex time dependence in the disk structure. The vertical temperature gradient governs the surface density through the condition of hydrostatic equilibrium. Remember that, even in the thin disk approximation, the disk thickness depends on T through the scale height. You can see this by using Σ as the dependent variable and rewriting the vertical radiative diffusion equation as

$$F = -\frac{4acT^3}{3\kappa}\frac{\partial T}{\partial \Sigma} \tag{18}$$

Since the opacity depends on both temperature and density, the vertical structure of the disk feeds back into the rate of mass accretion. This is the basis for the *disk oscillations* that are observed in nonlinear models for the disks of cataclysmic variables. These arise mainly because of recombination effects when the column density is high. The oscillations are the response to changes in the mass inflow.

With all sources of dissipation included, the total gravitational potential energy of each parcel of infalling gas will ultimately be radiated, so that an estimate for the total luminosity of the system is $L_{acc} \sim GM\dot{M}/R_\star$, where $R_\star$ is the stellar radius. The temperature should reach about $T_{acc} \sim GM/\kappa R_\star$. Thus, for a compact body, there are two important conclusions to be drawn from this exercise. The first is that a WD, neutron star, or black hole in a binary system should produce emissions ranging from soft to very hard X-rays. The second is that such a source can eat only so fast. The Eddington luminosity sets a limit on the emission a star can support simultaneously in hydrostatic and radiative equilibrium. For a black hole, whose radius depends only on its mass, this means that $\dot{M} \sim M$ so the luminosity of such a source is a measure of its mass accretion rate. For less extreme objects, the luminosity depends on the stellar radius and mass, so that $\dot{M} \sim R_\star$. For a white dwarf, for instance, this means that $\dot{M} \sim M^{-1/3}$. In other words, some of the mass will not wind up on the star if it is accreted too quickly. Mass loss from the binary may happen through jets driven from the disk. It is possible to observationally detect accretion disks, for instance in eclipsing binaries where the absorption lines from the environmental gas are seen in projection against the stellar photosphere. Emission lines from the circulating material have characteristic double peaks separated by the projected orbital velocity of the inner portions of the disk. Finally, the continuum emissions from a viscously heated disk show a power law with increasing flux

to shorter wavelengths, $F_\nu \sim \nu^{1/3}$ for frequency ν. This is because the emission is weighted toward the inner region by temperature but modified by the increased surface area for the cooler outer parts of the flow.

1. Magnetorotational Instability and Viscosity in Disks

Since disks rotate differentially, even a weak magnetic field can have profound consequences for their structure because of dynamo action. The field can be amplified and communicates a torque over a large distance. This produces a very strong instability that is inherent to all conducting media, the *magnetorotational* instability (MRI). Imagine that the disk is threaded by a weak magnetic field that lies in the ϕ and z directions (neglecting the radial component). The combined effects of buoyancy and rotation, which amplify any seed field with time, leads to the instability. Even ignoring buoyancy, the disk is still unstable *unless*

$$\frac{d\Omega}{dr} > 0 \tag{19}$$

Even Keplerian motion fails this stability criterion! The Rayleigh criterion, that the angular momentum must increase outward, may be met by the circulation, yet, even if the seed magnetic field is vanishingly small, it grows through shear and the MRI grows. Although it is not clear from our development how this leads to turbulence, the MRI clearly produces strong, growing fluctuations in the density and velocity that likely become turbulent. We have neglected dissipation, but as soon as the velocity fluctuations grow large enough they will certainly drive heating throughout the disk by the strong viscous coupling. We will close our treatment here with a few additional remarks. This instability is very general, and any ionized shearing medium is likely to experience some form of the MRI. This includes galactic-scale disks as well as protostellar accretion disks.

B. Cataclysmic Variables and Compact Objects in Close Binaries

The cataclysmic variables are a heterogeneous type of close binary, unified phenomenologically by their propensity for "violent" behavior. What they all have in common is that one of the components is a compact, degenerate star—a white dwarf, neutron star, or black hole. The array of types is as dizzying as with any other area of astronomy. Disk-dominated systems, for which the compact star (white dwarf) is overwhelmed with the accretion environment, include the SU Uma stars, for which the disk produces a complex light variation due to a *superhump* that arises from the hot spot in the disk where the stream impacts its periphery. The excess emission progresses in retrograde through the light curve (that is, it moves toward earlier phase with time) and the systems undergo occasional outbursts. These systems are some of the shortest period binaries known; WZ Sge has an orbital period of only about 1 hour. At least one nova, V1974 Cyg, appears to also be a related system. The U Gem systems are eruptive disk systems that undergo nearly periodic outbursts that obey a period–amplitude relation. Magnetic white dwarfs are the gainers in the AM Her systems, a category that also includes some novae such as V1500 Cyg and DQ Her.

Cataclysmics binaries with neutron stars include most low-mass X-ray binaries (LMXRB) such as Cyg X-2 and Sco X-1. These systems have low-mass (red dwarf) companions and resemble the white dwarf cataclysmics in many respects. A more massive system, Her X-1 = HZ Her, has a late A loser and a magnetized neutron star gainer. High-mass X-ray binaries (HMXRB) have more massive (OB) losers, such as Vela X-1, 4U 0900-17, and ϕ Per. The primary is sometimes a Be star, meaning that it shows either a strong wind or evidence for an extended circumstellar environment. Cataclysmics with white dwarf gainers include dwarf nova systems such as WZ Sge, U Gem, and SS Cyg. Classical and some recurrent novae fall into this category, as well, having cool low-mass losers supplying the mass.

C. Novae and X-Ray Bursts: Surface Nuclear Explosions

Classical novae are optically detected by swift increases in brightness of order 10 magnitudes from quiescence to maximum in a mere few days and decays on time scales of between weeks and months. They eject shells with masses between $10^{-7}M_\odot$ and $10^{-4}M_\odot$ at speeds between about 1000 to 10,000 km sec^{-1}. The *speed class* is distinguished by t_2 or t_3, the time required for the optical light curve to fall between 2 and 3 magnitudes from maximum. Such systems do not show repeated outbursts. If they do, on time scales of decades to centuries, they are classified as *recurrents*. Novae are all binary stars in which the gainer is a white dwarf. The loser can be a compact star; an M-type star that can be either a main sequence star or one that is hydrogen poor, indicative of some stripping before or during the evolution of the system; or a giant that is similar to the ones observed in symbiotic stars. No classical system is of the latter variety, but recurrents (V3890 Sgr, RS Oph, and T CrB) are. On the other hand, recurrents can overlap the properties of classical systems (LMC 1990 No. 2 and U Sco have compact companions). X-ray novae, which have much shorter optical duration

and emit most strongly at high energy, arise on accreting neutron stars in systems that otherwise closely resemble low mass X-ray binaries. Several are known to repeat, but on shorter recurrence intervals than recurrent novae. All show rapid rises in the optical and no opaque stage analogous to classical novae.

Piling mass onto a white dwarf produces a very different response than is seen for a normal star. The pressure increases by an amount ΔP at the base of the accreted layer of mass ΔM, depending on the star's surface gravity:

$$\Delta P = \frac{GM\Delta M}{4\pi R^4} \tag{20}$$

so the temperature steadily increases by compression. At a critical threshold, $\Delta P_c \approx 2 \times 10^{19}$ dyne cm^{-2}, the gas achieves the conditions that ignite nuclear reactions. Depending on the white dwarf's composition, these may be reactions on either carbon or heavier elements, but these are details. The critical value of ΔM depends on the mass of the white dwarf, since its radius is mass dependent through $R \sim M^{-1/3}$. The scenario results in the scaling relation $\Delta M \sim M^{-7/3}$, and the higher mass white dwarf does not have to accrete as much before the critical conditions are reached.

Nuclear reactions start at the base of the accreted layer, which is partially degenerate. The resulting release of energy by the reactions does not increase the pressure although the temperature rises continually, resulting in a thermonuclear runaway. You have seen this before, but here it is happening at the surface of a star! The Fermi energy, ϵ_F, corresponds to temperatures of approximately 10^8 K which can be reached as the nuclear source increases in luminosity. The layer crosses the degeneracy threshold when $T \geq \epsilon_F/\kappa$, and at this point the layer rapidly expands and the reactions stop. In a stellar interior, at the time of the helium flash, the bulk of the mass has such a long thermal time scale that the core expands but without a large change in the envelope. Not so here. The stellar radius swells and a deep convection zone forms. The convection is driven by the usual condition that, in order to transport the flux being produced by nuclear reactions, which is the Eddington luminosity, the temperature gradient must be superadiabatic. This turbulence drives deep mixing, which transports heavy elements into the burning zone, thereby providing fuel for the reactions and enhancing the nucleosynthesis. Products of the reactions, in particular ^{13}N and ^{15}O, are β unstable and decay on time scales of about 100 sec, short compared with the sound travel time through the now bloated envelope of the star. When these decay, they typically release 10 MeV/nucleon, sufficient to heat the accreted layer and blow it off the star. The crucial condition is, however, that there must be a substantial production of these nuclei, and this requires that the accreted gas be hydrogen rich and that the white dwarf must be overabundant in target CNO nuclei. Consequently, the layer is ejected very quickly, with speeds exceeding the escape velocity from the white dwarf of the order of several thousand km sec^{-1}. The rapid expansion is accompanied by a dramatic increase in the optical brightness of the star, the rise indicating the onset of the classical nova outburst. When this happens in a stellar core during the helium flash, the overlying layers respond to the increased energy release on the thermal time scale but otherwise remain essentially hydrostatic. In a nova, there is nothing to prevent the explosion from ejecting the outer envelope. Novae come in two distinct varieties based on ejecta abundance patterns, and these are the observational support for the nucleosynthesis we have invoked to explain the explosion. Most novae show strongly enhanced CNO elements and little else that has been changed. These appear to come from roughly 1-$M_\odot$ white dwarfs, with progenitor masses in the range of a few solar masses.These pre-nova stars must have been able to complete helium core burning but then can have their envelopes removed through mass loss caused by the binary. Another group, the ONeMg subclass, show strongly enhanced abundances of Ne and heavier elements, for which higher core temperatures in the pre-nova star are required and must therefore come from a relatively rarer progenitor. These more massive gainers are thought to be closer to the Chandrasekhar limit and therefore to represent possible precursors to SN Ia events.

In current scenarios for type Ia supernova explosions, gainers at or near the mass limit for a white dwarf, called the Chandrasekhar mass, appear to be involved. One of the likely possibilities is not substantially different than a nova explosion, with the key exception that the trigger is not hydrogen accretion but *carbon ignition*. The temperature sensitivity of this reaction is such that once it begins in the white dwarf, the energy release produces an explosion rather than a fizzle. The debate is currently over whether the nuclear reaction proceeds as a deflagration wave or as an actual detonation. In the latter, the expansion of the shock and the local energy release power the continued expansion and subsequent burning. In the former, the process is subsonic and leads to sufficient energy release that the star disrupts by thermal conduction.

On a neutron star, the same fundamental process occurs except that the degeneracy is far greater and therefore lifts only after much a much higher luminosity is reached for the nuclear source. This is the mechanism for X-ray bursts. Nucleosynthesis can proceed much further because of the higher densities and temperatures, since the Fermi level is $>10^9$ K, compared with $(1-3) \times 10^8$ K for the most massive white dwarf in a nova system. Rapid proton capture provides nuclear processing far beyond the iron peak, up to ^{95}Mo, without disrupting the star. Very low mass loss is

expected, and the luminosities reach the Eddington value for a few seconds. Even less mass will be ejected than for a white dwarf, approximately 10^{-11} $M_\odot$, and, consequently, the event is *much* faster than for a classical nova, seconds compared to days.

D. Black Holes in Binary Systems

The search for binary black hole candidates began in the late 1960s with the (now obvious) suggestion by Zel'dovich and Thorne that X-ray emission would be a signal of accretion onto a compact object. Using only radial velocities, a mass can be obtained for the unseen component, and if this exceeds the maximum possible for a stable neutron star, the only alternative must be a black hole. The first such system, Cyg X-1, was discovered in 1972. Radial velocity measurements of HD 269858, the optical counterpart, by Bolton and Webster and Murdin showed that this O star is a binary with a period of 5.6 days and a large mass function that exceeded reasonable upper limits for a neutron star companion. Since then, many such systems have been detected, often accompanying the observation of X-ray nova outbursts. These are listed in Table I. The X-ray signature includes a very hard source (with a power law that extends to >100 keV) and, for the X-ray novae, strong variability at all wavelengths, often including a nonthermal radio source and even relativistic jets.

A fundamental difference between a black hole as gainer and other compact objects is the inner boundary condition. While a neutron star or white dwarf presents a surface onto which the mass must ultimately settle, a black hole is an open flow. If the cooling time for the gas is long enough, specifically if the collision time becomes short compared to the freefall time at any point in the flow, the radiative efficiency of the gas will drop. This is the explanation for the lack of emission from the innermost parts of the disk, a model called *advection-dominated accretion flows*. Since the cross-section for thermal bremsstrahlung (light emitted following thermalizing collisions between ions and electrons) decreases with increasing particle kinetic energy, the gas becomes progressively hotter from collisions but cannot remove the extra energy as it collapses through the Schwarzschild radius. This is a generic plasma process and happens regardless of the mass of the accreting hole. In yet another way, binary systems become laboratories—they reveal processes that are also expected to dominate the observable emission from black holes in active galactic nuclei.

TABLE I Properties of Some Binary Black Hole Candidates[a]

System	$f(M)$ ($M_\odot$)	P (days)	M_x ($M_\odot$)
HDE 226868 = Cyg X-1	0.24	5.6	>7
GS2023+33 = V404 Cyg	6.08	6.46	10–14
G2000+25	5.0	0.34	6–18
H1705-25	4.9	0.52	5–8
J1655-40	3.24	2.62	6.5–7.8
A0620-00	3.0	0.32	6±3
GS 1124-68	3.01	0.43	4.5–6.2
GRO J0422+32	1.21	0.21	>4
4U 1543-47	0.22	1.12	2.7–7.5
LMC X-3	2.3	1.7	>7

[a] Data from Blandford, R., and Gehrels, N. (1999). *Phys. Today* 52(6), 41; Charles, P. (1999). *In* "Observational Evidence for Black Holes in the Universe" (S. K. Chakrabarti, ed.), p. 279, Kluwer Academic, Dordrecht/Norwell, MA; Grindlay, J. E. (2000). *In* "Astrophysical Quantities," 4th ed. (A. N. Cox, ed.), Springer–Verlag, Berlin.

The total luminosity is determined by the rate of infall, since it is assumed that accretion is the primary source for emission of radiation, and the mean temperature is given by the shock at the surface of the gainer. As an approximate estimator, the temperature can be obtained from the virial theorem

$$T_s = \frac{GM}{m_p k R} \approx 10^7 \frac{M}{R} K \tag{21}$$

where in the latter equation M and R are in solar units. The luminosity is given by the rate of accretion, so that

$$L = \frac{GM}{R}\dot{M} \tag{22}$$

where $\dot{M}$ is the mass accretion rate. The maximum luminosity that a source of radiation can emit without driving some of the material off by radiation pressure, the so-called *Eddington limit*, is given by

$$L_{\rm Edd} = 3 \times 10^4 \frac{M}{M_\odot} L_\odot \tag{23}$$

so that the accretion can only be stable as long as $L \leq L_{\rm Edd}$. Most mass estimates for the gainer in X-ray binaries are limited by this luminosity.

The temperature of the emergent spectrum can serve as a guide to the nature of the accreting star. The lower the temperature (that is, the softer the X-ray spectrum), the lower the M/R ratio. White dwarfs have $M/R \approx 10^2$, and neutron stars and black holes have $M/R \geq 10^5$ so that, in general, the more collapsed the gainer, the higher the temperature and the luminosity for the same rate of mass accretion. The continuous spectrum from an accretion disk can be approximated as a product of local emission processes. Dissipation depends on the shear and viscosity. If the disk is vertically geometrically thin and optically thick, each annulus radiates locally like a blackbody with a temperature that depends on the rate of shear. Since the luminosity due to viscous dissipation depends on the shear, $S_{r,\phi} = r(d\omega/dr)$, through $L_{\rm diss} \sim -\eta S_{r\phi}^2$, where

η is the bulk viscosity, then the rate of radiation varies as $L_{\text{diss}} \sim r^{-3}$ for Keplerian disks (those with $\omega \sim r^{-3/2}$). Assuming that the local temperature is given by the rate of radiative loss, $T(r) \sim r^{-3/4}$. The inner region of the disk radiates most of the energy but occupies a relatively smaller area than the outer disk, so its area-weighted contribution to the emergent spectrum is reduced. The frequency of the spectral peak of any annulus depends directly on the temperature so the inner portion of the disk contributes to the higher frequency portion of the spectrum. Integrating over the disk, the emergent spectrum is seen to be $F_\nu \sim \nu^{1/3}$, where ν is the frequency and F_ν is the monochromatic flux. The approximate power law dependence of the spectrum is an indication of the local nature of the emission process. More detailed models, including the effects of radiative transfer vertically through the disk, show that the emergent spectrum looks like a combination of a power law and a very hot stellar atmosphere. The importance of this spectral dependence is that different parts of the spectrum probe processes in different portions of the disk.

One fundamental feature of accretion disks in close binaries is that they are unstable over a wide range of temperatures and mass accretion rates. The theory of time-dependent accretion disks is still under development. However, it is clear that limit-cycle variations in temperature, and therefore in luminosity, are possible for a variety of disks. Especially applicable in cataclysmics such as SS Cyg and other dwarf novae, the oscillations of accretion disks are partly driven by the viscosity and its dependence on the local physical properties of the disk. If the mass-transfer rate should increase, the matter will be stored in the disk for some period of time, though the details are not well understood. Angular momentum must be dissipated before the material can accrete onto the central star. This is true regardless of the nature of the mass gainer. Viscosity acts both to dissipate energy and to redistribute angular momentum within the disk. As material falls onto the central body, the accretion produces a rise in temperature and luminosity of the shock and also of the inner part of the disk. The variation of the high-frequency portion of the spectrum relative to the lower frequencies therefore probes the rate of release of mass and energy in the disk during a mass-transfer event. The applicability of this picture to a wide variety of binary systems. Such as symbiotics, dwarf novae, and Algols, ensures that work will continue on this important physical process.

Disk formation is also different in different types of systems. The gravitational potential for a white dwarf is lower, so the circulation velocity of the material is also lower. The efficiency of heating being reduced, the disks are cooler and the opacity is also higher. Consequently, the disks are sensitive to small fluctuations in the local heating rate and may be thermally unstable. This results in both flickering and alteration of the vertical structure of the disk with time. Also, the heating at the inner boundary of both neutron stars and white dwarfs differs from that seen for black holes, since there is actually no surface in the latter on which matter can accumulate. Consequently, neutron stars and white dwarfs are more sensitive to the accumulation of matter on their surfaces and can undergo flash nuclear processing, the initiation of nuclear fusion under low-pressure, upper boundary conditions, which serves to blow matter off their surfaces on time scales corresponding to the rate of expansion of the outer layers because of the release of nuclear binding energy. In effect, these stars behave like the cores of stars without the presence of the overlying matter, while black holes, which may go through some slow nuclear processing of matter as it falls into the central region, cannot initiate flash processes. Thus, novas and dwarf novas appear to be due to non-black-hole systems.

The details of mass accretion can only be understood by detailed modeling of the accretion disk processes. These depend critically on the nature of the environment of the compact object, because the presence of strong magnetic fields appears to significantly alter the details of the accretion. A magnetic field exerts a pressure against the matter infalling to the surface and can serve to channel the matter to the poles of the accreter. In addition, it alters the pressure and density conditions at the interaction region in the plane of the accretion disk, forming an extended layer at large distance from the central star and can thus lower the effective temperature emerging from the shock region.

VII. FORMATION OF BINARY SYSTEMS

This is one of the major areas of study in stellar evolution theory, because at present, there are few good examples of pre-main-sequence binary stars so the field remains dominated by theoretical questions. Protostellar formation begins with the collapse of a portion of an interstellar cloud, which proceeds to form a massive disk. Through viscosity and interaction with the ambient magnetic field, the disk slowly dissipates its angular momentum. If the disk forms additional self-gravitating fragments, they will collide as they circulate and accrete to form more massive structures. Models show that such disks are unstable to the formation of a few massive members, which then accrete unincorporated material and grow.

Classical results for stability analysis of rapidly rotating homogeneous objects point to several possible alternatives for the development of the core object. One is the central star in the disk, if it is still rapidly rotating, may deform to a barlike shape, which can pinch off a low-mass

component. The core may undergo spontaneous fission into fragments, which then evolve separately. Simulations show, however, that the fission scenario does not yield nearly equal mass fragments. Such systems more likely result either from early protostar fragmentation and disk accretion in the first stages of star formation or of coalescence of fragments during some intermediate stage of disk fragmentation before the cores begin to grow.

Binary star formation appears to be one of the avenues by which collapsing clouds relieve themselves of excess angular momentum, replacing spin angular momentum with orbital motion of the components. However, while the distribution of q may be a clue to the mechanism of formation, even this observational quantity is very poorly determined. The discovery of debris disks around several intermediate-mass, main sequence stars, especially β Pic and α Lyr = Vega, has fueled the speculation that planetary systems may be an alternative to the formation of binary stars for some systems. Statistical studies show that radial velocity variations are observed in many low-mass, solar-type stars, but the period and mass ratio distribution for these systems is presently unknown.

VIII. CONCLUDING REMARKS

Binary stars form the cornerstone of the understanding of stellar evolution. They are the one tool with which we can determine masses and luminosities for stars independent of their distances. They present theaters in which many fascinating hydrodynamic processes can be directly observed, and they serve as laboratories for the study of the effects of mass transfer on the evolution of single stars. By the study of their orbital dynamics, we can probe the deep internal density structures of the constituent stars. Finally, binary stars provide examples of some of the most striking available departures from quiescent stellar evolution. The development of space observation, using X-ray (EXOSAT, EINSTEIN, GINGA. ASCF, ROSAT, and Chandra) and ultraviolet (IUE) satellites, has in the past decades, provided much information about the energetics and dynamics of these systems. The coming decade, especially due to observations made with the Hubble Space Telescope, NGSI, and SIRTF, promises to be a fruitful one for binary star research.

SEE ALSO THE FOLLOWING ARTICLES

GALACTIC STRUCTURE AND EVOLUTION • NEUTRON STARS • STAR CLUSTERS • STARS, MASSIVE • STELLAR SPECTROSCOPY • STELLAR STRUCTURE AND EVOLUTION • SUPERNOVAE

BIBLIOGRAPHY

Batten, A. (1973). "Binary and Multiple Star Systems," Pergamon, Oxford.

de Loore, C., and Doom, C. (1992). "Structure and Evolution of Single and Binary Stars," Kluwer Academic, Dordrecht/Norwell, MA.

Frank, J., King, A. R., and Raine, D. (1992). "Accretion Power in Astrophysics," 2nd ed., Cambridge Univ. Press, London.

Kenyon, S. J. (1986). "The Symbiotic Stars," Cambridge Univ. Press, London and New York.

Popper, D. M., Ulrich, R. K., and Plavec, M., eds. (1980). "Close Binary Systems: IAU Symp. 88," Reidel, Dordrecht, Netherlands.

Pringle, J. (1981). "Accretion disks in astrophysics," *Annu. Rev. Astron. Astrophys.* **19,** 137.

Shore, S. N., Livio, M., and van den Heuvel, E. P. J. (1994). "Interacting Binary Systems," (H. Nussbaumer and A. Orr, eds.), Springer–Verlag, Berlin.

Shu, F., and Lubrow, S. (1981). "Mass, angular momentum, and energy transfer in close binary systems," *Annu. Rev. Astron. Astrophys.* **19,** 277.

van der Kamp, P. (1986). "Dark Companions of Stars," Reidel, Dordrecht, Netherlands.

Cosmic Inflation

Edward P. Tryon
Hunter College and Graduate Center of The City University of New York

GLOSSARY

Comoving Moving in unison with the overall expansion of the universe.

Cosmic microwave background (CMB) Emitted when the universe was ~300,000 years old, this electromagnetic radiation fills all of space and contains detailed information about the universe at that early time.

Cosmic scale factor The distance between any two comoving points is a constant fraction of the cosmic scale factor $a(t)$, which grows in unison with the cosmic expansion. The finite volume of a closed universe is $V_u(t) = 2\pi^2 a^3(t)$.

Cosmological constant Physical constant Λ in the field equations of general relativity, corresponding to an effective force that increases in proportion to distance and is repulsive for positive Λ. Its effects are equivalent to those of constant vacuum energy and pressure.

Critical density Curvature of space is positive, zero, or negative according to whether the average density of mass-energy is greater than, equal to, or less than a critical density determined by the Hubble parameter and cosmological constant.

Curvature constant A constant k that is $+1$, 0, or -1 for positive, zero, or negative spatial curvature, respectively.

Doppler effect Change in observed frequency of any type of wave due to motion of the source relative to the observer: motion of the source away from an observer decreases the observed frequency by a factor determined by the relative velocity.

False vacuum Metastable or unstable state predicted by grand unified theories, in which space has pressure P and mass-energy density ρ related by $P = -c^2\rho$; the gravitational effects mimic a positive cosmological constant.

Field Type of variable representing quantities that may, in principle, be measured (or theoretically analyzed) at any point of space–time; hence, fields possess values

(perhaps zero in some regions) at every point of space–time.

Higgs fields In unified theories, field quantities that, when all are zero, have positive energy density and negative pressure giving rise to false vacuum. At least one Higgs field must be nonzero in the state of minimum energy (true vacuum), which breaks an underlying symmetry of the theory.

Horizon distance Continually increasing distance beyond which neither light nor any causal influence could yet have traveled since the universe began.

Hubble parameter Ratio H of the recessional velocity of a comoving object to its distance from a comoving observer; by Hubble's law, this ratio has the same value for all such objects and observers. Its present value is called *Hubble's constant.*

Inflaton fields In theories of elementary particles, the generic term for fields with a false vacuum state that could give rise to inflation; e.g., Higgs fields.

Riemannian geometry The geometry of any curved space or space-time that is approximately flat over short distances.

Robertson–Walker (RW) metric: In homogeneous, isotropic, cosmological models, describes space–time and its evolution in terms of a curvature constant and (evolving) cosmic scale factor.

Second-rank tensor A 4×4 array (i.e., matrix) of quantities whose values depend, in a characteristic way, on the reference frame in which they are evaluated.

Stress-energy tensor The source of gravity in general relativity, a second-rank tensor whose components include the energy and momentum densities, pressure, and other stresses.

Work-energy theorem Any change in energy of a system equals the work done on it; as a corollary, the energy of an isolated system is constant.

COSMIC INFLATION refers to a conjectured, early period of exponential growth, during which the universe is believed to have increased in linear dimensions by a factor exceeding $\sim 10^{30}$. The inflationary period ended before the universe was $\sim 10^{-30}$ seconds old, after which cosmic evolution proceeded in accord with the standard big bang model. Inflation is based on the general theory of relativity together with grand unified (and several other) theories of the elementary particles. The latter predict a peculiar state called the *false vacuum*, which rendered gravitation repulsive for a very early and brief time, thereby causing inflation. Several long-standing cosmic puzzles are resolved by this conjecture.

Standard theories of inflation predict that our universe is spatially flat. This prediction is strikingly supported by recent studies of the cosmic microwave background (CMB). Inflation also makes quantitative predictions about tiny deviations from perfect uniformity at very early times, the slight concentrations of matter that later grew into galaxies and larger structures. These predictions also appear to be confirmed by recent, detailed studies of the CMB.

I. EINSTEIN'S THEORY OF GRAVITATION

In Newton's theory of gravitation, mass is the source of gravity and the resulting force is universally attractive. The empirical successes of Newton's theory are numerous and familiar. All observations of gravitational phenomena prior to this century were consistent with Newtonian theory, with one exception: a minor feature of Mercury's orbit. Precise observations of the orbit dating from 1765 revealed an advance of the perihelion amounting to 43 arc sec per century in excess of that which could be explained by the perturbing effects of other planets. Efforts were made to explain this discrepancy by postulating the existence of small, perturbing masses in close orbit around the sun, and this type of explanation retained some plausibility into the early part of the twentieth century. Increasingly refined observations, however, have ruled out such a possibility: there is not sufficient mass in close orbit around the sun to explain the discrepancy.

In 1907, Einstein published the first in a series of papers in which he sought to develop a relativistic theory of gravitation. From 1913 onward, he identified the gravitational field with the metric tensor of Riemannian (i.e., curved) space–time. These efforts culminated in his general theory of relativity (GTR), which was summarized and published in its essentially final form in 1916. The scope and power of the theory made possible the construction of detailed models for the entire universe, and Einstein published the first such model in 1917. This early model was subsequently retracted, but the science of mathematical cosmology had arrived.

The GTR reproduces the successful predictions of Newtonian theory and furthermore explains the perihelion advance of Mercury's orbit. Measurement of the bending of starlight in close passage by the sun during the solar eclipse of 1919 resulted in further, dramatic confirmation of GTR, which has since been accepted as the standard theory of gravitation. Additional successes of GTR include correct predictions of gravitational time dilation and, beginning in 1967, the time required for radar echoes to return to Earth from Mercury and artificial satellites.

There are no macroscopic phenomena for which the predictions of GTR are known (or suspected) to fail. Efforts to quantize GTR (i.e., to wed GTR with the principles of quantum theory) have met with continuing frustration,

primarily because the customary perturbative method for constructing solutions to quantum equations yields divergent integrals whose infinities cannot be canceled against one another (the quantized version of GTR is not renormalizable). Hence there is widespread belief that microscopic gravitational phenomena involve one or more features differing from those of GTR in a quantized form. In the *macroscopic* realm, however, any acceptable theory of gravitation will have to reproduce the successful predictions of GTR. It is reasonable (though not strictly necessary) to suppose that all the macroscopic predictions of GTR, including those that remain untested, will be reproduced by any theory that may eventually supersede it.

For gravitational phenomena, the obvious borderline between macroscopic and microscopic domains is provided by the *Planck length* ($L_{\rm P}$):

$$L_{\rm P} \equiv (Gh/c^3)^{1/2} = 4.0 \times 10^{-33}\,{\rm cm} \qquad (1)$$

where G is Newton's constant of gravitation, h is Planck's constant, and c is the speed of light. Quantum effects of gravity are expected to be significant over distances comparable to or less than $L_{\rm P}$. According to the wave–particle duality of quantum theory, such short distances are probed only by particles whose momenta exceed Planck's constant divided by $L_{\rm P}$. Such particles are highly relativistic, so their momenta equal their kinetic energies divided by c. Setting their energies equal to a thermal value of order kT (where k is Boltzmann's constant and T the absolute temperature), the *Planck temperature* ($T_{\rm P}$) is obtained:

$$T_{\rm P} \equiv ch/kL_{\rm P} = 3.6 \times 10^{32}\,{\rm K} \qquad (2)$$

where "K" denotes kelvins (degrees on the absolute scale). For temperatures comparable to or greater than $T_{\rm P}$, effects of quantum gravity may be important.

The fundamental constants of quantum gravity can also be combined to form the *Planck mass* ($M_{\rm P}$):

$$M_{\rm P} \equiv (hc/G)^{1/2} = 5.5 \times 10^{-5}\,{\rm g} \qquad (3)$$

The Planck mass and length give rise to the *Planck density* $\rho_{\rm P}$:

$$\rho_{\rm P} \equiv M_{\rm P}\big/L_{\rm P}^3 = 8.2 \times 10^{92}\,{\rm g/cm^3} \qquad (4)$$

For mass densities comparable to or greater than $\rho_{\rm P}$, the effects of quantum gravity are expected to be significant. (Definitions of $L_{\rm P}$, $T_{\rm P}$, $M_{\rm P}$, and $\rho_{\rm P}$ sometimes differ from those given here by factors of order unity, but only the orders of magnitude are important in most applications.)

According to the standard big bang model, the very early universe was filled with an extremely hot, dense gas of elementary particles and radiation in (or near) thermal equilibrium. As the universe expanded, the temperature and density both decreased. Calculations indicate that the temperature fell below $T_{\rm P}$ when the universe was $\sim 10^{-45}$ sec. old, and the density fell below $\rho_{\rm P}$ at $\sim 10^{-44}$ sec. Hence currently unknown effects of quantum gravity may have been important during the first $\sim 10^{-44}$ sec or so of cosmic evolution. For all later times, however, the temperature and density were low enough that GTR should provide an adequate description of gravitation, and our theories of elementary particles together with thermodynamics and other established branches of physics should be applicable in a straightforward way. We shall assume this to be the case.

As remarked earlier, GTR identifies the gravitational field with a second-rank tensor, the metric tensor of curved space–time. The source of gravitation in GTR is therefore not simply mass (nor mass-energy), but another second-rank tensor: the stress-energy tensor. Hence stresses contribute to the source of gravitation in GTR, the simplest example of stress being pressure.

It is commonly said that gravitation is a universally attractive force, but such a sweeping statement is not supported by the field equations of GTR. It is true that the kinds of objects known (or believed) to constitute the present universe attract each other gravitationally, but GTR contains another possibility: granted an extraordinary medium wherein pressure is negative and sufficiently large, the resulting gravitational field would be repulsive. This feature of GTR lies at the heart of cosmic inflation: it is conjectured that during a very early and brief stage of cosmic evolution, space was characterized by a state called a false vacuum with a large, negative pressure. A false vacuum would mimic the effects of a positive cosmological constant in the field equations, as will be explained in some detail in Section IV.G. Cosmic dimensions would have increased at an exponential rate during this brief period, which is called the *inflationary era*.

The concept of negative pressure warrants a brief explanation. A medium characterized by positive pressure, such as a confined gas, has an innate tendency to expand. In contrast, a medium characterized by negative pressure has an innate tendency to contract. For example, consider a solid rubber ball whose surface has been pulled outward in all directions, stretching the rubber within. More generally, for a medium characterized by pressure P, the internal energy U contained within a variable volume V is governed by $\Delta U = -P\Delta V$ for adiabatic expansion or contraction (a result of the work-energy theorem, where ΔU and ΔV denote small changes in U and V, respectively). If U increases when V increases, then P is negative.

The possibility of an early false vacuum with resulting negative pressure and cosmic inflation was discovered in 1979 by Alan H. Guth, while he was studying grand unified theories (GUTs) of elementary particles and their interactions. (The relevant features of GUTs will be described in Section VII.C.) The resulting inflation is believed to have

increased cosmic linear dimensions by a factor of $\sim 10^{30}$ in the first $\sim 10^{-30}$ sec, and probably by very much more.

Over the two decades since Guth's initial discovery, it has been noted that inflation is possible in many different versions of GUTs. Furthermore, superstring theory and several other types of particle theory contain features similar to those of GUTs, and give rise to comparable inflation. Inflation therefore encompasses a *class* of theories, including at least fifty distinct versions that have been proposed since Guth's initial model.

All the theories of elementary particles that predict inflation remain speculative, as does inflation itself. We shall see, however, that inflation would explain several observed features of the universe that had previously defied explanation. Inflation also makes a quantitative prediction about deviations from perfect uniformity in the early universe, deviations that later gave rise to the clumping of matter into galaxies and clusters of galaxies. This prediction has recently been strikingly confirmed by detailed studies of the cosmic background radiation. Empirical evidence for an early period of cosmic inflation is therefore quite substantial, whatever the details of the underlying particle theory might prove to be.

In common with the initial model of Guth, the great majority of inflationary theories predict that the average curvature of space is (virtually) zero, corresponding to *flat* universes. Such theories are called "standard inflation." A few theories predict nonzero, negative (*hyperbolic*) spatial curvature, which, if small, cannot be ruled out by present observations. The latter theories are sometimes called "open inflation" (a potentially misleading name, because flat and hyperbolic universes are both infinite, and traditionally have both been called "open"). Most predictions of standard and open inflation are virtually the same, however, so we shall only distinguish between them where they differ.

We shall next enumerate and briefly describe some heretofore unexplained features of the universe that may be understood in terms of cosmic inflation, and then we will proceed to a detailed presentation of the theory.

II. COSMOLOGICAL PUZZLES

A. The Horizon Problem

Both the special and general theories of relativity preclude the transmission of any causal agent at a speed faster than light. Accepting the evidence that our universe has a finite age, each observer is surrounded by a *particle horizon*: no comoving particle beyond this horizon could have yet made its presence known to us in any way, for there has not yet been sufficient time for light (or anything else) to have traveled the intervening distance. Regions of the universe outside each others' horizons are causally disconnected, having had no opportunity to influence each other. The distance to the horizon naturally increases with the passage of time. In the standard big bang model, an order-of-magnitude estimate for the horizon distance is simply the speed of light multiplied by the time that has elapsed since the cosmos began. (Modifications to this simple estimate arise because the universe is expanding and because space–time is curved; see Section VI.A.)

Astronomical observations entail the reception and interpretation of electromagnetic radiation, that is, visible light and, in recent decades, radiation from nonvisible parts of the spectrum such as radio waves, X-rays, and γ-rays. We shall refer to that portion of the cosmos from which we are now receiving electromagnetic radiation as the *observable universe*; it is that fraction of the whole which may be studied by contemporary astronomers.

The universe was virtually opaque to electromagnetic radiation for times earlier than $\sim 300{,}000$ yr (as will be explained). Therefore, the present radius of the observable universe is less than the present horizon distance—$\sim 300{,}000$ ly less in the standard big bang model, or very much less if cosmic inflation occurred at the very early time characteristic of inflationary models. The cosmos is currently ~ 14 billion years old, and the present radius of the observable universe is ~ 50 billion light-years. (Sources of light have been moving away from us while the light traveled toward us.) Approximately 10^{11} galaxies lie within the observable universe. About 10% of these galaxies are found in groups called *clusters*, some of which form larger groups called *superclusters* (i.e., clusters of clusters).

When averaged over suitably large regions, the universe appears to be quite similar in all directions at all distances, i.e., isotropic and homogeneous. The distribution of matter is considerably less homogeneous than was supposed only twenty years ago, however. Over the last two decades, continually improving technology has made possible far more detailed surveys of the sky, and surprising features have been discovered.

Over distances of roughly 100 million light-years, the concentration of matter into galaxies, clusters, and superclusters is roughly a fractal pattern. No fractal pattern holds over larger distances, but new and larger structures have become apparent. Clusters and superclusters are arrayed in filaments, sheets, and walls, with nearly empty voids between them. Diameters of the voids typically range from 600 to 900 million light-years. One sheet of galaxies, the "Great Wall," is about 750 million light-years long, 250 million light-years wide, and 20 million light-years thick. The *largest* voids are nearly 3 *billion* light-years across. In very rough terms, the distribution of matter resembles an irregular foam of soap bubbles, except

that many of the voids are interconnected as in a sponge. No pattern is evident in the locations or sizes of the walls and voids—their distribution appears to be random. In a *statistical* sense, the universe still appears to be isotropic and homogeneous.

More dramatic and precise evidence for isotropy is provided by the cosmic background radiation, whose existence and properties were anticipated in work on the early universe by George Gamow and collaborators in the late 1940s. These investigators assumed that the early universe was very hot, in which case a calculable fraction of the primordial protons and neutrons would have fused together to form several light species of nuclei [^{2}H (*deuterium*), ^{3}He, ^{4}He, and ^{7}Li] during the first few minutes after the big bang. (Heavier nuclei are believed to have formed later, in the cores of stars and during stellar explosions.) The abundances predicted for these light nuclei agreed with observations, which confirmed the theory of their formation.

After this early period of nucleosynthesis, the universe was filled with a hot plasma (ionized gas) of light nuclei and electrons, together with a thermal distribution of photons. When the universe had expanded and cooled for ∼300,000 yr, the temperature dropped below 3000 K, which enabled the nuclei and electrons to combine and form electrically neutral atoms that were essentially transparent to radiation. (At earlier times the plasma was virtually opaque to electromagnetic radiation, because the photons could travel only short distances before being scattered by charged particles: the observable universe extends to the distance where we now see it as it was at the age of ∼300,000 years. This is called the *time of last scattering*.) The thermal photons that existed when the atoms formed traveled freely through the resulting gas, and later through the vast reaches of empty space between the stars and galaxies that eventually formed out of that gas. These photons have survived to the present, and they bathe every observer who looks toward the heavens. (If a television set is tuned to a channel with no local station and the brightness control is turned to its lowest setting, ∼1% of the flecks of "snow" on the screen are a result of this cosmic background radiation.)

The photons now reaching earth have been traveling for ∼14 billion yr and hence were emitted by distant parts of the universe receding from us with speeds near that of light. The Doppler effect has redshifted these photons, but preserved the blackbody distribution; hence the radiation remains characterized by a temperature.

Extending earlier work with Gamow, R. A. Alpher and R. C. Herman described the properties of this cosmic background radiation in 1948, estimating its present temperature to be ∼5 K (with its peak in the *microwave* portion of the spectrum). No technology existed, however, for observing such radiation at that time. In 1964, Arno A. Penzias and Robert W. Wilson accidently discovered this radiation (at a single wavelength) with a large horn antenna built to observe the Echo telecommunications satellite. Subsequent observations by numerous workers established that the spectrum is indeed blackbody, with a temperature of 2.73 K. The present consensus in favor of a hot big bang model resulted from this discovery of the relic radiation.

The cosmic background radiation is isotropic within ∼one part in 10^5 (after the effect of earth's velocity with respect to distant sources is factored out). This high degree of isotropy becomes all the more striking when one realizes that the radiation now reaching earth from opposite directions in the sky was emitted from regions of the cosmos that were over 100 times farther apart than the horizon distance at the time of emission, assuming the horizon distance was that implied by the standard big bang model. How could different regions of the cosmos that had never been in causal contact have been so precisely similar?

Of course the theoretical description of any dynamical system requires a stipulation of initial conditions; one might simply postulate that the universe came into being with perfect isotropy as an initial condition. An initial condition so special and precise begs for an explanation, however; this is called the *horizon problem*.

B. The Flatness Problem

In a sweeping extrapolation beyond the range of observations, it is often assumed that the entire cosmos is homogeneous and isotropic. This assumption is sometimes called the *Cosmological Principle*, a name that is perhaps misleading in suggesting greater sanctity for the assumption than is warranted either by evidence or by logic. (Indeed, current scenarios for cosmic inflation imply that the universe is distinctly different at some great distance beyond the horizon, as will be explained in Section VIII.B.) The existence of a horizon implies, however, that the observable universe has evolved in a manner independent of very distant regions. Hence the evolution of the observable universe should be explicable in terms of a model that assumes the universe to be the same everywhere. In light of this fact and the relative simplicity of models satisfying the Cosmological Principle, such models clearly warrant study.

In 1922, Alexandre Friedmann discovered three classes of solutions to the field equations of GTR that fulfill the Cosmological Principle. These three classes exhaust the possibilities (within the stated assumptions), as was shown in the 1930s by H. P. Robertson and A. G. Walker. The three classes of Friedmann models (also called

Friedmann–Robertson–Walker or Friedmann–Lemaître models) are distinguished from one another by the curvature of space, which may be positive, zero, or negative. The corresponding spatial geometries are *spherical*, *flat*, and *hyperbolic*, respectively. Einstein's field equations correlate the spatial geometry with the average density ρ of mass-energy in the universe. There is a critical density ρ_c such that spherical, flat, and hyperbolic geometries correspond to $\rho > \rho_c$, $\rho = \rho_c$, and $\rho < \rho_c$, respectively. A spherical universe has finite volume (space curves back on itself), and is said to be "closed"; flat and hyperbolic universes have infinite volumes and are said to be "open."

Astronomical observations have revealed no departures from Euclidean geometry on a cosmic scale. Furthermore, recent measurements of the cosmic microwave background reveal space to be so nearly flat that $\rho = (1.0 \pm 0.1)\rho_c$. This is a striking feature of our universe, given that the ratio ρ/ρ_c need only lie in the range from zero to infinity. The ratio can be arbitrarily large in closed universes, which furthermore can be arbitrarily small and short-lived. The ratio can be arbitrarily close to zero in hyperbolic universes, with arbitrarily little matter. Flat universes are quite special, having the highest possible value for ρ/ρ_c of any infinite space. Our universe is remarkably close to flat, and the question of why is called the *flatness problem* (first pointed out, in detail, by Robert H. Dicke and P. James E. Peebles in 1979).

C. The Smoothness Problem

Although the observable universe is homogeneous on a large scale, it is definitely lumpy on smaller scales: the matter is concentrated into stars, galaxies, and clusters of galaxies. The horizon problem concerns the establishment of large-scale homogeneity, whereas the smoothness problem concerns the formation of galaxies and the chain of events that resulted in the observed sizes of galaxies. That a problem exists may be seen in the following way.

A homogeneous distribution of matter is rendered unstable to clumping by gravitational attraction. Any chance concentration of matter will attract surrounding matter to it in a process that feeds on itself. This phenomenon was first analyzed by Sir James Jeans, and is called the *Jeans instability*. If the universe had evolved in accord with the standard big bang model since the earliest moment that we presume to understand (i.e., since the universe was 10^{-44} sec old), then the inevitable thermal fluctuations in density at that early time would have resulted in clumps of matter *very* much larger than the galaxies, clusters, and superclusters that we observe. This was first noted by P. J. E. Peebles in 1968, and the discrepancy is called the *smoothness problem*.

D. Other Problems

In addition to the questions or problems thus far described, there are others of a more speculative nature that may be resolved by cosmic inflation. We shall enumerate and briefly explain them here. The first is a problem that arises in the context of GUTs.

Grand unified theories predict that a large number of magnetic monopoles (isolated north or south magnetic poles) would have been produced at a very early time (when the universe was $\sim 10^{-35}$ sec old). If the universe had expanded in the simple way envisioned by the standard big bang model, then the present abundance of monopoles would be vastly greater than is consistent with observations. (Despite extensive searches, we have no persuasive evidence that *any* monopoles exist.) This is called the *monopole problem* (see Sections VII.C and VIII.A).

The remaining questions and problems that we shall mention concern the earliest moment(s) of the cosmos. In the standard big bang model, one assumes that the universe began in an extremely dense, hot state that was expanding rapidly. In the absence of quantum effects, the density, temperature, and pressure would have been infinite at time zero, a seemingly incomprehensible situation sometimes referred to as the *singularity problem*. Currently unknown effects of quantum gravity might somehow have rendered finite these otherwise infinite quantities at time zero. It would still be natural to wonder, however, why the universe came into being in a state of rapid expansion, and why it was intensely hot.

The early expansion was not a result of the high temperature and resulting pressure, for it can be shown that a universe born stationary but with arbitrarily high (positive) pressure would have begun immediately to contract under the force of gravity (see Section IV.D). We know that any initial temperature would decrease as the universe expands, but the equations of GTR are consistent with an expanding universe having no local random motion: the temperature could have been absolute zero for all time. Why was the early universe hot, and why was it expanding?

A final question lies at the very borderline of science, but has recently become a subject of scientific speculation and even detailed model-building: How and why did the big bang occur? Is it possible to understand, in scientific terms, the creation of a universe *ex nihilo* (from nothing)?

The theory of cosmic inflation plays a significant role in contemporary discussions of all the aforementioned questions and/or problems, appearing to solve some while holding promise for resolving the others. Some readers may wish to understand the underlying description of cosmic geometry and evolution provided by GTR. Sections

III and IV develop the relevant features of GTR in considerable detail, using algebra and, occasionally, some basic elements of calculus. Section V describes the content, age, and future of the universe. Section VI explains the horizon and flatness problems in detail. Most of the equations are explained in prose in the accompanying text, so that a reader with a limited background in mathematics should nevertheless be able to grasp the main ideas. Alternatively, one might choose to skip these sections (especially in a first reading) and proceed directly to Section VII on unified theories and Higgs fields, without loss of continuity or broad understanding of cosmic inflation (the glossary may be consulted for the few terms introduced in the omitted sections).

III. GEOMETRY OF THE COSMOS

A. Metric Tensors

General relativity presumes that special relativity is valid to arbitrarily great precision over sufficiently small regions of space–time. GTR differs from special relativity, however, in allowing for space–time to be *curved* over extended regions. (Curved space-time is called *Riemannian*, in honor of the mathematician G. F. B. Riemann who, in 1854, made a seminal contribution to the theory of curved spaces.) Another difference is that while gravitation had been regarded as merely one of several forces in the context of special relativity, it is conceptually not a force at all in GTR. Instead, all effects of gravitation are imbedded in, and expressed by, the curvature of space–time.

In special relativity, there is an *invariant interval* s_{12} separating events at t_1, $\mathbf{r}_1$ and t_2, $\mathbf{r}_2$ given by

$$(s_{12})^2 = -c^2(t_1 - t_2)^2 + (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \qquad (5)$$

where x, y, and z denote Cartesian components of the position vector $\mathbf{r}$. If $t_1 = t_2$, $|s_{12}|$ is the distance between $\vec{r}_1$ and $\vec{r}_2$. If $\vec{r}_1 = \vec{r}_2$, then $|s_{12}|/c$ is the time interval between t_1 and t_2. [Some scientists define the overall sign of $(s_{12})^2$ to be opposite from that chosen here, but the physics is not affected so long as consistency is maintained.]

The remarkable relationship between space and time in special relativity is expressed by the fact that the algebraic form of s_{12} does not change under *Lorentz transformations*, which express the space and time coordinates in a moving frame as linear *combinations* of space and time coordinates in the original frame. This mixing of space with time implies a dependence of space and time on the reference frame of the observer, which underlies the counter-intuitive predictions of special relativity. The *flatness* of the space-time of special relativity is implied by the existence of coordinates t and $\vec{r}$ in terms of which s_{12} has the algebraic form of Eq. (5) *throughout* the space-time (called *Minkowski space–time*, after Hermann Minkowski).

The most familiar example of a curved space is the surface of a sphere. It is well-known that the surface of a sphere cannot be covered with a coordinate grid wherein all intersections between grid lines occur at right angles (the lines of longitude on earth are instructive). Similarly, a curved space–time of arbitrary (nonzero) curvature cannot be spanned by any rectangular set of coordinates such as those appearing in Eq. (5). One can, however, cover any *small* portion of a spherical surface with *approximately* rectangular coordinates, and the rectangular approximation can be made arbitrarily precise by choosing the region covered to be sufficiently small. In a like manner, any small region of a curved space–time can be spanned by rectangular coordinates, with the origin placed at the center of the region in question. The use of such coordinates (called *locally flat* or *locally inertial*) enables one to evaluate any infinitesimal interval ds between nearby space–time points by application of Eq. (5).

To span an extended region of a curved space–time with a single set of coordinates, it is necessary that the coordinate grid be curved. Since one has no *a priori* knowledge of how space–time is curved, the equations of GTR are typically expressed in terms of arbitrary coordinates x^μ ($\mu = 0, 1, 2,$ and 3). The central object of study becomes the *metric tensor* $g_{\mu\nu}$, defined implicitly by

$$(ds)^2 = g_{\mu\nu}\, dx^\mu\, dx^\nu \qquad (6)$$

together with the assumption that space–time is *locally* Minkowskian (and subject to an axiomatic proviso that $g_{\mu\nu}$ is symmetric, i.e., $g_{\mu\nu} = g_{\nu\mu}$).

In principle, one can determine $g_{\mu\nu}$ at any given point by using Eq. (5) with locally flat coordinates to determine infinitesimal intervals about that point, and then requiring the components of $g_{\mu\nu}$ to be such that Eq. (6) yields the same results in terms of infinitesimal changes dx^μ in the arbitrary coordinates x^μ. The values so determined for the components of $g_{\mu\nu}$ will clearly depend on the choice of coordinates x^μ, but it can be shown that the variation of $g_{\mu\nu}$ over an *extended* region also determines (or is determined by) the curvature of space–time. It is customary to express models for the geometry of the cosmos as models for the metric tensor, and we shall do so here.

B. The Robertson–Walker Metric

Assuming the universe to be homogeneous and isotropic, coordinates can be chosen such that

$$(ds)^2 = -c^2(dt)^2 + a^2(t)\left\{\frac{(dr)^2}{1-kr^2} + r^2[(d\theta)^2 + \sin^2\theta(d\varphi)^2]\right\} \tag{7}$$

as was shown in 1935 by H. P. Robertson and, independently, by A. G. Walker. The ds of Eq. (7) is called the *Robertson–Walker* (RW) *line element*, and the corresponding $g_{\mu\nu}$ (with coordinates x^μ identified as ct, r, θ, and φ) is called the *Robertson–Walker metric*. The cosmological solutions to GTR discovered by Friedmann in 1922 are homogeneous and isotropic, and may therefore be described in terms of the RW metric.

The spatial coordinates appearing in the RW line element are *comoving*; that is, r, θ, and φ have constant values for any galaxy or other object moving in unison with the overall cosmic expansion (or contraction). Thus t corresponds to proper (ordinary clock) time for comoving objects. The radial coordinate r and the constant k are customarily chosen to be dimensionless, in which case the *cosmic scale factor* $a(t)$ has the dimension of length. The coordinates θ and φ are the usual polar and azimuthal angles of spherical coordinates. Cosmic evolution in time is governed by the behavior of $a(t)$. [Some authors denote the cosmic scale factor by $R(t)$, but we shall reserve the symbol R for the curvature scalar appearing in Section IV.A.]

The constant k may (in principle) be any real number. If $k = 0$, then three-dimensional space for any single value of t is Euclidean, in which case neither $a(t)$ nor r is uniquely defined: their product, however, equals the familiar radial coordinate of flat-space spherical coordinates. If $k \neq 0$, we may define

$$\bar{r} \equiv |k|^{1/2} r \quad \text{and} \quad \bar{a}(t) \equiv a(t)/|k|$$

in terms of which the RW line element becomes

$$(ds)^2 = -c^2(dt)^2 + \bar{a}^2(t)\left\{\frac{(d\bar{r})^2}{1 \mp \bar{r}^2} + \bar{r}^2[(d\theta)^2 + \sin^2\theta(d\varphi)^2]\right\} \tag{8}$$

where the denominator with ambiguous sign corresponds to $1 - (k/|k|)\bar{r}^2$. Note that Eq. (8) is precisely Eq. (7) with $a = \bar{a}$, $r = \bar{r}$, $k = \pm 1$. Hence there is no loss of generality in restricting k to the three values 1, 0, and -1 in Eq. (7), and we shall henceforth do so (as is customary). These values for k correspond to positive, zero, and negative spatial curvature, respectively.

C. The Curvature of Space

If $k = \pm 1$, then the spatial part of the RW line element differs appreciably from the Euclidean case for objects with r near unity, so that the kr^2 term cannot be neglected. The proper (comoving meter-stick) distance to objects with r near unity is of order $a(t)$. Hence $a(t)$ is roughly the distance over which the curvature of space becomes important. This point may be illustrated with the following examples.

The proper distance $r_{\rm p}(t)$ to a comoving object with radial coordinate r is

$$r_{\rm p}(t) = a(t)\int_0^r \frac{dr'}{(1-kr'^2)^{1/2}} \tag{9}$$

The integral appearing in Eq. (9) is elementary, being $\sin^{-1} r$ for $k = +1$, r for $k = 0$, and $\sinh^{-1} r$ for $k = -1$. Thus the radial coordinate is related to proper distance and the scale factor by $r = \sin(r_{\rm p}/a)$ for $k = +1$, $r_{\rm p}/a$ for $k = 0$, and $\sinh(r_{\rm p}/a)$ for $k = -1$.

Two objects with the same radial coordinate but separated from each other by a small angle $\Delta\theta$ are separated by a proper distance given by Eq. (7) as

$$\Delta s = ar\Delta\theta \tag{10}$$

For $k = 0$, $r = r_{\rm p}/a$ and we obtain the Euclidean result $\Delta s = r_{\rm p}\Delta\theta$. For $k = +1$ (or -1), Δs differs from the Euclidean result by the extent to which $\sin(r_{\rm p}/a)$ [or $\sinh(r_{\rm p}/a)$] differs from $r_{\rm p}/a$. Series expansions yield

$$\Delta s = r_{\rm p}\Delta\theta\left\{1 - \frac{k}{6}(r_{\rm p}/a)^2 + \mathbb{O}\left[(r_{\rm p}/a)^4\right]\right\} \tag{11}$$

where "$\mathbb{O}[x^4]$" denotes additional terms containing 4th and higher powers of "x" that are negligible when x is small compared to unity. Equation (11) displays the leading-order departure from Euclidean geometry for $k = \pm 1$, and it also confirms that this departure becomes important only when $r_{\rm p}$ is an appreciable fraction of the cosmic scale factor. Standard inflation predicts the scale factor to be enormously greater than the radius of the observable universe, in which case no curvature of space would be detectable even if $k \neq 0$. Open inflation could produce a wide range of values for the scale factor, with no upper limit to the range.

Equation (7) implies that the proper area A of a spherical surface with radial coordinate r is

$$A = 4\pi a^2 r^2 \tag{12}$$

and that the proper volume V contained within such a sphere is

$$V = 4\pi a^3 \int_0^r dr' \frac{r'^2}{(1-kr'^2)^{1/2}} \tag{13}$$

The integral appearing in Eq. (13) is elementary for all three vlaues of k, but will not be recorded here.

When $k=0$, Eqs. (9), (12), and (13) reproduce the familiar Euclidean relations between A, V, and the proper radius $r_{\rm p}$. For $k=\pm1$, series expansions of A and V in terms of $r_{\rm p}$ yield

$$A = 4\pi r_{\rm p}^2\left\{1 - \frac{k}{3}(r_{\rm p}/a)^2 + \mathbb{O}\left[(r_{\rm p}/a)^4\right]\right\} \qquad (14)$$

$$V = \frac{4\pi}{3} r_{\rm p}^3\left\{1 - \frac{k}{5}(r_{\rm p}/a)^2 + \mathbb{O}\left[(r_{\rm p}/a)^4\right]\right\} \qquad (15)$$

For any given $r_{\rm p}$, the surface area and volume of a sphere are smaller than their Euclidean values when $k=+1$, and larger than Euclidean when $k=-1$.

The number of galaxies within any large volume of space is a measure of its proper volume (provided the distribution of galaxies is truly uniform). By counting galaxies within spheres of differing proper radii (which must also be measured), it is possible in principle to determine whether proper volume increases less rapidly than $r_{\rm p}^3$, at the same rate, or more rapidly; hence whether $k=+1$, 0, or -1. If $k\neq0$, detailed comparison of observations with Eq. (15) would enable one to determine the present value of the cosmic scale factor.

Other schemes also exist for comparing observations with theory to determine the geometry and, if $k\neq0$, the scale factor of the cosmos. Thus far, however, observations have revealed no deviation from flatness. Furthermore, recent precise measurements of the cosmic microwave background indicate that space is flat or, if curved, has a scale factor at least as large as the radius of the observable universe ($\sim$50 billion light-years, explained in Section VI.A).

For $k=0$ and $k=-1$, the coordinate r spans the entire space. This cannot be true for $k=+1$, however, because $(1-kr^2)$ is negative for $r>1$. To understand this situation, a familiar analog in Euclidean space may be helpful. Consider the surface of a sphere of radius R, e.g., earth's surface. Let r, θ, and φ be standard spherical coordinates, with $r=0$ at Earth's center and $\theta=0$ at the north pole. Longitudinal angle then corresponds to φ. Line elements on the surface can be described by $R\,d\theta$ and $R\sin\theta\,d\varphi$, or by $Rd\rho/\sqrt{1-\rho^2}$ and $R\rho\,d\varphi$, where $\rho=\sin\theta$. In the latter case, the radial line element is singular at $\rho=1$, which corresponds to the equator at $\theta=\pi/2$. This singularity reflects the fact that all lines of constant longitude are parallel at the equator. It also limits the range of ρ to $0\le\rho\le\pi/2$, the northern hemisphere.

The Robertson-Walker metric for $k=1$ is singular at $r=1$ because radial lines of constant θ and φ become parallel at $r=1$. In fact this three-dimensional space is precisely the 3-D "surface" of a hypersphere of radius $a(t)$ in 4-D Euclidean space. With $r=0$ at the "north pole," the "equator" lies at $r=1$, so the coordinate r only describes the "northern hemisphere." A simple coordinate for describing the *entire* 3-D space is $u\equiv\sin^{-1}r$, which equals $\pi/2$ at the "equator" and π at the "south pole." The analogs of Eqs. (7)–(15) may readily be found by substituting $\sin u$ for r, wherever r appears. The entire space is then spanned by a finite range of u, namely $0\le u\le\pi$.

Such a universe is said to be *spherical* and *closed*, and $a(t)$ is called the *radius of the universe*. Though lacking any boundary, a spherical universe is finite. The proper circumference is $2\pi a$, and the proper volume is $V(t)=2\pi^2a^3(t)$.

In a curved space, lines as straight as possible are called *geodesics*. Such lines are straight with respect to *locally* flat regions of space surrounding each segment, but curved over longer distances. In spaces described by a Robertson-Walker metric, lines with constant θ and φ are geodesics (though by no means the only ones). On earth's surface, the geodesics are great circles, e.g., lines of constant longitude. Note that adjacent lines of longitude are parallel at the equator, intersect at both poles, and are parallel again at the equator on the opposite side. A spherical universe is very similar. Geodesics (e.g., rays of light) that are parallel in some small region will intersect at a distance of one-fourth the circumference, in either direction, and become parallel again at a distance of half the circumference. If one follows a geodesic for a distance *equal* to the circumference, one returns to the starting point, from the opposite direction.

For $k=0$ and $k=-1$, space is infinite in linear extent and volume; such universes are said to be *open*. As remarked earlier, a universe with $k=0$ has zero spatial curvature for any single value of t; such a universe is said to be *Euclidean* or *flat*. The RW metric with $k=0$ does not, however, correspond to the flat space–time of Minkowski. The RW coordinates are comoving with a spherically symmetric expansion (or contraction), so that the RW time variable is measured by clocks that are moving relative to one another. Special relativity indicates that such clocks measure time differently from the standard clocks of Minkowski space–time. In particular, a set of comoving clocks does not run synchronously with any set of clocks at rest relative to one another. Furthermore, the measure of distance given by the RW metric is that of comoving meter sticks, and it is known from special relativity that moving meter sticks do not give the same results for distance as stationary ones.

Since both time and distance are measured differently in comoving coordinates, the question arises whether the differences compensate for each other to reproduce Minkowski space–time. For $k=0$, the answer is negative: such a space–time can be shown to be curved, provided

the scale factor is changing in time so that comoving instruments actually are moving.

If $k = -1$, space has negative curvature: geodesics (e.g., light rays) that are parallel nearby diverge from each other as one follows them away. An imperfect analogy is the surface of a saddle; two "straight" lines that are locally parallel in the seat of the saddle will diverge from each other as one follows them toward the front or back of the saddle. (The analogy is imperfect because the curvature of space is presumed to be the same everywhere, unlike the curvature of a saddle, which varies from one region to another.) A universe with $k = -1$ is called *hyperbolic*.

D. Hubble's Law

It is evident from Eq. (9) that the proper distance r_p to a comoving object depends on time only through a multiplicative factor of $a(t)$. It follows that

$$\dot{r}_p/r_p = \dot{a}/a \equiv H(t) \tag{16}$$

$$\dot{r}_p = Hr_p \tag{17}$$

where dots over the symbols denote differentiation with respect to time, for example, $\dot{f} \equiv df/dt$. Thus $\dot{r}_p$ equals the velocity of a comoving object relative to the earth (assuming that earth itself is a comoving object), being positive for recession or negative for approach. An important proviso, however, arises from the fact that the concept of velocity relative to earth becomes problematic if the object is so distant that curvature of the intervening space–time is significant. Also, r_p is (hypothetically) defined in terms of comoving meter sticks laid end to end. All of these comoving meter sticks, except the closest one, are moving relative to earth, and are therefore perceived by us to be shortened because of the familiar length contraction of special relativity. Thus r_p deviates from distance in the familiar sense for objects receding with velocities near that of light. For these reasons, $\dot{r}_p$ only corresponds to velocity in a familiar sense if r_p is less than the distance over which space–time curves appreciably and $\dot{r}_p$ is appreciably less than c.

A noteworthy feature of Eq. (17) is that $\dot{r}_p$ exceeds c for objects with $r_p > c/H$. There is no violation of the basic principle that c is a limiting velocity, however. As just noted, $\dot{r}_p$ only corresponds to velocity when small compared to c. Furthermore, a correct statement of the principle fo limiting velocity is that no object may *pass by* another object with a speed greater than c. In Minkowski space–time the principle may be extended to the relative velocity between distant objects, but no such extension to distant objects is possible in a curved space–time.

Equations (16) and (17) state that if $\dot{a} \neq 0$, then all comoving objects are receding from (or approaching) earth with velocities that are proportional to their distances from us, subject to the provisos stated above. One of the most direct points of contact between theoretical formalism and observation resides in the fact that, aside from local random motions of limited magnitude, the galaxies constituting the visible matter of our universe are receding from us with velocities proportional to their distances. This fact was first reported by Edwin Hubble in 1929, based on determinations of distance (later revised) made with the 100-in. telescope that was completed in 1918 on Mt. Wilson in southwest California. Velocities of the galaxies in Hubble's sample had been determined by V. M. Slipher, utilizing the redshift of spectral lines caused by the Doppler effect.

The function $H(t)$ defined by Eq. (16) is called the Hubble parameter. Its present value, which we shall denote by H_0, is traditionally called *Hubble's constant* (a misnomer, since the "present" value changes with time, albeit very slowly by human standards). Equation (17), governing recessional velocities, is known as *Hubble's law*.

The value of H_0 is central to a quantitative understanding of our universe. Assuming that redshifts of spectral lines result entirely from the Doppler effect, it is straightforward to determine the recessional velocities of distant galaxies. Determinations of distance, however, are much more challenging.

Most methods for measuring distances beyond the Milky Way are based on the appearance from earth of objects whose intrinsic properties are (believed to be) known. Intrinsic luminosity is the simplest property used in this way. Apparent brightness decreases at increasing distance, so the distance to any (visible) object of known luminosity is readily determined. Objects of known intrinsic luminosity are called "standard candles," a term that is sometimes applied more generally to *any* kind of distance indicator.

Given a set of identical objects, the relevant properties of (at least) *one* of them must be measured in order to use them as standard candles. This can be a challenge. For example, the brightest galaxies in rich clusters are probably similar, representing an upper limit to galactic size. Since they are luminous enough to be visible at great distances (out to $\sim 10^{10}$ ly), they would be powerful standard candles. Determining their intrinsic luminosity (i.e., *calibrating* them) was difficult, however, because the nearest rich cluster is (now known to be) $\sim 5 \times 10^7$ light-years away (the *Virgo cluster*, containing $\sim$2500 galaxies). This distance is much too great for measurement by trigonometric parallax, which is ineffective beyond a few thousand light-years. Also note that even when calibrated, the brightest galaxies are useless for measuring distances less that $\sim 5 \times 10^7$ light-years, the distance to the nearest rich cluster.

In order to calibrate sets of standard candles spanning a wide range of distances, a *cosmic distance ladder* has been constructed. Different sets of standard candles form successive rungs of this ladder, with intrinsic luminosity increasing at each step upward. The bottom rung is based on parallax and other types of stellar motion. The resulting distances have been used to determine the intrinsic luminosities of main-sequence stars of various spectral types, which form the second rung. Every rung above the first is calibrated by using standard candles of the rung immediately below it, in a bootstrap fashion. Since any errors in calibration are compounded as one moves up the ladder, it is clearly desirable to minimize the number of rungs required to reach the top.

A key role has been played by *variable stars* (the third rung), among which *Cepheids* are the brightest and most important. Cepheids are supergiant, pulsating stars, whose size and brightness oscillate with periods ranging from (roughly) 1 to 100 days, depending on the star. In 1912, a discovery of historic importance was published by Henrietta S. Leavitt, who had been observing 25 Cepheids in the Small Magellenic Cloud (a small galaxy quite near the Milky Way).

Leavitt reported that the periods and apparent brightnesses of these stars were strongly correlated, being roughly proportional. Knowing that all of these stars were about the same distance from earth, she concluded that the *intrinsic* luminosities of Cepheids are roughly proportional to their periods. She was unable to determine the constant of proportionality, however, because no Cepheids were close enough for the distance to be measured by parallax.

Over the next decade, indirect and laborious methods were used by others to estimate the distances to several of the nearest Cepheids. Combined with the work of Leavitt, a *period-luminosity relation* was thereby established, enabling one to infer the distance to any Cepheid from its period and apparent brightness. Only a small fraction of stars are Cepheids, but every galaxy contains a great many of them. When Cepheids were observed (by Hubble) in Andromeda in 1923, the period-luminosity relation revealed that Andromeda lies far outside the Milky Way. This was the first actual proof that other galaxies exist.

Until recently, telescopes were only powerful enough to observe Cepheids out to $\sim$8 million light-years. At this distance (and considerably beyond), random motions of galaxies obscure the systematic expansion described by Hubble's law. Two more rungs were required in the distance ladder in order to reach distances where motion is governed by the Hubble flow. (This partly explains why Hubble's original value for H_0 was too large by $\sim$700%.) Different observers favored different choices for the higher rungs, and made different corrections for a host of complicating factors. During the two decades prior to launching of the Hubble Space Telescope (HST) in 1990, values for H_0 ranging from 50 to 100 km/sec/Mpc were published (where Mpc denotes *megaparsec*, with $1\ \text{Mpc} = 3.26 \times 10^6\ \text{ly} = 3.09 \times 10^{19}$ km). The reported uncertainties were typically near $\pm$10km/sec/Mpc, much too small to explain the wide range of values reported. Unknown systematic errors were clearly present.

The initially flawed mirror of the HST was corrected in 1993, and the reliability of distance indicators has improved substantially since then. Our knowledge of Cepheids, already good, is now even better. Far more important, the HST has identified Cepheids in galaxies out to $\sim$80 million light-years, nearly10 times farther than had previously been achieved. This tenfold increase in the range of Cepheids has placed them next to the highest rung of the distance ladder. Every standard candle of greater luminosity and range has now been calibrated with Cepheids, substantially improving their reliabilities.

Ground-based astronomy has also become much more powerful over the last decade. Several new and very large telescopes have been built, with more sensitive light detectors (charge-coupled devices), and with computer software that minimizes atmospheric distortions of the images. Large areas of the sky are surveyed and, when objects of special interest are identified, the HST is also trained on them. Collaboration of this kind has made possible the calibration of a new set of standard candles, supernovae of type Ia (SNe Ia, or SN Ia when singular).

Type SNe Ia occur in binary star systems containing a white dwarf and a bloated, nearby giant. The dwarf gradually accretes matter from its neighbor until a new burst of nuclear fusion occurs, with explosive force. The peak luminosity equals that of an *entire galaxy*, and is nearly the same for all SNe Ia. Furthermore, the rate at which the luminosity decays is correlated with its value in a known way, permitting a measurement of distance within $\pm$6% for *each* SN Ia. This precision exceeds that of any other standard candle.

Because of their enormous luminosities, SNe Ia can be studied out to very great distances (at least 12 billion light-years with present technology). In 1997, observations of SNe Ia strongly suggested that our universe is expanding at an *increasing* rate, hence that $\Lambda > 0$ (where Λ denotes the cosmological constant). Subsequent studies have bolstered this conclusion. An accelerating expansion is the most dramatic discovery in observational cosmology since the microwave background (1964), and perhaps since Hubble's discovery of the cosmic expansion (1929).

Many calculations in cosmology depend on the value of H_0 , which still contains a range of uncertainty (primarily due to systematic errors, implied by the fact that different standard candles yield somewhat different results). It is

therefore useful (and customary) to characterize H_0 by a dimensionless number h_0 (or simply h), of order unity, defined by

$$H_0 = h_0 \times 100\,\text{km/sec/Mpc} \tag{18}$$

An HST Key Project team led by Wendy L. Freedman, Robert C. Kennicutt, Jr., and Jeremy R. Mould has used nine different secondary indicators whose results, when combined, yield $h_0 = 0.70 \pm 0.07$. Type SNe Ia were one of their nine choices and, by themselves, led to $h_0 = 0.68 \pm 0.05$. The preceding values for h_0 would be reduced by $\sim$0.05 if a standard (but debatable) adjustment were made to correct for differences in metallicity of the Cepheids used as primary indicators.

As previously suggested, there are many devils in the details of distance measurements, and highly respected observers have dealt with them in different ways. For several decades Allan Sandage, with Gustav A. Tammann and other collaborators, has consistently obtained results for h_0 between 0.5 and 0.6. Using SNe Ia as secondary indicators, their most recent result is $h_0 = 0.585 \pm 0.063$. Several other teams have also obtained h_0 from SNe Ia, with results spanning the range from 0.58 to 0.68.

In light of the preceding results and other studies far too numerous to mention, it is widely believed that h_0 lies in the range

$$h_0 = 0.65 \pm 0.10 \tag{19}$$

Before leaving this section on cosmic geometry, we note that a space–time may contain one or more finite regions that are internally homogeneous and isotropic while differing from the exterior. Within such a homogeneous region, space–time can be described by an RW metric (which, however, tells one nothing about the exterior region). This type of space–time is in fact predicted by theories of cosmic inflation; the homogeneity and isotropy of the observable universe are not believed characteristic of the whole.

IV. DYNAMICS OF THE COSMOS

A. Einstein's Field Equations

A theory of cosmic evolution requires dynamical equations, taken here to be the field equations of GTR:

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = 8\pi G c^{-4} T_{\mu\nu} \tag{20}$$

where $R_{\mu\nu}$ denotes the Ricci tensor (contracted Riemann curvature tensor), R denotes the curvature scalar (contracted Ricci tensor), Λ denotes the cosmological constant, and $T_{\mu\nu}$ denotes the stress-energy tensor of all forms of matter and energy excluding gravity. The Ricci tensor and curvature scalar are determined by $g_{\mu\nu}$ together with its first and second partial derivatives, so that Eq. (20) is a set of second-order (nonlinear) partial differential equations for $g_{\mu\nu}$. As with all differential equations, initial and/or boundary conditions must be assumed in order to specify a unique solution. (A definition of the Riemann curvature tensor lies beyond the scope of this article. We remark, however, that some scientists define it with an opposite sign convention, in which case $R_{\mu\nu}$ and R change signs and appear in the field equations with opposite signs from those above.)

All tensors appearing in the field equations are symmetric, so there are 10 independent equations in the absence of simplifying constraints. We are assuming space to be homogeneous and isotropic, however, which assures the validity of the RW line element. The RW metric contains only one unknown function, namely the cosmic scale factor $a(t)$, so we expect the number of independent field equations to be considerably reduced for this case.

Since the field equations link the geometry of space–time with the stress-energy tensor, consistency requires that we approximate $T_{\mu\nu}$ by a homogeneous, isotropic form. Thus we imagine all matter and energy to be smoothed out into a homogeneous, isotropic fluid (or gas) of dust and radiation that moves in unison with the cosmic expansion. Denoting the proper mass-energy density by ρ (in units of mass/volume) and the pressure by P, the resulting stress-energy tensor has the form characteristic of a "perfect fluid":

$$T_{\mu\nu} = (\rho + P/c^2) U_\mu U_\nu + P g_{\mu\nu} \tag{21}$$

where U_μ denotes the covariant four-velocity of a comoving point: $U_\mu = g_{\mu\nu}\, dx^\nu / d\tau$, with τ denoting proper time ($d\tau = |ds|/c$).

Under the simplifying assumptions stated above, only two of the 10 field equations are independent. They are differential equations for the cosmic scale factor and may be stated as

$$\left(\frac{\dot a}{a}\right)^2 = \frac{8\pi G}{3}\rho + \frac{c^2\Lambda}{3} - k\left(\frac{c}{a}\right)^2 \tag{22}$$

$$\frac{\ddot a}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{c^2\Lambda}{3} \tag{23}$$

where $\dot a$ and $\ddot a$ denote first and second derivatives, respectively, of $a(t)$ with respect to time. Note that the left side of Eq. (22) is precisely H^2, the square of the Hubble parameter.

B. Local Energy Conservation

A remarkable feature of the field equations is that the *covariant divergence* (the curved–space-time generalization of Minkowski four-divergence) of the left side is identically zero (*Bianchi identities*). Hence the field equations

imply that over a region of space–time small enough to be flat, the Minkowski four-divergence of the stress-energy tensor is zero. A standard mathematical argument then leads to the conclusion that energy and momentum are conserved (locally, which encompasses the domain of experiments known to manifest such conservation). Hence within the context of GTR, energy and momentum conservation are not separate principles to be adduced, but rather are consequences of the field equations to the full extent that conservation is indicated by experiment or theory.

If a physical system is spatially localized in a space–time that is flat (i.e., Minkowskian) at spatial infinity, then the total energy and momentum for the system can be defined and their constancy in time proven. The condition of localization in asymptotically flat space–time is not met, however, by any homogeneous universe, for such a universe lacks a boundary and cannot be regarded as a localized system. The net energy and momentum of an unbounded universe defy meaningful definition, which precludes any statement that they are conserved. (From another point of view, energy is defined as the ability to do work, but an unbounded universe lacks any external system upon which the amount of work might be defined.)

In the present context, local momentum conservation is assured by the presumed homogeneity and isotropy of the cosmic expansion. Local energy conservation, however, implies a relation between ρ and P, which may be stated as

$$\frac{d}{dt}(\rho a^3) = -\frac{P}{c^2}\frac{d}{dt}(a^3) \tag{24}$$

[That Eq. (24) follows from the field equations may be seen by solving Eq. (22) for ρ and differentiating with respect to time. The result involves $\ddot{a}$, which may be eliminated by using Eq. (23) to obtain a relation equivalent to Eq. (24).]

To see the connection between Eq. (24) and energy conservation, consider a changing spherical volume V bounded by a comoving surface. Equation (13) indicates that V is the product of a^3 and a coefficient that is independent of time. Multiplying both sides of Eq. (24) by that coefficient and c^2 yields

$$\frac{d}{dt}(\rho c^2 V) = -P\frac{dV}{dt} \tag{25}$$

The Einstein mass-energy relation implies that ρc^2 equals the energy per unit volume, so $\rho c^2 V$ equals the energy U internal to the volume V. Hence Eq. (25) is simply the work-energy theorem as applied to the adiabatic expansion (or contraction) of a gas or fluid: $\Delta U = -P\Delta V$. Since energy changes only to the extent that work is done, (local) energy conservation is implicit in this result.

By inverting the reasoning that led to Eq. (24), it can be shown that the second of the field equations follows from the first and from energy conservation; that is, Eq. (23) follows from Eqs. (22) and (24). [To reason from Eqs. (23) and (24) to (22) is also possible; k appears as a constant of integration.] Hence evolution of the cosmic scale factor is governed by Eqs. (22) and (24), together with an equation of state that specifies the relation between ρ and P.

C. Equation for the Hubble Parameter

The present expansion of the universe will continue for however long H remains positive. Since $H \equiv \dot{a}/a$, Eq. (22) may be written as

$$H^2 = \frac{8\pi G}{3}\rho + \frac{c^2\Lambda}{3} - k\left(\frac{c}{a}\right)^2 \tag{26}$$

The density ρ is inherently positive. If $\Lambda \geq 0$ (as seems certain), then $H > 0$ for all time if $k \leq 0$. Open universes, flat or hyperbolic, therefore expand forever. A closed, finite universe may (or may not) expand forever, depending on the relation between its total mass M_u and Λ (as will be discussed).

If $P = 0$, Eq. (24) implies that ρ varies inversely with the cube of $a(t)$, which simply corresponds to a given amount of matter becoming diluted as the universe expands:

$$P = 0: \qquad \rho = \rho_0\left(\frac{a_0}{a}\right)^3 \tag{27}$$

where ρ_0 and a_0 denote present values.

The density ρ has been dominated by matter under negligible pressure (i.e., $P \ll \rho c^2$) since the universe was $\sim 10^6$ years old, so Eq. (27) has been a good approximation since that early time. Recalling again that $H \equiv \dot{a}/a$, Eqs. (26) and (27) imply

$$\dot{a}^2 = \frac{8\pi G\rho_0 a_0^3}{3a} + \frac{c^2\Lambda a^2}{3} - kc^2 \tag{28}$$

The scale factor $a(t)$ increases as the universe expands. Open universes expand forever, and their scale factors increase without limit. If $\Lambda = 0$, Eq. (28) implies that $\dot{a}^2 \to -kc^2$ as $t \to \infty$. From Eqs. (9) and (16), we see that every comoving object has $\dot{r}_p \to 0$ in a flat ($k = 0$) universe, while $\dot{r}_p \to c \times \sinh^{-1} r$ in a hyperbolic ($k = -1$) universe.

If $\Lambda > 0$ in an open universe, then $\dot{a}/a \to c\sqrt{\Lambda/3}$ as $a \to \infty$. In this limit,

$$\dot{a} = \Gamma a \qquad (\text{with } \Gamma \equiv c\sqrt{\Lambda/3}) \tag{29}$$

The solution is

$$a(t) = a(t_{in})\exp[\Gamma(t - t_{in})] \tag{30}$$

where t_{in} denotes any initial time that may be convenient. Equation (29) will be satisfied to good approximation after some minimum time $t_{\min}$ has passed, so $a(t)$ will be given for all $t \geq t_{\min}$ by Eq. (30), with $t_{in} = t_{\min}$.

Equation (30) implies an exponential rate of growth for r_p in the final stage of expansion, with a doubling time t_d given by

$$t_d = \frac{\ln 2}{c}\sqrt{\frac{3}{\Lambda}} \tag{31}$$

Equation (29) then implies exponential growth for $\dot{r}_p$ as well, with the same doubling time.

Next we consider the closed, $k=+1$ case in some detail, where $a(t)$ is called the *radius* of the universe. The total mass of such a universe is finite and given by $M_u = \rho V_u = \rho(2\pi^2 a^3)$. Assuming zero pressure for simplicity, M_u is constant, and Eq. (26) implies

$$H^2 = \frac{4GM_u}{3\pi a^3} + \frac{c^2\Lambda}{3} - \frac{c^2}{a^2} \tag{32}$$

If $\Lambda = 0$, H will shrink to zero when the radius $a(t)$ reaches a maximum value

$$a_{\max} = \frac{4GM_u}{3\pi c^2} \tag{33}$$

The time required to reach $a_{\max}$ can be calculated, and is given by

$$t(a_{\max}) = \frac{2GM_u}{3c^3} \tag{34}$$

It is remarkable that both $a_{\max}$ and the time required to reach it are determined by M_u. The dynamics and geometry are so interwoven that the rate of expansion does not enter as an independent parameter.

After reaching $a_{\max}$, the universe will contract ($H<0$) at the same rate as it had expanded. Equation (23) ensures that the scale factor will not remain at $a_{\max}$, and Eq. (28) implies that the magnitude of H is determined by the value of $a(t)$.

If the universe is closed and $\Lambda > 0$, a more careful analysis is required. A detailed study of Eq. (32) reveals that the present expansion will continue forever if

$$\Lambda \geq \left(\frac{\pi c^2}{2GM_u}\right)^2 \tag{35}$$

If the *in*equality holds, the size of the universe will increase without limit. As for the open case, $\dot{a}/a \to c\sqrt{\Lambda/3}$ as $a \to \infty$. After some time t_0 has passed, Eq. (30) will describe $a(t)$, and the final stage of expansion will be the same as for the open case. Both r_p and $\dot{r}_p$ will grow at exponential rates, with a doubling time given by Eq. (31).

If the *equality* in Eq. (35) holds, then the universe will expand forever but approach a maximum, limiting size $a_{\max}$ as $t \to \infty$, with $a_{\max} = 2GM_u/\pi c^2$ (exactly 3/2 as large as for $\Lambda = 0$.) If the inequality (35) is *violated* (small Λ), then the universe will reach a maximum size in a finite time, after which it will contract (as for the special case where $\Lambda = 0$).

We note in passing that despite the name *hyperbolic*, a completely empty universe with $\Lambda = 0$ and $k=-1$ has precisely the flat space-time of Minkowski. Empty Minkowski space-time is homogeneous and isotropic, and therefore must correspond to an RW metric. With $\rho = \Lambda = 0$, Eq. (28) implies that $k=-1$ [and that $\dot{a}=c$, in which case Eq. (23) implies that $P=0$, as does Eq. (24).] The difference in appearance between the Minkowski metric and the equivalent RW metric results from a fact noted earlier, namely that RW coordinates are comoving. The clocks and meter sticks corresponding to the RW metric are moving with respect to each other, so they yield results for time and distance that differ from those ordinarily used in a description of Minkowski space-time.

To make explicit the equivalence between empty, hyperbolic space-time and Minkowski space-time, we recall that Eq. (28) implies $\dot{a}=c$. Since $a=0$ at $t=0$, it follows that $a(t)=ct$ for the hyperbolic case in question. It may readily be verified that the Minkowski metric transforms into the corresponding RW metric with $r_M = rct$, $t_M = t(r^2+1)^{1/2}$, where r_M and t_M denote the usual coordinates for Minkowski space-time.

D. Cosmic Deceleration or Acceleration

We saw in Section IV.B that Eq. (23) contains no information beyond that implied by Eq. (22) and local energy conservation. Equation (23) makes a direct and transparent statement about cosmic evolution, however, and therefore warrants study. Recalling again that the proper distance $r_{\rm p}(t)$ to a comoving object depends on time only through a factor of $a(t)$, we see that Eq. (23) is equivalent to a statement about the acceleration $\ddot{r}_{\rm p}$ of comoving objects (where again the dots refer to time derivatives):

$$\ddot{r}_{\rm p} = \left\{-\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{c^2\Lambda}{3}\right\} r_{\rm p} \tag{36}$$

Note that $\ddot{r}_{\rm p}$ is proportional to $r_{\rm p}$, which is essential to the preservation of Hubble's law over time. A positive $\ddot{r}_{\rm p}$ corresponds to acceleration *away* from us, whereas a negative $\ddot{r}_{\rm p}$ corresponds to acceleration *toward* us (e.g., deceleration of the present expansion). The same provisions that conditioned our interpretation of $\dot{r}_{\rm p}$ as recessional velocity apply here; $\ddot{r}_{\rm p}$ equals acceleration in the familiar sense only for objects whose distance is appreciably less than the scale factor (for $k \neq 0$) and whose $\dot{r}_{\rm p}$ is a small fraction of c.

As a first step in understanding the dynamics contained in Eq. (34), let us consider what Newton's laws would predict for a homogeneous, isotropic universe filling an infinite space described by Euclidean geometry. In particular, let us calculate the acceleration $\ddot{r}_{\rm p}$ (relative to a comoving observer) of a point mass at a distance $r_{\rm p}$ away

from the observer. (The distinction of "proper" distance is superfluous in Newtonian physics, but the symbol $r_{\rm p}$ is being used to maintain consistency of notation.)

In Newton's theory of gravitation, the force F between two point masses m and M varies inversely as the square of the distance d between them: $F = GmM/d^2$. If more than two point masses are present, a principle of superposition states that the net force on any one mass is the vector sum of the individual forces resulting from the other masses.

While it is not obvious (indeed, Newton invented calculus to prove it), the inverse–square dependence on distance together with superposition implies that the force exerted by a uniform sphere of matter on any exterior object is the same as if the sphere's mass were concentrated at its center. It may also be shown that if a spherically symmetric mass distribution has a hollow core, then the mass exterior to the core exerts no net force on any object inside the core. (These theorems may be familiar from electrostatics, where they also hold because Coulomb's law has the same mathematical form as Newton's law of gravitation.)

The above features of Newtonian gravitation may be exploited in this study of cosmology by noting that in the observer's frame of reference, the net force on an object of mass m (e.g., a galaxy) at a distance $r_{\rm p}$ from the observer results entirely from the net mass M contained within the (hypothetical) sphere of radius $r_{\rm p}$ centered on the observer:

$$M = \left(\tfrac{4}{3}\pi r_{\rm p}^3\right)\rho \tag{37}$$

The force resulting from M is the same as if it were concentrated at the sphere's center, and there is no net force from the rest of the universe exterior to the sphere. Combining Newton's law of gravitation with his second law of motion (that is, acceleration equals F/m) yields

$$\ddot{r}_{\rm p} = -GM/r_{\rm p}^2 = -\frac{4\pi G}{3}\rho r_{\rm p} \tag{38}$$

where Eq. (35) has been used to eliminate M, and the minus sign results from the attractive nature of the force.

Note that Eq. (36) agrees precisely with the contribution of ρ to $\ddot{r}_{\rm p}$ in the Einstein field equation (36); this part of relativistic cosmology has a familiar, Newtonian analog. In GTR, however, stresses also contribute to gravitation, as exemplified by the pressure term in Eq. (36). If $\Lambda \neq 0$, then a cosmological term contributes as well.

In ordinary circumstances, pressure is a negligible source of gravity. To see why, we need only compare the gravitational effect of mass with the effect of pressure filling the same volume. For example, lead has a density $\rho = 11.3 \times 10^3\ \mathrm{kg/m^3}$. In order for pressure filling an equal volume to have the same effect, the pressure would need to be $P = c^2\rho/3 = 3.39 \times 10^{20}\ \mathrm{N/m^2}$ (where $1\ \mathrm{N} = 0.225\ \mathrm{lb}$). This is $\sim 3 \times 10^{15}$ times as great as atmospheric pressure, and $\sim 10^9$ times as great as pressure at the center of the earth.

The density ρ appearing in the field equations is positive, so its contribution to $\ddot{r}_p$ in Eq. (36) is negative: the gravitation of mass-energy is attractive. Positive pressure also makes a negative contribution to $\ddot{r}_p$, but pressure may be either positive or negative. Equation (36) implies that negative pressure causes gravitational repulsion.

If $\Lambda \neq 0$, its contribution to $\ddot{r}_p$ has the same sign as Λ. If $\Lambda > 0$ and a time arrives when

$$\Lambda > \frac{4\pi G}{c^2}\left(\rho + \frac{3P}{c^2}\right) \tag{39}$$

then $\ddot{r}_p > 0$, which corresponds to an *accelerating* expansion. The density and pressure both decrease as the universe expands, whereas Λ is constant. After the first moment when the inequality (39) is satisfied, the universe will expand forever, at an accelerating rate. A time will therefore come when Eq. (29) is satisfied, after which r_p and $\dot{r}_p$ will both grow at exponential rates, in accordance with Eqs. (29)–(31). There is recent evidence that our universe has in fact entered such a stage, as will be discussed in Section V.E.

The *name* of the big bang model is somewhat misleading, for it suggests that the expansion of the cosmos is a result of the early high temperature and pressure, rather like the explosion of a gigantic firecracker. As just explained, however, positive pressure leads not to acceleration "outward" of the cosmos, but rather to deceleration of any expansion that may be occurring. Had the universe been born stationary, positive pressure would have contributed to immediate collapse. Since early high pressure did not contribute to the expansion, early high temperature could not have done so either. (It is perhaps no coincidence that the phrase *big bang* was initially coined as a term of derision, by Fred Hoyle—an early opponent of the model. The model has proven successful, however, and the graphic name remains with us.)

E. The Growth of Space

It is intuitively useful to speak of ρ, P, and Λ as leading to gravitational attraction or repulsion, but the geometrical role of gravitation in GTR should not be forgotten. The geometrical significance of gravitation is most apparent for a closed universe, whose total proper volume is finite and equal to $2\pi^2 a^3(t)$. The field equations describe the change in time of $a(t)$, hence of the size of the universe. From a naive point of view, Hubble's law describes cosmic expansion in terms of recessional velocities of galaxies; but more fundamentally, galaxies are moving apart because the space between them is expanding. How else an increasing volume for the universe?

The field equation for $\ddot{a}(t)$ implies an equation for $\ddot{r}_{\rm p}$, which has a naive interpretation in terms of acceleration and causative force. More fundamentally, however, $\ddot{a}$ and

$\ddot{r}_p$ represent the derivative of the rate at which space is expanding. Our intuitions are so attuned to Newton's laws of motion and causative forces that we shall often speak of velocities and accelerations and of gravitational attraction and repulsion, but the geometrical nature of GTR should never be far from mind.

The interpretation of gravity in terms of a growth or shrinkage of space resolves a paradox noted above, namely, that positive pressure tends to make the universe collapse. Cosmic pressure is the same everywhere, and uniform pressure pushes equally on all sides of an object leading to no net force or acceleration. Gravity, however, affects the rate at which space is growing or shrinking, and thereby results in relative accelerations of distant objects. One has no *a priori* intuition as to whether positive pressure should lead to a growth or shrinkage of space; GTR provides the answer described above.

F. The Cosmological Constant

For historical and philosophical reasons beyond the scope of this article, Einstein believed during the second decade of this century that the universe was static. He realized, however, that a static universe was dynamically impossible unless there were some repulsive force to balance the familiar attractive force of gravitation between the galaxies. The mathematical framework of GTR remains consistent when the cosmological term $\Lambda g_{\mu\nu}$ is included in the field equations (20), and Einstein added such a term to stabilize the universe in his initial paper on cosmology in 1917.

Although the existence of other galaxies outside the Milky Way was actually not established until 1923, Einstein (in his typically prescient manner) had incorporated the assumptions of homogeneity and isotropy in his cosmological model of 1917. With these assumptions, the field equations reduce to (22) and (23). The universe is static if and only if $\dot{a}(t) = 0$, which requires (over any extended period of time) that $\ddot{a}(t) = 0$ as well. Applying these conditions to Eqs. (22) and (23), it is readily seen that the universe is static if and only if

$$\Lambda = \frac{4\pi G}{c^2}\left(\rho + \frac{3P}{c^2}\right) \tag{40}$$

$$k\left(\frac{c}{a}\right)^2 = 4\pi G\left(\rho + \frac{P}{c^2}\right) \tag{41}$$

Since our universe is characterized by positive ρ and positive (albeit negligible) P, the model predicts positive values for Λ and k. Hence the universe is closed and finite, with the scale factor (radius of the universe) given by Eq. (37) with $k = 1$. For example, if $\rho \sim 10^{-30}$ g/cm^3 (roughly the density of our universe) and $P = 0$, then Eq. (41) yields a cosmic radius of $\sim 10^{10}$ ly, $\sim 1/5$ the size of our observable universe.

The static model of Einstein has of course been abandoned, for two reasons. In 1929, Edwin Hubble reported that distant galaxies are in fact receding in a systematic way, with velocities that are proportional to their distance from us. On the theoretical side, it was remarked by A. S. Eddington in 1930 that Einstein's static model was intrinsically unstable against fluctuations in the cosmic scale factor: if $a(t)$ were momentarily to decrease, ρ would increase and the universe would collapse at an accelerating rate under the newly dominating, attractive force of ordinary gravity. Conversely, if $a(t)$ momentarily increased, ρ would decrease and the cosmological term would become dominant, thereby causing the universe to expand at an accelerating rate. For this combination of reasons, Einstein recommended in 1931 that the cosmological term be abandoned (i.e., set $\Lambda = 0$); he later referred to the cosmological term as "the biggest blunder of my life." From a contemporary perspective, however, only the static model was ill-conceived. The cosmological term remains a viable possibility, and two recent observations indicate that $\Lambda > 0$ (as will be discussed in Section V.E).

We have seen that Einstein was motivated by a desire to describe the cosmos when he added the $\Lambda g_{\mu\nu}$ term to the field equations, but there is another, more fundamental reason why this addendum is called the cosmological term. Let us recall the structure of Newtonian physics, whose many successful predictions are reproduced by GTR. Newton's theory of gravity predicts that the gravitational force on any object is proportional to its mass, while his second law of motion predicts that the resulting acceleration is inversely proportional to the object's mass. It follows that an object's mass cancels out of the acceleration caused by gravity. Hence the earth (or any other source of gravity) may be regarded as generating an *acceleration field* in the surrounding space. This field causes all objects to accelerate at the same rate, and may be regarded as characteristic of the region of space.

The GTR regards gravitational accelerations as more fundamental than gravitational forces, and focuses attention on the acceleration field in the form of $g_{\mu\nu}$ as the fundamental object of study. The acceleration field caused by a mass M is determined by the field equations (20), wherein M contributes (in a way depending on its position and velocity) to the stress-energy tensor $T_{\mu\nu}$ appearing on the right side. The solution for $g_{\mu\nu}$ then determines the resulting acceleration of any other mass that might be present, mimicking the effect of Newtonian gravitational force. The term $\Lambda g_{\mu\nu}$ in Eq. (20), however, does not depend on the position or even the existence of any matter. Hence the influence of $\Lambda g_{\mu\nu}$ on the solution for $g_{\mu\nu}$ cannot be interpreted in terms of forces between objects.

To understand the physical significance of the cosmological term, let us consider a special case of the field equations, namely their application to the case at hand:

a homogeneous, isotropic universe. As we have seen, the field equations are then equivalent to Eqs. (26) and (34). From the appearance of Λ on the right sides, it is clear that Λ contributes to the recessional velocities of distant galaxies (through H), and to the acceleration of that recession. This part of the acceleration cannot be interpreted in terms of forces acting among the galaxies for, as just remarked, the cosmological term does not presume the existence of galaxies.

The proper interpretation of the cosmological term resides in another remark made earlier: that the recession of distant galaxies should be understood as resulting from growth of the space between them. We conclude that the cosmological term describes a growth of space that is somehow intrinsic to the nature of space and its evolution in time. Such a term clearly has no analog in the physics predating GTR.

G. Equivalence of Λ to Vacuum Energy and Pressure

An inspection of Eqs. (22) and (23) reveals that the contributions of Λ are precisely the same as if all of space were characterized by constant density and pressure given by

$$\rho_V = \frac{c^2\Lambda}{8\pi G} \tag{42}$$

$$P_V = -\frac{c^4\Lambda}{8\pi G} = -c^2\rho_V \tag{43}$$

From a more basic point of view, Eqs. (42), (43), and (21) yield

$$T_{\mu\nu} = P_V g_{\mu\nu} = -c^2\rho_V g_{\mu\nu} \tag{44}$$

Comparing Eq. (44) with the field equations in their most fundamental form, i.e., Eq. (20), the equivalence between Λ and the above ρ_V and P_V is obvious. We also note that local energy conservation is satisfied by a constant ρ_V with $P_V = -c^2\rho_V$, as may be seen from Eq. (24). In fact all known principles permit an interpretation of Λ in terms of the constant ρ_V and P_V given by Eqs. (42) and (43).

Is there any reason to suppose that $T_{\mu\nu}$ might contain a term such as that described above, equivalent to a cosmolgical constant? Indeed there is. In relativistic quantum field theory (QFT), which describes the creation, annihilation, and interaction of elementary particles, the vacuum state is not perfectly empty and void of activity, because such a perfectly ordered state would be incompatible with the uncertainty principles of quantum theory. The vacuum is defined as the state of minimum energy, but it is characterized by spontaneous creation and annihilation of particle–antiparticle pairs that sometimes interact before disappearing.

For most purposes, only changes in energy are physically significant, in which case energy is only defined within an arbitrary additive constant. The energy of the vacuum is therefore often regarded as zero, the simplest convention. In the context of general relativity, however, a major question arises. Does the vacuum of QFT really contribute nothing to the $T_{\mu\nu}$ appearing in Einstein's field equations, or do quantum fluctuations contribute and thereby affect the metric $g_{\mu\nu}$ of spacetime?

Let us consider the possibility that quantum fluctuations act as a source of gravity in the field equations of GTR. These fluctuations occur throughout space, in a way governed by the local physics of special relativity. The energy density of the vacuum should therefore not change as the universe expands. Denoting the vacuum density by ρ_V [with $\rho_V = (\text{energy density})/c^2$)], the corresponding pressure P_V can be deduced from Eq. (24). Assuming that ρ_V is constant, Eq. (24) implies $P_V = -c^2\rho_V$, in perfect accordance with Eq. (43).

If $\Lambda \geq 0$ (as seems certain on empirical grounds), then the corresponding vacuum has $\rho_V \geq 0$ and $P_V \leq 0$. Negative pressure has a simple interpretation in this context. Consider a (hypothetical) spherical surface expanding in unison with the universe, so that all parts of the surface are at rest with respect to comoving coordinates. Assuming the vacuum energy density to be constant, the energy inside the sphere will increase, because of the increasing volume. Local energy conservation implies that positive work is being done to provide the increasing energy, hence that an outward force is being exerted on the surface.

No such force would be required unless the region inside the sphere had an innate tendency to contract, which is precisely what is meant by negative pressure. The required *outward* force on the surface is provided by negative pressure in the *surrounding* region, which exactly balances the internal pressure (as required in order for every part of the surface to remain at rest with respect to comoving coordinates). A constant ρ_V and the P_V of Eq. (43) describe this situation perfectly. It is not yet known how to calculate ρ_V, so the value of ρ_V and the corresponding Λ remain empirical questions.

The theory of cosmic inflation rests on the assumption that during a very early period of cosmic history, space was characterized by an unstable state called a "false vacuum." Over the brief period of its existence, this false vacuum had an enormous density $\rho_{FV} \sim 10^{80}$ g/cm^3, with $P_{FV} \cong -c^2\rho_{FV}$. This caused the cosmic scale factor to grow at an exponential rate [as in Eqs. (29)–(31)], with a doubling time $t_d \sim 10^{-37}$ sec.

Even if the false vacuum lasted for only 10^{-35} sec, about 100 doublings would have occurred, causing the universe to expand by a factor of $\sim 2^{100} \approx 10^{30}$. Recall that in curved spaces, departures from flatness at a distance r_p are of order $(r_p/a)^2$ (as explained in Section III.C). Standard

inflation predicts that even if space *is* curved, the scale factor is now so large that the *observable* universe is virtually flat. [Hyperbolic ("open") inflation is assumed to have occurred in a space with negative curvature. The inflation, while substantial, may not have lasted long enough to render the observable universe ~ flat.]

V. CONTENTS, AGE, AND FUTURE OF THE COSMOS

A. The Critical Density

The field equation (26) implies a relation between H, ρ, Λ, and the geometry of space, since the latter is governed by k. For present purposes, it is convenient (and customary) to express Λ in terms of the equivalent vacuum density ρ_V, given by Eq. (42). The sum of all *other* contributions will be denoted by ρ_M. ("M" is short for matter *and radiation*. The latter is negligible at present, but dominated ρ_M at very early times.) We next consider Eq. (26) with $\Lambda = 0$ and $\rho = \rho_M + \rho_V$.

Setting $k = 0$ in Eq. (26) and solving for ρ, we obtain ρ for the intermediate case of a *flat* universe. Its value is called the *critical density*, and is presently given by

$$\rho_c = \frac{3H_0^2}{8\pi G} = (1.88 \times 10^{-29}\text{g/cm}^3)h_0^2 \qquad (45)$$

where h_0 is defined by Eq. (18). Inspection of Eq. (26) reveals that k is positive, zero, or negative if ρ is greater than, equal to, or less than ρ_c, respectively. We note in passing that ρ_c is extremely small. For $h_0 \cong 0.65$, ρ_c is comparable to a density of five hydrogen atoms per cubic meter of space. This is far closer to a perfect vacuum than could be achieved in a laboratory in the foreseeable future.

To facilitate a comparison of individual densities with the critical density, it is customary to define a set of numbers Ω_i by

$$\rho_i = \Omega_i \rho_c \qquad (46)$$

where the index i can refer to any of the particular kinds of density. For example, $\rho_M = \Omega_M \rho_c$, so $\Omega_M = \rho_M/\rho_c$ is the fraction of ρ_c residing in matter (and radiation). The *total* density is described at present by

$$\Omega_0 \equiv \Omega_M + \Omega_V \qquad (47)$$

Equation (26) now implies that $k > 0$, $k = 0$, or $k < 0$ if $\Omega_0 > 1$, $\Omega_0 = 1$, or $\Omega_0 < 1$, respectively. If $k \neq 0$, the present value of the scale factor is readily seen to be

$$k = \pm 1: \qquad a_0 = \frac{c}{H_0}\sqrt{\frac{k}{\Omega_0 - 1}} \qquad (48)$$

(Recall from Section III.B that if $k = 0$, then a_0 has no *physical* significance, and can be chosen arbitrarily.)

Standard inflation does not predict the geometry of the *entire* universe, only that the *observable* universe is *very close* to flat. As was shown in Section III.C, this would be true for either type of curved space if a_0 were very much larger than the radius of the observable universe. Equation (48) reveals that even if $k = \pm 1$, the observable universe can be arbitrarily close to flat if Ω_0 is sufficiently close to 1.

Note that inflation does not predict whether the universe is *finite* ($k = +1$) or *infinite* ($k \leq 0$), for reasons stated above and also because the observable universe is not regarded as representative of the whole. In the latter case, the RW metric (upon which our discussion is based) would not be applicable to the entire universe.

B. Kinds of Matter

The nature and distribution of matter are complex subjects, only partially understood. Broadly speaking, the principal sources of ρ_M are ordinary matter and other (*exotic*) matter. Electrons and atomic nuclei (made of protons and neutrons) are regarded as ordinary, whether they are separate or bound together in atoms. The universe has zero net charge, so electrons and protons are virtually identical in their abundance. (Charged particles are easily detected, and all other charged particles that we know of are short-lived.)

Protons and neutrons have nearly equal masses, and are ~1800 times as massive as electrons. Virtually all the mass of ordinary matter therefore resides in nuclei. Protons and neutrons are members of a larger group of particles called baryons, and cosmologists often refer to ordinary matter as *baryonic matter* (despite the fact that electrons are not baryons). All other baryons are very short-lived and rare, so any significant amount of other matter is *nonbaryonic*. We shall denote the densities of baryonic and nonbaryonic matter by ρ_B and ρ_{NB}, respectively. The total, present density of matter (and radiation) may now be expressed as $\rho_M \cong \rho_B + \rho_{NB}$, where the approximation is valid within a tiny fraction of 1%.

We have no evidence that stars and planets are constructed of anything but ordinary matter. It follows that any other matter whose density is a significant fraction of ρ_c must be electrically neutral, and also immune to the strong nuclear force (the *strong interaction*): its presence would otherwise be conspicuous in stars and planets. Such matter can interact only weakly with ordinary matter, through gravity and perhaps through the weak nuclear force (the *weak interaction*), and conceivably through some other weak force that remains to be identified.

Neutrinos are particles with the above properties. They are also abundant, having been copiously produced

during the early period of nucleosynthesis studied by Gamow and coworkers (*big bang nucleosynthesis*, or BBN). Their rest masses were long thought to be zero (a convoluted piece of jargon meaning that, like photons, they always travel at the speed of light). The expansion and cooling of the universe would have reduced the energy of such relic neutrinos to a negligible amount, as has happened with the relic photons that now make up the cosmic microwave background. Recent experiments have revealed, however, that neutrinos have small but nonzero rest masses. Even with very small masses, neutrinos are so abundant that their density ρ_ν could be a significant fraction of ρ_c. [The Greek letter ν (nu) is a standard symbol for neutrinos.]

There is gravitational evidence (to be described) for the existence of electrically neutral, weakly interacting particles *other* than neutrinos. No such particles have yet been observed in laboratories, so their identity and detailed properties remain unknown. Such matter is therefore called *exotic*, and its density will be denoted by ρ_X (however many varieties there may be). We then have $\rho_{NB} = \rho_\nu + \rho_X$. (Neutrinos are sometimes regarded as exotic matter, but the present distinction is useful.)

C. Amounts of Matter

Stars are made of baryonic matter, as are interstellar and intergalactic clouds of gas and dust. Bright stars (brighter than red dwarfs, the dimmest of main sequence stars) are easily seen and studied, and are found to contain only $\sim$0.5% of ρ_c. Clouds of gas and dust contain several times as much, but the amount has been difficult to measure. In fact our best estimate for the total amount of baryonic matter is based on the relative abundances of nuclei formed during the early period of BBN.

The fraction of primordial baryonic matter converted into ^{2}H (deuterium) during BBN depended strongly on the density of baryons at that early time, in a (presumably) known way. Intergalactic clouds of gas have never been affected by nucleosynthesis in stars, so they should be pure samples of the mixture produced by BBN. By studying light from distant quasars that has passed through intergalactic clouds, the relative abundance of deuterium was measured in 1996 by David R. Tytler, Scott Burles, and coworkers. On the basis of that and subsequent measurements, it is now believed that

$$\Omega_B = (0.019 \pm 0.001)/h_0^2 = 0.045 \pm 0.009 \qquad (49)$$

The amount of baryonic matter is therefore $\sim$9 times greater than appears in bright stars.

It has become apparent in recent decades that the universe contains a great deal of *dark* (nonluminous) *matter*. As indicated above, only $\sim 1/9$ of ordinary matter is in bright stars, and most of the remainder is cold (or dilute) enough to be dark. There is also *gravitational* evidence for $\sim$7 times as much dark matter as there is baryonic matter. The evidence resides primarily in the dynamics of galaxies and galactic clusters, which are held together by gravitational forces. The motions of stars in galaxies, and of galaxies in clusters, reveal the strength of gravity within them. The total amount of mass can then be inferred.

We first consider orbital motion around some large, central mass, such as the motion of planets around our sun. Under the simplifying assumption of circular orbits, Newton's laws of gravitation and motion (the latter being $\vec{F} = m\vec{a}$, with centripetal acceleration given by $a = v^2/r$) imply that

$$v = \sqrt{GM_\odot/r} \qquad (50)$$

where v and r denote the speed around, and radius of, the orbit, and $M_\odot$ denotes the mass of the sun.

We can test this model for the solar system by seeing whether orbital speeds and radii of the nine planets are related by the proportionality $v \propto 1/\sqrt{r}$, as predicted by Eq. (50). The orbits pass this test (making allowances for the fact that orbits are actually ellipses). Larger orbits are indeed traversed at smaller speeds, in the way predicted by Newton's laws. Since Eq. (50) implies that $M_\odot = v^2 r/G$, the sun's mass can be determined from the speed and radius of a *single* orbit, with confidence that all orbits yield the same answer (because $v \propto 1/\sqrt{r}$). This is how we know the value of $M_\odot$.

The Milky Way has a small nucleus that is very densely populated with stars (and may contain a black hole of $\sim 10^6 M_\odot$). Centered about this nucleus is a "nuclear bulge" that is also densely populated, with a radius of $\sim 10^4$ ly. The visible disk has a radius of $\sim 5 \times 10^4$ ly, and it becomes thinner and more sparsely populated as one approaches the edge. For the preceding reasons, it was long supposed that most of the galaxy's mass is less than 2×10^4 ly from the center. For stars in circular orbits outside this central mass, orbital speeds should then be a decreasing function of radius, roughly satisfying $v \propto 1/\sqrt{r}$ (as for planets in the solar system).

Beginning in the 1970s, Vera Rubin and others began systematic studies of orbital speeds, out to the edge of our galaxy and beyond (where a relatively small number of stars are found). To everyone's surprise, the speeds did *not* decrease with increasing distance. In fact the speeds at 5×10^4 ly (the visible edge) are $\sim$10% *greater* than at 2×10^4 ly. Furthermore, the speeds are even larger (by a small amount) at 6×10^4 ly, *beyond* the visible edge. Studies of this kind indicate that beyond $\sim 2 \times 10^4$ ly, the amount of mass within a distance r of the center is roughly proportional to r, out to 10^5 ly or so (twice the radius of the visible disk).

Studies of other spiral galaxies have revealed similar properties. The Milky Way and other spirals are now believed to have roughly spherical distributions of dark matter, containing most of the mass and extending far beyond the disk, with densities falling off (roughly) like $1/r^2$ beyond the nuclear bulge. Most of this dark matter is probably nonbaryonic (as will soon be explained).

The largest clusters contain thousands of galaxies, so it is plausible (but not certain) that little dark matter extends beyond their visible boundaries. There are three distinct ways to estimate the total masses of large clusters. The original method (first used by Fritz Zwicky in 1935) is based on the motions of galaxies within a cluster, from which the strength of gravity and total mass can be inferred. The second method exploits the presence of hot, X-ray emitting, intracluster gas. The high temperature presumably results from motion caused by gravity, so the total mass can be inferred from the spectrum of the X-rays. The third method is based on gravitational lensing. The light from more distant galaxies bends as it passes by (or through) a cluster, and detailed observations of this effect enable one to estimate the total mass. All three methods yield similar results.

The total luminosities of nearby clusters are easily determined, and the mass-to-luminosity ratios are about the same for all large clusters that have been studied. If one assumes this ratio to be representative of *all* galaxies (only $\sim$10% of which reside in clusters), then the average density of matter is readily obtained from the total luminosity of all galaxies within any large, representative region of space. This line of reasoning has been applied to the relevant data, yielding $\Omega_M = 0.19 \pm 0.06$. This value for Ω_M would be an underestimate, however, if the matter inside large clusters were more luminous than is typical of matter. This would be true if an appreciable amount of dark matter extends beyond the visible boundaries of large clusters, or if an atypically large fraction of baryonic matter inside of large clusters has formed bright stars. Both possibilities are plausible, and at least one appears to be the case.

In addition to emitting X-rays, the hot intracluster gas slightly alters (by scattering) the cosmic microwave radiation passing through it (the Sunyaev-Zel'dovich effect). Since only charged matter generates or affects radiation, the X-rays and slightly diminished CMB reaching earth both contain information about the amount of baryonic matter. The most *precise* inference from such studies is the *ratio* of baryonic to nonbaryonic mass in large clusters; namely $\rho_B/\rho_M = (0.075 \pm 0.002)/h_0^{3/2}$. Assuming this ratio to be typical of matter everywhere, it can be combined with Eq. (49) to obtain $\Omega_M = (0.253 \pm 0.02)/\sqrt{h_0} = 0.31 \pm 0.03$. This is regarded as the most reliable value for Ω_M. There is a long history of underestimating uncertainties in this field, however, so it seems appreciably safer to say that

$$\Omega_M = 0.31 \pm 0.06 = (6.9 \pm 1.9)\Omega_B \qquad (51)$$

In terms of mass, only $\sim 1/7$ of matter is baryonic, and only $\sim 1/60$ is in bright stars. Luminous matter is quite special, even less revealing than the tips of icebergs (which display $\sim 1/40$ of the whole, and reveal the composition of the remainder).

D. Nonbaryonic Matter

By mass, $\sim 6/7$ of matter is nonbaryonic. Neutrinos have been studied for decades, and are part of the "standard model" of elementary particles. As mentioned previously, there is recent evidence that neutrinos have nonzero rest masses (the Super-Kamiokande detection of neutrino oscillations, beyond the scope of this article). The present data place a lower limit on Ω_ν, namely $\Omega_\nu \geq 0.003$. Laboratory experiments have not ruled out the possibility that neutrinos account for *all* the nonbaryonic matter, but there are reasons to believe otherwise.

Indirect evidence for the nature of the nonbaryonic matter arises in computer models for the progressive concentration of matter into galaxies, clusters of galaxies, superclusters, sheets (or *walls*) of galaxies, and voids. All of these structures evolved as gravity progressively amplified small deviations from perfect uniformity, beginning at a very early time. (The original inhomogeneities are believed to have been microscopic quantum fluctuations that were magnified by inflation.)

When the temperature dropped below $\sim$3,000 K, the electrons and nuclei were able to remain together as electrically neutral atoms, forming a transparent gas. This happened $\sim$300,000 yr after the big bang, the *time of last scattering*. The photons then present have been traveling freely ever since, and now form the cosmic microwave background. Studies of the CMB reveal that deviations from perfect uniformity at the time of last scattering were on the order of $\delta\rho_M/\rho_M \sim 10^{-5}$. Those slight inhomogeneities have since been amplified by gravity into all the structures that we see today.

Computer models for the growth of structure require assumptions about the exotic dark matter. This matter is usually assumed to consist of stable elementary particles that are relics from a very early time. Because they are weakly interacting, only gravity should have affected them appreciably (gravity is the only weak force acting over large distances).

Since space was virtually transparent to exotic particles, their speeds would have played a key role in structure formation. High speeds would have inhibited their being captured by growing concentrations of matter,

unless those concentrations were unusually extended and massive. Computer simulations indicate that fast weakly-interacting particles would only have formed very large concentrations, of cluster and even supercluster size, whereas slow particles would first have formed smaller concentrations of galactic size. The scenario wherein concentrations of cluster size preceded the formation of galaxies is called "top down" growth of structure. If galaxies formed first and later clumped into clusters, the growth is called "bottom up."

Since light travels at a finite speed, astronomers see things as they were in the increasingly distant past as they look out to increasingly greater distances. The evidence is very clear that galaxies formed first, with only ~10% of them clumping later into clusters. Structure formation occurred from the bottom up. It follows that nonbaryonic particles were predominately slow. Of course "fast" and "slow" are relative terms, and the average speeds of *all* types of massive particles decreased as the universe expanded. These simple labels are useful in a brief discussion, however, and they have fairly definite meanings among workers in the field. In further jargon, fast nonradiating (uncharged) particles are referred to as "hot dark matter," and slow ones as "cold dark matter." Bottom-up structure formation implies that dark matter is predominately cold.

There are three species of neutrino (electron, muon, and tau-neutrinos). All three species should have reached thermal equilibrium with other matter and photons before the universe was one second old. (The early, high density of matter more than compensated for the weakness of neutrino interactions.) For this and other reasons, the present abundance of relic neutrinos is believed to be known, and Ω_ν can be expressed as

$$\Omega_\nu = \sum_i \frac{m_i c^2}{93.5 h_0^2\,\mathrm{eV}} = \sum_i \frac{(0.025 \pm 0.005) m_i c^2}{1\,\mathrm{eV}} \qquad (52)$$

where the index i refers to the three species (1 eV = 1 electronvolt $= 1.60 \times 10^{-19}$ joule).

If only *one* species had a nonzero mass, with $m_i c^2 \cong 39$ eV, the result would be $\Omega_\nu \cong 1$. (For comparison, the lightest particle of ordinary matter is the electron, with $m_e c^2 = 5.11 \times 10^5$ eV.) Since $\Omega_{NB} \cong 0.26$, the upper limit on neutrino mass-energies is on the order of 10 eV. It is the combination of weak interactions with very small masses that has frustrated efforts to determine the mass of even one species of neutrino. The Super-Kamiokande experiment is only sensitive to *differences* in mass between different species. The observed mass difference implies $\Omega_\nu \geq 0.003$, where the equality would hold if only one species had nonzero mass.

Because neutrinos are so light and were in thermal equilibrium at an early, hot time, they are the paradigm of "hot dark matter." They slowed as the universe expanded, but at any *given* temperature, the lightest particles were the fastest. Computer simulations rule out neutrinos as the predominate form of nonbaryonic matter. Recalling the lower limit discussed previously, we conclude that

$$0.003 \leq \Omega_\nu < \tfrac{1}{2}\Omega_{NB} \qquad (53)$$

The two most favored candidates for exotic matter are *axions* and *neutralinos*. These arise in speculative, but plausible, theories that go beyond the standard model of elementary particles. (Neutralinos occur in *supersymmetric* theories, which receive tentative support from a very recent experiment on the precession of muons.) For quite disparate reasons, both qualify as cold dark matter.

Axions are predicted to be much lighter than neutrinos. Unlike neutrinos, however, they would never have been in thermal equilibrium, and would *always* have moved slowly (for abstruse reasons beyond the scope of this article). In sharp contrast, neutralinos would have been in thermal equilibrium when the universe was very young and hot. Neutralinos (if they exist) are much more massive than protons and neutrons, however, and their *large masses* would have rendered them (relatively) slow-moving at any given temperature. Other candidates for exotic "cold dark matter" have been proposed, but will not be mentioned here.

E. Vacuum Energy

In 1998, two teams of observers, one led by Saul Perlmutter and the other by Brian P. Schmidt, announced that the cosmic expansion appears to be *accelerating*. This revolutionary claim was based on observations of type SNe Ia over a wide range of distances, extending out to $\sim 6 \times 10^9$ ly. Distances approximately twice as great have since been explored using SNe Ia as standard candles.

Redshifts are usually described in terms of a red-shift parameter $z = \lambda_0/\lambda_e - 1$, where λ_0 denotes the observed wavelength and λ_e the emitted wavelength. With spacetime described by a RW metric, it is readily shown that $z = a(t_0)/a(t_e) - 1$, where t_0 and t_e denote the (present) time of observation and the earlier time of emission, respectively. Supernovae of type Ia have been observed with redshifts as large as $z \approx 1$, corresponding to $a(t_0) \approx 2a(t_e)$. Such light was emitted when the universe was only half its present size, and roughly half its present age.

Since the intrinsic luminosity of SNe Ia is known, the apparent brightness reveals the distance of the source, from which the time of emission can be inferred (subject to possible corrections for spatial curvature). The evolution

over time of $a(t)$ has been deduced in this way, and indicates an accelerating expansion. Since ρ_M has decreased over time while ρ_V remained constant, they have affected $a(t)$ in different ways. Thus Ω_M and Ω_V can both be deduced from detailed knowledge of $a(t)$. Present SNe Ia data indicate that $\Omega_M \approx 0.3$ and $\Omega_V \approx 0.7$, with uncertainties in both cases on the order of ± 0.1.

The CMB radiation provides a snapshot of the universe at the time of last scattering, ~300,000 yr after the big bang. The CMB reaching us now has come from the outermost limits of the observable universe, and has a redshift $z \sim 1100$. It has an extremely uniform temperature of $T = (2.728 \pm 0.002)$K in all directions, except for tiny deviations on the order of $\delta T/T \sim 10^{-5}$ over small regions of the sky. These deviations are manifestations of slight inhomogeneities in density of order $\delta\rho_M/\rho_M \sim 10^{-5}$ at the time of last scattering. Gravity produced these slight concentrations of matter from much smaller ones at earlier times, in ways that are believed to be well understood. In particular, the *spectrum* (probability distribution) of different *sizes* of inhomogeneity is calculable in terms of much earlier conditions.

The angle subtended by an object of known size at a known distance depends on the curvature of space, as discussed in Section III.C. (For example, the circumference of a circle differs from $2\pi r$ in a curved space.) In the present context, the "objects of known size" were the *largest* concentrations of matter, whose diameters were governed by the distance that density waves could have traveled since the big bang (the *sound horizon* at the time of last scattering).

The speed of density waves is a well-understood function of the medium through which they travel, and the sound horizon at last scattering has been calculated with (presumably) good accuracy. The angular diameters of the largest concentrations of matter, manifested by $\delta T/T$, therefore reveal the curvature of space. Within observational uncertainties, the data confirm the foremost prediction of standard inflation, namely that *space is very close to flat*. There is no evidence of curvature. The CMB data indicate that space is flat or, if curved, that $a_0 \geq 3c/H_0 \sim 45$ billion light-years. Equation (48) then implies that

$$\Omega_0 = 1.0 \pm 0.1 \tag{54}$$

From Eq. (51) and the SNe Ia result cited above, it follows that

$$\Omega_V = 0.69 \pm 0.12 \tag{55}$$

Spatial flatness and vacuum energy are the most important and dramatic discoveries in observational cosmology since the CMB (1964).

There has been some speculation that the mass-energy density filling all of space might not be *strictly* constant, but change slowly over time. (Such an energy field is sometimes called "quintessence.") Denoting the density of such a hypothetical field by ρ_Q, observations would imply a large, negative pressure, $P_Q < -0.6c^2\rho_Q$. Present data are wholly consistent with the simplest possibility, however, the one described here in detail: a constant ρ_V with $P_V = -c^2\rho_V$, perfectly mimicking a cosmological constant.

F. Age and Future of the Cosmos

We now consider the evolution over time of a flat universe, under the simplifying assumption that $P_M = 0$ (an excellent approximation since the first 10 million years or so). Equations (22), (27), (42), and (43) then imply

$$t = \frac{2}{3H_0}\int_0^{x(t)} \frac{dx'}{\sqrt{\Omega_V x'^2 + \Omega_M}} \tag{56}$$

where t denotes the time since the big bang, $x(t) \equiv [a(t)/a_0]^{3/2}$, denotes the present value of the cosmic scale factor $a(t)$, and Ω_M and Ω_V are *present* values, with $\Omega_M + \Omega_V = 1$.

The scale of time is set by H_0, where $t_H \equiv 1/H_0$ is called the *Hubble time*. To understand its significance, suppose that every distant galaxy had always been moving away from us at the same speed it has now. For motion at constant speed v, the time required to reach a distance d is $t = d/v$. According to Hubble's law, $v = H_0 d$ [Eq. (17)], so the time required for *every* galaxy to have reached its present distance from us would be $d/v = 1/H_0 = t_H$. Thus t_H would be the age of the universe if galactic speeds had never been affected by gravity. (It is sometimes called the "Hubble age.") With H_0 given by Eqs. (18) and (19),

$$t_H \equiv \frac{1}{H_0} = (15.1 \pm 2.4) \times 10^9 \text{ yr} \tag{57}$$

Denoting the *actual* age of the universe by t_0 (the *present* time), we note that $x(t_0) = 1$. For the simple case where $\Omega_M = 1$ (and $\Omega_V = 0$), Eq. (56) yields $t_0 = \frac{2}{3}t_H \sim 10 \times 10^9$ yr (less than the age of the oldest stars). For $\Omega_V > 0$, Eq. (56) implies

$$\frac{t}{t_H} = \frac{2}{3\sqrt{\Omega_V}} \ln\left(\frac{\sqrt{\Omega_V(a/a_0)^3} + \sqrt{\Omega_V(a/a_0)^3 + \Omega_M}}{\sqrt{\Omega_M}}\right) \tag{58}$$

expressing t/t_H as a function of a/a_0. Using the observed value $\Omega_M = 0.31 \pm 0.06$ (with $\Omega_V = 1 - \Omega_M$) and setting $a = a_0^4$, we obtain the present age

$$t_0 = (0.955 \pm 0.053)t_H = (14.4 \pm 2.4) \times 10^9 \text{ yr} \tag{59}$$

The oldest stars are believed to be $(12 \pm 2) \times 10^9$ years old, and they are expected to have formed $\sim 2 \times 10^9$ yr after

the big bang. The discovery of substantial vacuum energy (or something very similar) has increased the theoretical age of the universe comfortably beyond that of the oldest stars, thereby resolving a long-standing puzzle.

The time when the cosmic expansion began accelerating is readily determined. Replacing Λ in Eq. (23) with the ρ_V and P_V of Eqs. (42) and (43), we obtain

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho_M - 2\rho_V + \frac{3P_M}{c^2}\right) \tag{60}$$

With $P_M \cong 0$, the acceleration ($\ddot{a} > 0$) began when ρ_M fell below ρ_V. Using Eq. (27) for ρ_M, this happened when $(a/a_0)^3 = \Omega_M/2\Omega_V \simeq 0.22$. Denoting the corresponding time by t_a, Eq. (58) yields $t_a \simeq 0.52 t_H \simeq 0.55 t_0$.

When a/a_0 reaches 2, ρ_M will have shrunk (by dilution) to ~6% of ρ_V. This will happen at $t \simeq 1.7 t_H \simeq 1.8 t_0$. From roughly that time onward, the universe will grow at an exponential rate, as described by Eqs. (29)–(31) and (42). The doubling time will be

$$t_d = \frac{\ln 2}{\sqrt{\Omega_V}} t_H \approx 13 \times 10^9 \text{ yr} \tag{61}$$

The ratio $a(t)/a_0$ is displayed in Fig. 1, which should be reliable for $t/t_0 \geq 10^{-3}$.

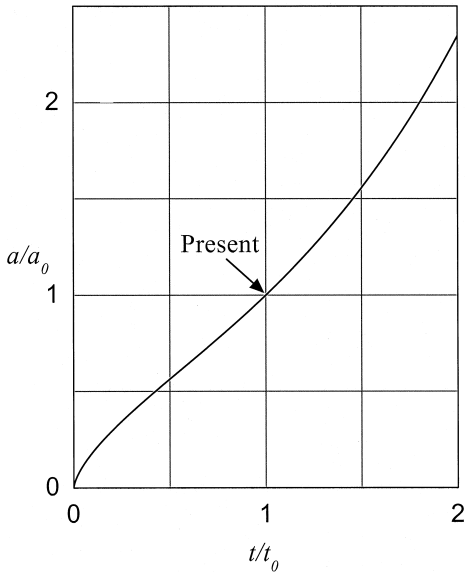


FIGURE 1 Cosmic expansion: the distance between any pair of comoving objects is proportional to the cosmic scale factor $a(t)$. Its present value is denoted by a_0, and the present time by t_0. For example, two objects were half as far apart as now when $a/a_0 = 1/2$, which occurred at $t \approx 0.4 t_0$. They will be twice as far apart as now when $a/a_0 = 2$, which happen at $t \approx 1.8 t_0$. The graph is only valid for $t/t_0 \geq 10^{-3}$ (after inflation).

VI. HORIZON AND FLATNESS PROBLEMS

A. The Horizon Problem

The greatest distance that light could (in principle) have traveled since the universe began is called the *horizon distance*. Since no causal influence travels faster than light, any two regions separated by more than the horizon distance could not have influenced each other in any way. Observations reveal the universe to be flat (or nearly so) out to the source of the CMB now reaching us. We shall assume for simplicity that space is precisely flat ($k = 0$, $\Omega = 1$).

Every light path is a "null geodesic," with $(ds)^2 = 0$ in Eq. (7). With $k = 0$, $c\,dt = \pm a(t)\,dr$. Let us place ourselves at $r = 0$, and determine the proper distance $r_p = a_0 r$ to an object whose light is now reaching us with redshift $z = (a_0/a) - 1$. Since $dt = da/\dot{a}$ and $\dot{a} = aH$,

$$r_p = a_0 c \int_a^{a_0} \frac{da'}{a'^2 H(a')} \tag{62}$$

where $H(a)$ is given by Eq. (26) in terms of $\rho(a)$. We shall initially assume the pressure of matter and radiation to be negligible, in which case $\rho \propto 1/a^3$. Equation (26) then implies

$$H^2 = H_0^2\left[\Omega_M\left(\frac{a_0}{a}\right)^3 + \Omega_V\right] \tag{63}$$

where Λ is being represented by ρ_V as in Eq. (42) (and, as before, Ω_M and Ω_V denote *present* values). Equations (62) and (63) imply

$$\begin{aligned} r_p(z) &= \frac{c}{H_0} f(z), \\ f(z) &\equiv 2\int_{1/\sqrt{z+1}}^{1} \frac{dy}{\sqrt{\Omega_M + \Omega_V y^6}} \end{aligned} \tag{64}$$

where $y \equiv \sqrt{a/a_0}$.

The integral defining $f(z)$ cannot be evaluated in terms of elementary functions. For small z, however, one can define $x \equiv 1 - y$, expand the integrand in powers of x, and thereby obtain

$$f(z) = z\left[1 - \tfrac{3}{4}\Omega_M z + \mathbb{O}(z^2)\right] \tag{65}$$

where $\mathbb{O}(z^2)$ denotes corrections of magnitude $\sim z^2$. The two terms displayed in Eq. (65) are accurate within 1% for $z \leq 1$ and $\Omega_M \sim 0.3$. (Despite appearances, Eq. (65) involves Ω_V, through our assumption that $\Omega_V = 1 - \Omega_M$.)

Next we note that $f(\infty) - f(z)$ is twice the integral over y from 0 to $1/\sqrt{z+1}$. For large z, y remains small,

and the integrand can be expanded as a power series in y. We thereby obtain

$$f(z) = f(\infty) - \frac{2}{\sqrt{\Omega_M(z+1)}}\left\{1 - O\left[\frac{\Omega_V}{14\Omega_M(z+1)^3}\right]\right\} \tag{66}$$

where the correction shown is *exact* below order $1/(z+1)^6$. If one keeps this correction term, Eq. (66) is valid within 0.2% for $z \geq 1$ and $\Omega_M \sim 0.3$. The value of $f(\infty)$ is readily determined by numerical integration of Eq. (64).

For times earlier that $t_{EQ} \sim 3 \times 10^5$ yr (the time of *equality* between low and high-pressure ρ_M), radiation and highly relativistic particles contribute more to ρ_M than does matter with negligible pressure. (It appears to be coincidental that this is close to the time of last scattering, when the universe became transparent and the CMB set forth on its endless journey.) The transition is, of course, a smooth one. For simplicity, however, we assume that $P_M = 0$ for $t > t_{EQ}$, in which case Eq. (64) would be valid. For $t < t_{EQ}$, we assume that $P_M = c^2\rho_M/3$, which is the pressure exerted by radiation and highly relativistic particles.

We also impose continuity on a and $\dot{a}$ at t_{EQ} or, equivalently, on a and H. Finally, the CMB was emitted at $t \sim t_{EQ}$, and has $z_{CMB} \sim 1100$. The value of z_{EQ} can only be estimated, but it seems certain that $z_{EQ} \geq 100$. It follows that $\rho_V \leq 10^{-6}\rho_M$ at t_{EQ}, and we shall neglect it for $t \leq t_{EQ}$.

For $P_M = c^2\rho_M/3$, Eq. (24) implies that $\rho_M \propto 1/a^4$. It then follows from Eq. (26) that $2a\dot{a} = da^2/dt$ is constant for $t \leq t_{EQ}$. We are presently concerned with the standard big bang (without inflation), wherein $\rho \propto 1/a^4$ all the way back to $t = 0$ (when $a = 0$). We therefore have

$$t \leq t_{EQ}: \qquad \left(\frac{a}{a_{EQ}}\right)^2 = \frac{t}{t_{EQ}}, \qquad H \equiv \frac{\dot{a}}{a} = \frac{1}{2t} \tag{67}$$

where $a_{EQ} \equiv a(t_{EQ})$.

Equation (67) holds at early times *regardless* of what the present universe is like, an extraordinary fact of great simplicity and usefulness.

Equation (67) implies that a^2H is constant for $t \leq t_{EQ}$, so the integral in Eq. (60) is trivial for $t \leq t_{EQ}$. Imposing continuity on H, Eq. (63) can be used to express H_{EQ} in terms of H_0, Ω_M, and a_0 (neglecting Ω_V at t_{EQ}). We thereby obtain

$$z \geq z_{EQ}:$$

$$r_p(z) = \frac{c}{H_0}\left[f(\infty) - \frac{1}{\sqrt{\Omega_M(z_{EQ}+1)}}\frac{(z+z_{EQ}+2)}{z+1}\right] \tag{68}$$

For $\Omega_M = 0.31$ (with $\Omega_V = 0.69$), a numerical integration yields $f(\infty) = 3.26$. As z_{EQ} ranges from 100 to the (unrealistic) value of ∞, $r_p(\infty)$, varies by less than 6%, and $r_p(z_{EQ})$ differs from $r_p(\infty)$ by less than 6%. Note that $r_p(\infty)$ is the present *standard horizon* distance $d_{SH}(t_0)$, i.e., the horizon distance if there had been no inflation. For $z_{EQ} \sim 1000$, we obtain

$$d_{SH}(t_0) = r_p(\infty) \approx 3.2\frac{c}{H_0} \approx 48 \times 10^9 \text{ ly} \tag{69}$$

A 20% uncertainty in Ω_M would only render the factor of 3.2 uncertain by ~10%. It may seem puzzling that $d_{SH}(t_0) \sim 3 \times$ (age of universe), but the cosmic expansion has been moving the source away from us while the light traveled toward us. This effect has been amplified by acceleration of the expansion.

The distance to the source of the CMB is of particular interest, for reasons that will become apparent. Since $z_{CMB} \approx 1100$, the percentage difference between $r_p(z_{CMB})$ and $d_{SH}(t_0)$ is small, and we have seen that it varies little with any choice of $z_{EQ} \geq 100$. For $z_{EQ} = z_{CMB} = 1100$, Eq. (68) yields $r_p(z_{CMB}) \approx 3.1(c/H_0) \approx 47 \times 10^9$ ly. The CMB sources in opposite directions from us are therefore separated by $\sim 90 \times 10^9$ ly. Proper distances between comoving points scale like $a/a_0 = z + 1$, so sources in opposite directions were $\sim 8 \times 10^7$ ly apart when they emitted the CMB we now observe. This is a fact of extreme interest, as we shall see.

The CMB reaching us now began its journey at $t_{LS} \approx 300{,}000$ yr, where t_{LS} denotes the time of last scattering. We wish to determine $d_{SH}(t_{LS})$, the (standard) horizon distance at that time. We consider light emitted from $r = 0$ at $t = 0$, and consider the proper distance $d_H(t) = a(t)r(t)$ it has traveled by time t:

$$d_H(t) = ca(t)\int_0^t \frac{dt'}{a(t')} \tag{70}$$

Using Eq. (67), we obtain

$$t \leq t_{EQ}: \qquad d_{SH}(t) = 2ct, \tag{71}$$

a wonderfully simple result. If $t_{LS} \leq t_{EQ}$, Eq. (71) would imply that $d_{SH}(t_{LS}) \approx 600{,}000$ ly.

For a considerable time later than t_{EQ}, a/a_0 remained small enough that $\rho_V \ll \rho_M$. (For example, with $\Omega_V/\Omega_M \approx 7/3$, $\rho_V \leq 10^{-2}\rho_M$ for $a/a_0 \leq \frac{1}{6}$. This is satisfied for $t \leq 0.7t_0 \approx 10^{10}$ year, as may be seen from Fig. 1.) With $\rho \propto 1/a^3$, Eq. (63) implies that $da^{3/2}/dt$ is constant, hence that $a^{3/2} = \alpha t + \beta$ for constant α and β. Continuity of a and H at t_{EQ} determines these constants, with the results

$$t \geq t_{EQ},\ \rho_V = 0: \quad a(t) = a_{EQ}\left(\frac{3t + t_{EQ}}{4t_{EQ}}\right)^{2/3}, \tag{72}$$

$$d_{SH}(t) = c(3t + t_{EQ})\left\{1 - \left[\frac{t_{EQ}}{2(3t + t_{EQ})}\right]^{1/3}\right\} \tag{73}$$

Now recall that if $t_{LS} \le t_{EQ}$, Eq. (71) yields $d_{SH}(t_{LS}) \approx$ 600,000 ly. If $t_{LS} \ge t_{EQ}$, Eq. (73) yields $d_{SH} \approx 630{,}000$ ly (within 5%), for any $t_{EQ} \ge 50{,}000$ yr (a very conservative lower bound). In either case, $d_{SH}(t_{LS})$ was less than 1% of 80 million ly, the distance between sources of the CMB in opposite directions from us when the CMB was omitted. This is the *horizon problem* described in Section II.A.

Next consider a simple inflationary model. Let t_b and t_e denote the times when inflation begins and ends, respectively, with $t_e \ll 1$ sec. For $t \le t_b$, $(a/a_b)^2 = t/t_b$, with $H = 1/2t$ as in Eq. (67). For $t_b \le t \le t_e$, $a = a_b \exp\Gamma(t - t_b)$, with $H = \Gamma$, a constant [recall Eqs. (29) and (30)]. Continuity of H implies $\Gamma = 1/2t_b$. The doubling time t_d is related to Γ by $\Gamma = \ln 2/t_d$, so $t_d = (2 \ln 2)t_b$.

Inflation ends at t_e with $a_e = a_b \exp\Gamma \Delta t$ and $H = \Gamma$, where $\Delta t \equiv t_e - t_b$. The density then reverts to its earlier behavior $\rho \propto 1/a^4$, and Eq. (26) again implies that da^2/dt is constant. The general solution is $a^2 = \xi t + \eta$, where ξ and η are fixed by continuity of a and H at t_e. One readily finds that $a^2 = a_e^2[2\Gamma(t - t_e) + 1]$, but a more elegant solution is at hand.

From the end of inflation onward, the universe evolves precisely like a model without inflation, but which had the same conditions at t_e. Denoting time in the latter model by t', $H = 1/2t'$, and $H = \Gamma$ at $t'_e = 1/2\Gamma$. The scale factor was zero at $t' = 0$ in this model, so $(a/a_e)^2 = t'/t'_e$. Since $t'_e = 1/2\Gamma$ when $t = t_e$, the relation is $t' = (t - t_e) + t_d/\ln 4$.

Using $dr = c\,dt/a(t)$, it is straightforward to calculate $r(t_b)$, $r(t_e) - r(t_b)$, and $r(t') - r(t'_e)$. We thereby obtain

$$\frac{d_H(t')}{2ct'} = (e^{\Gamma \Delta t} - 1)\sqrt{\frac{2t_d}{t' \ln 2}} + 1 \qquad (74)$$

(presuming the *actual* horizon distance d_H is given by the inflationary model). The horizon problem is solved if $d_H(t_{LS}) \gg r_p(z_{CMB})/z_{CMB}$. This inequality needs to be a strong one, in order to explain how the CMB arriving from opposite directions can be so *precisely* similar. Combined with Eq. (74), this requires inflation by a factor

$$e^{\Gamma \Delta t} \gg \frac{r_p(z_{CMB})}{z_{CMB} c\sqrt{t_d t_{LS}}} \qquad (75)$$

where factors of order unity have been omitted. For $t_d \sim 10^{-37}$ sec,

$$e^{\Gamma \Delta t} \gg 10^{28} \qquad (76)$$

is required. Resolution of the horizon problem is displayed in Figure 2.

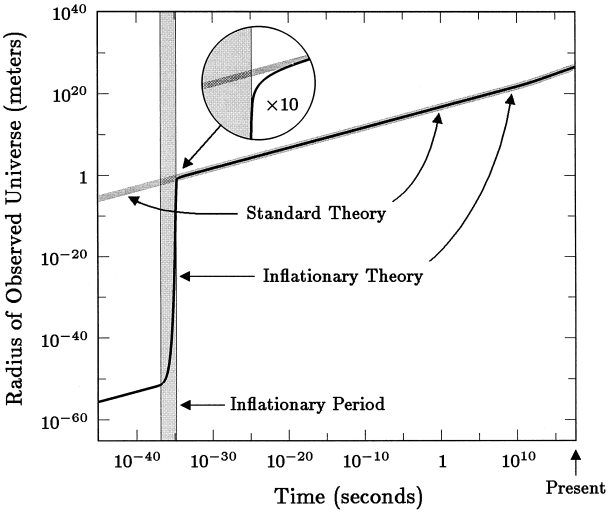


FIGURE 2 Solution to the horizon problem: the size of the observed universe in the standard and inflationary theories. The vertical axis shows the radius of the region that evolves to become the presently observed universe, and the horizontal axis shows the time. In the inflationary theory, the universe is filled with a false vacuum during the inflationary period, marked on the graph as a vertical gray band. The gray line describing the standard hot big bang theory is drawn slightly thicker than the black line for the inflationary theory, so that one can see clearly that the two lines coincide once inflation has ended. Before inflation, however, the size of the universe in the inflationary theory is far smaller than in the standard theory, allowing the observed universe to come to a uniform temperature in the time available. The inset, magnified by a factor of 10, shows that the rapid expansion of inflation slows quickly but smoothly once the false vacuum has decayed. (The numerical values shown for the inflationary theory are not reliable, but are intended only to illustrate how inflation can work. The precise numbers are highly uncertain, since they depend on the unknown details of grand unified theories.) [From "The Inflationary Universe" by Alan Guth. Copyright © 1997 by Alan Guth. Reprinted by permission of Perseus Books Publishers, a member of Perseus Books, L.L.C.]

B. The Flatness Problem

Let us now define

$$\Omega(t) \equiv \rho(t)/\rho_c(t) \qquad (77)$$

where $\rho_c(t)$ is the critical density given by Eq. (45) with H_0 replaced by $H(t) = \dot{a}/a$. As emphasized by Robert Dicke and P. James E. Peebles in 1979, any early deviation of Ω from unity would have grown very rapidly with the passage of time, as Figure 3 displays. Replacing Λ by the equivalent ρ_V, Eq. (26) implies

$$\frac{\Omega - 1}{\Omega} = \frac{3kc^2}{8\pi G\rho a^2} \le \frac{3kc^2}{8\pi G\rho_M a^2} \qquad (78)$$

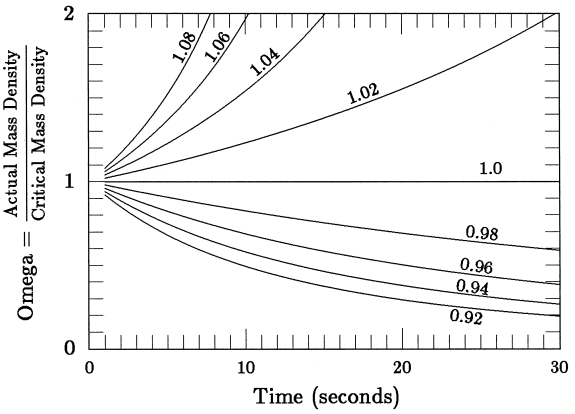


FIGURE 3 The evolution of omega for the first 30 sec. The curves start at one second after the big bang, and each curve represents a different starting value of omega, as indicated by the numbers shown on the graph. [From "The Inflationary Universe" by Alan Guth. Copyright © 1997 by Alan Guth. Reprinted by permission of Perseus Books Publishers, a member of Perseus Books, L.L.C.]

where the inequality holds for $\rho_V \geq 0$. Since $\rho = (\Omega/\Omega_M)\rho_M$,

$$\frac{\Omega_0 - 1}{\Omega_0} = \left(\frac{\Omega_{M0}}{\Omega_0}\right)\frac{3kc^2}{8\pi G\rho_{M0}a_0^2} \tag{79}$$

where the subscript "0" denotes present values.

The pressure P_M of matter (and radiation) became negligible when the universe was $\sim 10^6$ years old. Since that early time, $\rho_M \simeq \rho_{M0}(a_0/a)^3$ [as in Eq. (27)]. Equations (78) and (79) therefore imply

$$\frac{\Omega - 1}{\Omega} \leq \frac{\Omega_0 - 1}{\Omega_{M0}}\left(\frac{a}{a_0}\right) \tag{80}$$

which actually holds for *all* times since inflation, because the left side decreases even more rapidly than the right side as one goes back into the period when $P_M > 0$. As explained in Section V, it appears that $\Omega_{M0} \geq 0.2$ and $|\Omega_0 - 1| \leq 0.2$, in which case

$$\frac{|\Omega - 1|}{\Omega} \leq \frac{a}{a_0} \tag{81}$$

for all times since inflation. The CMB began its journey at $t_{LS} \approx 300{,}000$ yr, and now has $z_{CMB} \approx 1100$. It follows that $|\Omega_{LS} - 1| < 10^{-3}$.

We do not receive any light emitted before t_{LS}. To determine Ω at earlier times, we must study the time-dependence of the scale parameter a. We begin by noting that Eq. (26) (with $\Lambda = 0$) can be written as $H^2 = \Omega H^2 - k(c/a)^2$, which implies that $k(c/a)^2 = [(\Omega - 1)/\Omega] \times \Omega H^2$. The k-term in Eq. (26) is therefore negligible when $|\Omega - 1| \ll 1$, and we shall drop it for $t \leq t_{LS}$.

For all times earlier than t_{LS}, ρ_V was completely negligible. It is also known that for times earlier than $t_{ER} \sim 40{,}000$ yr (the *end* of the *radiation* era), radiation and/or relativistic particles dominated ρ_M, exerting a pressure $P_M \simeq \rho_M c^2/3$. Thus $\rho \simeq \rho_{ER}(a_{ER}/a)^4$, and Eq. (78) implies

$$t \leq t_{ER}: \qquad \frac{\Omega - 1}{\Omega} = \frac{\Omega_{ER} - 1}{\Omega_{ER}}\left(\frac{a}{a_{ER}}\right)^2 \tag{82}$$

We are here considering the standard model (without inflation), so $a^2 \propto t$.

Neither matter nor radiation dominated for times between t_{LS} and t_{ER}. Since $t_{ER} < t_{LS}$, however, $|\Omega_{ER} - 1| < |\Omega_{LS} - 1| < 10^{-3}$. Equation (82) then implies

$$t \leq t_{ER}: \qquad |\Omega - 1| < 10^{-3}\frac{t}{t_{ER}} \tag{83}$$

Since $t_{ER} = 1.3 \times 10^{12}$ sec, it follows that $|\Omega - 1| < 10^{-15}$ at $t = 1$ sec.

It is striking that Ω was so close to unity at 1 sec, but there is no physical significance to the time of 1 sec. Had there been no inflation, the inequality (83) should have been valid all the way back to the Planck time $t_P \equiv L_P/c \sim 10^{-43}$ sec. Quantum gravity would have been significant at earlier times, and is only partially understood, but Eq. (83) implies

$$t = t_{Pl}: \qquad |\Omega - 1| < 10^{-58} \tag{84}$$

We have no reason to suppose that quantum gravity would lead to a value for Ω near unity, let alone so *extremely* close. The present age of the universe is $\sim 10^{60} \times t_{Pl}$, and it is a profound mystery why Ω now differs from unity (if at all) by less than 20%. This is the "flatness problem."

The Hubble parameter $H \equiv \dot{a}/a$ remains constant during any period of exponential growth [recall Eqs. (29) and (30)]. If $a(t)$ increased by a factor of 10^N during a period of exponential growth, a generalization of Eq. (48) to arbitrary times would imply that $|\Omega - 1|$ *decreased* by a factor of 10^{2N}. To remove the necessity for Ω to have been close to unity at the Planck time, inflation by a factor $\geq 10^{30}$ is required. It seems unlikely that such growth would have been the *bare minimum* required to explain why Ω is near unity after $10^{60} \times t_{Pl}$, so standard inflation predicts growth by *more* than 30 powers of 10 (already required to solve the horizon problem), probably *many* more. It would follow that Ω_0 is extremely close to unity, which standard inflation predicts.

VII. UNIFIED THEORIES AND HIGGS FIELDS

A. Quantum Theory of Forces

In quantum field theory (QFT), forces between particles result from the exchange of *virtual quanta*, i.e., transient,

particlelike manifestations of some field that gives rise to the force. Electromagnetic forces arise from the exchange of virtual photons, where *photon* is the name given to a quantum or energy packet of the electromagnetic field. The weak nuclear force results from the exchange of *weak vector bosons*, of which there are three kinds called W^+, W^-, and Z. The strong nuclear force results, at the most fundamental level, from the exchange of eight *colored gluons* (a whimsical, potentially misleading name in that such particles are without visible qualities; *color* refers here to a technical property). From the viewpoint of QFT, gravitation may be regarded as resulting from the exchange of virtual *gravitons*. The four types of force just enumerated encompass all of the fundamental interactions currently known between elementary particles.

It is well-known that a major goal of Einstein's later work was to construct a unified field theory in which two of the existing categories of force would be understood as different aspects of a single, more fundamental force. Einstein failed in this attempt; in retrospect, the experimental knowledge of elementary particles and their interactions was too fragmentary at the time of his death (1955) for him to have had any real chance of success. The intellectual heirs of Einstein have kept his goal of unification high on their agendas, however. With the benefit of continual, sometimes dramatic experimental advances, and mathematical insights from a substantial number of theorists, considerable progress in unification has now been achieved. A wholly unexpected by-product of contemporary unified theories is the possibility of cosmic inflation. To understand how this came about, let us briefly examine what is meant by a unification of forces.

Each of the four categories of force enumerated above has several characteristics that, taken together, serve to distinguish it from the other types of force. One obvious characteristic is the way in which the force between two particles depends on the distance between them. For example, electrostatic and (Newtonian) gravitational forces reach out to arbitrarily large distances, while decreasing in strength like the inverse square of the distance. Forces of this nature (there seem to be only the two just named) are said to be *long range*. The weak force, in contrast, is effectively zero beyond a very short distance ($\sim 2.5 \times 10^{-16}$ cm). The strong force between protons and neutrons is effectively zero beyond a short range of $\sim 1.4 \times 10^{-13}$ cm, but is known to be a residual effect of the more fundamental force between the quarks (elementary, subnuclear particles) of which they are made. It is the force between quarks that is mediated by colored gluons (in a complicated way with the remarkable property that, under special conditions, the force between quarks is virtually independent of the distance between them).

The range of a force is intimately related to the mass of the virtual particle whose exchange gives rise to the force. Let us consider the interaction of particles A and B through exchange of virtual particle P. For the sake of definiteness, suppose that particle P is emitted by A and later absorbed by B, with momentum transmitted from A to B in the process.

The emission of particle P appears to violate energy conservation, for P did not exist until it was emitted. This process is permitted, however, by the time-energy uncertainty principle of quantum theory: energy conservation may be temporarily violated by an amount ΔE for a duration Δt, provided that $\Delta E \times \Delta t \approx \hbar$ (where $\hbar$ denotes Planck's constant divided by 2π). If particle P has a rest-mass m_p, then $\Delta E \geq m_p c^2$, so the duration is limited to $\Delta t \leq \hbar / m_p c^2$. The farthest that particle P can travel during this time is $c\Delta t \leq \hbar / m_p c$. This limits the range of the resulting force to a distance of roughly $\hbar / m_p c$ (called the *Compton wavelength* of particle P).

The preceding discussion is somewhat heuristic, but is essentially correct. If a particle being exchanged has $m > 0$, then the resulting force drops rapidly to zero for distances exceeding its Compton wavelength; conversely, long-range forces (i.e., electromagnetism and gravitation) can only result from exchange of "massless" particles. (Photons and gravitons are said to be massless, meaning zero rest-mass. Calling them "massless" is a linguistic convention, because such particles always travel at the speed of light and hence defy any measurement of mass when at rest. They may have arbitrarily little energy, however, and they behave as theory would lead one to expect of a particle with zero rest-mass.)

There are other distinguishing features of the four types of force, such as the characteristic strengths with which they act, the types of particles on which they act, and the force direction. We shall not pursue these details further here, but simply remark that two (or more) types of force would be regarded as unified if a single equation (or set of equations) were found, based on some unifying principle, that described all the forces in question. Electricity and magnetism were once regarded as separate types of force, but were united into a single theory by James Clerk Maxwell in 1864 through his equations which show their interdependence. From the viewpoint of QFT, electromagnetism is a single theory because both electrical and magnetic forces result from the exchange of photons; forces are currently identified in terms of the virtual quanta or particles that give rise to them.

B. Electroweak Unification via the Higgs Mechanism

When seeking to unify forces that were previously regarded as being of different types, it is natural to exploit any features that the forces may have in common. Einstein hoped to unify electromagnetism and gravitation,

in considerable measure because both forces shared the striking feature of being long-range. As we have noted, Einstein's efforts were not successful. In the 1960s, however, substantial thought was given to the possibility of unifying the electromagnetic and *weak* interactions.

The possibility of unifying the weak and electromagnetic interactions was suggested by the fact that both interactions affected all electrically charged particles (albeit in distinctly different ways), and by the suspicion that weak interactions might result from exchange of a hypothetical set of particles called "weak vector bosons." (*Vector bosons* are particles with intrinsic angular momentum, or "spin," equal to $\hbar$, like the photon.)

The electromagnetic and weak forces would be regarded as unified if an intimate relationship could be found between the photon and weak vector bosons, and between the strengths of the respective interactions. (More specifically, in regrettably technical terms, it was hoped that the photon and an uncharged weak vector boson might transform into one another under some symmetry group of the fundamental interactions, and that coupling strengths might be governed by eigenvalues of the group generators. The symmetry group would then constitute a unifying principle.)

The effort to unify these interactions faced at least two obstacles, however. The most obvious was that photons are massless, whereas the weak interaction was known to be of very short range. Thus weak vector bosons were necessrily very massive, which seemed to preclude the intimate type of relationship with photons necessary for a unified theory. A second potential difficulty was that while experiments had suggested (without proving) the existence of electrically charged weak bosons, now known as the W^+ and W^-, unification required a third, electrically neutral weak boson as a candidate for (hypothetical) transformations with the neutral photon. There were, however, no experimental indications that such a neutral weak boson (the Z) existed.

Building on the work of several previous investigators, Steven Weinberg (in 1967) and Abdus Salam (independently in 1968) proposed an ingenious unification of the electromagnetic and weak interactions. The feature of principle interest to us here was the remarkable method for unifying the massless photon with massive weak bosons. The obstacle of different masses was overcome by positing that, at the most fundamental level, there was no mass difference. Like the photon, the weak vector bosons were assumed to have no *intrinsic* mass. How, then, could the theory account for the short range of the weak force and corresponding large effective masses of the weak bosons? The theory did so by utilizing an ingenious possibility discovered by Peter Higgs in 1964, called the *Higgs mechanism* for (effective) mass generation.

A QFT is defined by the choice of fields appearing in it and by the *Lagrangian* chosen for it. The Lagrangian is a function of the fields (and their derivatives) that specifies the kinetic and potential energies of the fields, and of the particles associated with them (the *field quanta*). In more general terms, the Lagrangian specifies all intrinsic masses, and all interactions among the fields and particles of the theory. The Langrangian contains two kinds of physical constants that are basic parameters of the theory: the masses of those fields that represent massive particles, and *coupling constants* that determine the strengths of interactions.

A theory wherein the Higgs mechanism is operative has a scalar field variable $\Phi(\mathbf{r}, t)$ (a *Higgs field*). This Φ appears in the Lagrangian in places where a factor of mass *would* appear for certain particles, if those particles *had* intrinsic mass. The potential energy of the Higgs field is then defined in such a way that its state of minimum energy has a nonzero, constant value Φ_0. This state of lowest energy corresponds to the *vacuum*. Particles with no *intrinsic* mass thereby acquire *effective* masses proportional to Φ_0. These effective masses also contain numerical factors that depend on the type of particle, giving rise to the differences in mass that are observed.

The Higgs mechanism seems like a very artificial way to explain so basic a property as mass. When first proposed, it was regarded by many as a mathematical curiosity with no physical significance. The Higgs mechanism plays a central role in the Weinberg–Salam electroweak theory, however, for the weak bosons are assumed to acquire their large effective masses in precisely this way. Furthermore, the Weinberg–Salam theory has been experimentally confirmed in striking detail (including the existence of the neutral weak Z boson, with just the mass and other properties predicted by the theory). Verification of the theory led to Weinberg and Salam sharing the 1979 Nobel Prize in Physics with Sheldon Lee Glashow, who had laid some of the ground-work for the theory.

Only a single Higgs field has been mentioned thus far, but the Higgs mechanism actually requires at least two: one that takes on a nonzero value in the vacuum, plus one for each vector boson that acquires an effective mass through the Higgs mechanism. [There are four real (or two complex) Higgs fields in the Weinberg–Salam theory, since the W^+, W^-, and Z weak bosons all acquire effective masses in this way.] Furthermore, there are many different states that share the same minimum energy, and hence many possible vacuum states for the theory.

As a simple example, consider a theory with two Higgs fields Φ_1 and Φ_2 and a potential energy–density function V given by

$$V(\Phi_1, \Phi_2) = b\left(\Phi_1^2 + \Phi_2^2 - m^2\right)^2 \qquad (85)$$

where b and m are positive constants. (We shall see that m corresponds to the Φ_0 mentioned previously.) The minimum value of V is zero, which occurs when $\Phi_1^2 + \Phi_2^2 = m^2$. This is the equation for a circle of radius m, in a plane where the two axes measure Φ_1 and Φ_2, respectively. Every point on this circle corresponds to zero energy, and could serve as the vacuum.

A trigonometric identity states that $\cos^2\psi + \sin^2\psi \equiv 1$, for arbitrary angle ψ. The circle of zero energy in the Φ_1, Φ_2, plane is therefore described by

$$\begin{aligned} \Phi_1 &= m\cos\psi \\ \Phi_2 &= m\sin\psi \end{aligned} \tag{86}$$

where ψ ranges from 0 to 360°. Any *particular* angle ψ_0 corresponds to a particular choice of vacuum state. It is useful to consider a pair of Higgs fields Φ_1' and Φ_2' defined by

$$\begin{aligned} \Phi_1' &\equiv \Phi_1\sin\psi_0 - \Phi_2\cos\psi_0 \\ \Phi_2' &\equiv \Phi_1\cos\psi_0 + \Phi_2\sin\psi_0 \end{aligned} \tag{87}$$

Comparing Eq. (86) for $\psi = \psi_0$ with Eq. (87), we see that $\Phi_1' = 0$ and $\Phi_2' = m$. Hence one may say that "the vacuum" is characterized by $\Phi_1' = 0$ and $\Phi_2' = m$, so long as one bears in mind that some *particular* vacuum corresponding to a *particular* value of ψ has been selected by nature, from a continuous set of possible vacuums. Some intrinsically massless particles acquire *effective* masses from the fact that $\Phi_2' = m$ under ordinary conditions, as discussed previously.

Theories containing Higgs fields are always constructed so that the Higgs fields transform among one another under the action of a symmetry group, i.e., a group of transformations that leaves the Langrangian unchanged. [For the simple example just described, with V given by Eq. (85), the symmetry group is that of rotations in a plane with perpendicular axes corresponding to Φ_1 and Φ_2: the group is called $U(1)$.]

Higgs fields are always accompanied in a theory by (fundamentally) massless vector bosons that transform among one another under the same symmetry group. The process whereby nature chooses some particular vacuum (corresponding in our simple example to a particular value for ψ) is essentially random, and is called *spontaneous symmetry breaking*. Once the choice has been made, the original symmetry among all the Higgs fields is broken by the contingent fact that the actual vacuum realized in nature has a nonzero value for some particular Higgs field but not for the others. At the same time, the symmetry among the massless vector bosons is broken by the generation of effective mass for one (or more) of them, which in turn shortens the range of the force resulting from exchange of one (or more) of the vector bosons.

Theories involving the Higgs mechanism are said to possess a *hidden symmetry*. Such theories contain a symmetry that is not apparent in laboratories, where nature has *already* selected some *particular* vacuum, thereby concealing the existence of other possible vacuums. Furthermore, every possible vacuum endows a subset of intrinsically massless particles with effective masses, thereby obscuring their fundamental similarity to other particles that remain massless. Exchanges of the effectively massive particles result in forces of limited range ($\sim\hbar/mc$), whereas any particles that remain massless generate long-range forces (electromagnetism being the only one, except for gravity). Extraordinary imagination was required to realize that electromagnetism and the weak interaction might be different aspects of a single underlying theory. Actual construction of the resulting *electroweak theory* may reasonably be regarded as the work of geniuses.

It seems possible, even likely, that in other regions of space–time far removed from our own, the choice of vacuum has been made differently. In particular, there is no reason for the selection to have been made in the same way in regions of space–time that were outside each others' horizons at the time when selection was made. The fundamental laws of physics are believed to be the same everywhere, but the spontaneous breaking of symmetry by random selections of the vacuum state should have occurred differently in different "domains" of the universe.

We have remarked that the state of minimum energy has a nonzero value for one (or more) of the Higgs fields. If we adopt the convention that this state (the vacuum) has zero energy, it follows that any region of space where all Higgs fields vanish must have a mass-energy density $\rho_H > 0$.

Our example with the Higgs potential described by Eq. (85) would have

$$\rho_H = bm^4/c^2 \tag{88}$$

in any region where $\Phi_1 = \Phi_2 = 0$. Work would be required to expand any such region, for doing so would increase the net energy. Such a region is therefore characterized by negative pressure $P_{\rm H}$ that, by the work-energy theorem, is related to $\rho_{\rm H}$ by $P_{\rm H} = -c^2\rho_{\rm H}$ [as may be seen from Eq. (25) for constant ρ and P].

Of all the fields known or contemplated by particle theorists, only Higgs fields have positive energy when the fields themselves vanish. This feature of Higgs fields runs strongly against intuition, but the empirical successes of the Weinberg-Salam theory provide strong evidence for Higgs fields. If all Higgs fields were zero throughout the cosmos, the resulting ρ_H and negative P_H would perfectly mimic a positive cosmological constant, causing an exponential rate of growth for the universe. The original theory

of cosmic inflation is based on the premise that all Higgs fields were zero at a very early time. The rate of growth would have been enormous, increasing the size of the universe by many powers of ten during the first second of cosmic history. This extraordinary period of *cosmic inflation* ended when the Higgs fields assumed their present values, corresponding to our vacuum.

The Higgs fields relevant to inflation arise in *grand unified theories* (GUTs) of elementary particles and their interactions. The possibility of inflation was first discovered while analyzing a GUT, and many of the subsequent models for inflation are based on GUTs. Inflation also occurs in *supersymmetric* and *superstring* theories of elementary particles, where Higgs fields have analogs called *inflaton fields* (inflation-causing fields). The Higgs fields of GUTs have most of the essential features, however, and we shall discuss them first.

C. Grand Unified Theories (GUTs)

As evidence was mounting for the Weinberg–Salam theory in the early 1970s, a consensus was also emerging that the strong force between quarks is a result of the exchange of eight colored gluons, as described by a theory called *quantum chromodynamics*. Like the photon and the weak bosons, gluons are vector bosons, which makes feasible a unification of all three forces. [Gravitons are not vector bosons (they have a spin of twice $\hbar$), so a unification including gravity would require a different unifying principle.]

A striking feature of the standard model is that the strengths of all three forces are predicted to grow closer as the energy increases, becoming *equal* at a collision energy of $\sim 2 \times 10^{16}$ GeV (somewhat higher than reported earlier, because of recent data). This merging of all three strengths at a single energy suggested very strongly that the forces are unified at higher energies, by some symmetry that is spontaneously broken at $\sim 2 \times 10^{16}$ GeV. A symmetry of this kind would also explain why electrons and protons have precisely equal electrical charges (in magnitude). Experiments have established that any difference is less than one part in 10^{21}, a fact that has no explanation unless electrons are related by some symmetry to the quarks of which protons are made.

Guided by these considerations and by insights gained from the earlier electroweak unification, Howard M. Georgi and Sheldon Glashow proposed the first GUT of all three forces in 1974 [a group called SU(5) was assumed to describe the underlying symmetry]. The Higgs mechanism was adduced to generate effective masses for all particles except photons in the theory (gravitons were not included). In the symmetric phase (all Higgs fields zero), electrons, neutrinos, and quarks would be indistinguishable. The theory also predicted an (effectively) supermassive X boson, whose existence was required to achieve the unification.

Following the lead of Georgi and Glashow, numerous other investigators have proposed competing GUTs, differing in the underlying symmetry group and/or other details. There are now many such theories that reproduce the successful predictions of electroweak theory and quantum chromodynamics at the energies currently accessible in laboratories, while making different predictions for very high energies where we have little or no data of any kind. Thus we do not know which (or indeed whether any) of these GUTs is correct. They typically share three features of significance for cosmology, however, which we shall enumerate and regard as general features of GUTs.

Grand unified theories involve the Higgs mechanism (with at least 24 Higgs fields), and contain numerous free parameters. The energy where symmetry is broken may be set at the desired value ($\sim 2 \times 10^{16}$ GeV) by assigning appropriate values to parameters of the theory. This energy determines the gross features of the Higgs potential, and also the approximate value of ρ_H (the mass-energy density when all Higgs fields are zero). The precise value of ρ_H is somewhat different for different GUTs, and also depends on some parameters whose values can only be estimated. Numerical predictions for ρ_H and the resulting inflation therefore contain substantial uncertainties. We shall see, however, that the observable predictions of inflation are scarcely affected by these uncertainties.

The density ρ_H is given roughly (within a few powers of ten) by

$$\rho_H \sim (10^{16}\ \text{GeV})^4/(\hbar^3 c^5) \sim 10^{80}\ \text{g/cm}^3 \qquad (89)$$

This density is truly enormous, $\sim 10^{47}$ times as great as if our entire sun were compressed into a cubic centimeter. In fact ρ_H is comparable to the density that would arise if all the matter in the observable universe were compressed into a volume the size of a hydrogen atom. As indicated by Eq. (89), this enormous value for ρ_H results from the very high energy at which symmetry appears to be broken.

If all Higgs fields were zero throughout the cosmos, then Eqs. (42) and (89) would imply an enormous, effective cosmological constant

$$\Lambda_H \sim 10^{57}\ \text{m}^{-2} \qquad (90)$$

Any universe whose expansion is governed by a positive Λ will grow at an exponential rate. From Eq. (31), we see that the doubling time corresponding to Λ_H is

$$t_d \sim 10^{-37}\ \text{sec} \qquad (91)$$

At this rate of growth, only 10^{-35} sec would be required for 100 doublings of the cosmic scale factor (which governs

distances between comoving points, i.e., points moving in unison with the overall expansion). Note that $2^{100} \sim 10^{30}$. If inflation actually occurred, it is quite plausible that our universe grew by a factor much larger than 10^{30}, during a very brief time.

We have remarked that the value of ρ_H is uncertain by a few powers of ten. If ρ_H were only 10^{-10} times as large as indicated by Eq. (89), then t_d would be 10^5 times larger than given by Eq. (91). Even with $t_d \sim 10^{-32}$ sec, however, 100 doublings would occur in 10^{-30} sec. Whether 10^{-30} sec were required, or only 10^{-35} sec, makes no observable difference.

A second important feature of GUTs concerns *baryon number*. Baryon number is a concept used to distinguish certain types of particles from their antiparticles. Protons and neutrons are assigned a baryon number of $+1$, in contrast with -1 for antiprotons and antineutrons. Grand unified theories do not conserve baryon number, which means that matter and antimatter need not be created (or destroyed) in equal amounts.

Nonconservation of baryon number is a high-energy process in GUTs, because it results from emission or absorption of X, bosons, whose mc^2 energy is $\sim 10^{16}$ GeV. This is far beyond the range of laboratories, so it cannot be studied directly. When the universe was very young, however, it was hot enough for X bosons to have been abundant. A universe created with equal amounts of matter and antimatter could have evolved at a very early time into one dominated by matter, such as our own. Such a process is called *baryogenesis*.

In 1964, Steven Weinberg presciently remarked that there was no apparent reason for baryon number to be exactly conserved. He had recently shown that the QFT of any long-range force is inconsistent, unless the force is generated by a strictly conserved quantity. Electromagnetism is just such a force, which "explains" why electric charge is precisely conserved. Gravitation is another (the only other) such force, which "explains" why energy is conserved (more precisely, why $\partial^\mu T_{\mu\nu} = 0$). There is no long-range force generated by a baryon number, however, so it would be puzzling if it were exactly conserved. Weinberg discussed the possibility of baryogenesis in the early universe in general terms, but was limited by the absence of any detailed theory for such processes. Andrei Sakharov published a crude model for baryogenesis in 1967, but no truly adequate theory was available before GUTs.

In 1978, Motohiko Yoshimura published the first detailed model for baryogenesis based on a GUT, and many others soon followed. It was typically assumed that the universe contained equal amounts of matter and antimatter prior to the spontaneous breaking of symmetry, and that baryogenesis occurred as thermal energies were dropping below the critical value of $\sim 10^{16}$ GeV. Such calculations seem capable of explaining how matter became dominant over antimatter, and also why there are $\sim 10^9$ times as many photons in the cosmic microwave background as there are baryons in the universe.

The apparent success (within substantial uncertainties) of such calculations may be regarded as indirect evidence for GUTs (or theories with similar features, such as supersymmetric and superstring theories). More direct and potentially attainable evidence would consist of observing proton decay, which is predicted by various GUTs to occur with a half-life on the order of 10^{29}–10^{35} yr (depending on the particular GUT and on the values of parameters that can only be estimated). No proton decay has yet been observed with certainty, however, and careful experiments have shown the proton half-life to exceed 10^{32} yr (which rules out some GUTs that have been considered).

A third feature of GUTs results from the fact that there is no single state of lowest energy but rather a *multiplicity* of possible vacuum states. The choice of vacuum was made by nature as the universe cooled below $\sim 10^{29}$ K, when thermal energies dropped below $\sim 10^{16}$ GeV at a cosmic age of $\sim 10^{-39}$ sec. Since different parts of the universe were outside each others' horizons and had not yet had time to interact in any way, the choice of vacuum should have occurred differently in domains that were not yet in causal contact. These mismatches of the vacuum in different domains would correspond to surfacelike defects called *domain walls* and surprisingly, as was shown in 1974 by G. t' Hooft and (independently) by A. M. Polyakov, also to pointlike defects that are magnetic monopoles.

If the universe has evolved in accord with the standard big bang model (i.e., without inflation), then such domain walls and magnetic monopoles should be sufficiently common to be dramatically apparent. None have yet been observed with certainty, however, which may either be regarded as evidence against GUTs or in favor of cosmic inflation: inflation would have spread them so far apart as to render them quite rare today.

D. Phase Transitions

The notion of phase is most familiar in its application to the three states of ordinary matter: solid, liquid, and gas. As heat is added to a solid, the temperature gradually rises to the melting point. Addition of further energy equal to the latent heat of fusion results in a phase transition to the liquid state, still at the melting point. The liquid may then be heated to the boiling point, where addition of the latent heat of vaporization results in a transition to the gaseous

phase. The potential energy resulting from intermolecular forces is a minimum in the solid state, distinctly greater in the liquid state, and greater still for a gas. The latent heats of fusion and vaporization correspond to the changes in potential energy that occur as the substance undergoes transition from one phase to another.

In QFT, the state of minimum energy is the vacuum, with no particles present. If the theory contains Higgs fields, then one of them is nonzero in the vacuum. The vacuum (if perfect) has a temperature of absolute zero, for any greater temperature would entail the presence of a thermal (blackbody) distribution of photons. A temperature of absolute zero is consistent, of course, with the presence of any number of massive particles.

If heat is added to some region of space it will result in thermal motion for any existing particles, and also in a thermal distribution of photons. If the temperature is raised to a sufficiently high level, then particle–antiparticle pairs will be created as a result of thermal collisions: the mc^2 energies of the new particles will be drawn from the kinetic energies of preexisting particles. The Higgs field that was a nonzero constant in the vacuum would retain that value over a wide range of temperatures, however, and this condition may be regarded as defining a phase: it implies that certain vector bosons have effective masses, so that the underlying symmetry of the theory is broken in a particular way.

If heat continues to be added, then the temperature and thermal energy will eventually rise to a point where the Higgs fields no longer remain in (or near) their state of lowest energy: they will undergo random thermal fluctuations about a central value of zero. When this happens, the underlying symmetry of the theory is restored: those vector bosons that previously had effective masses will behave like the massless particles they fundamentally are. This clearly constitutes a different phase of the theory, and we conclude that nature undergoes a phase transition as the temperature is raised above a critical level sufficient to render all Higgs fields zero.

Roughly speaking, this symmetry-restoring phase transition occurs at the temperature where the mass-energy per unit volume of blackbody radiation equals the $\rho_{\rm H}$ corresponding to the vanishing of all Higgs fields; an equipartition of energy effects the phase transition at this temperature when latent heat equal to $\rho_{\rm H}c^2$ is added. For the $\rho_{\rm H}$ given by Eq. (89), the phase transition occurs at a temperature near 10^{29} K. Like all phase transitions, it can proceed in either direction, depending on whether the temperature is rising or falling as it passes through the critical value.

The relation between phase transitions and symmetry breaking may be elucidated by noting that ordinary solids break a symmetry of nature that is restored when the solid melts. The laws of physics have rotational symmetry, as do liquids: there is no preferred direction in the arrangement of molecules in a liquid. When a liquid is frozen, however, the rotational symmetry is broken because solids are crystals. The axes of a crystal may be aligned in any direction; but once a crystal has actually formed, its particular axes single out particular directions in space.

It is also worth noting that rapid freezing of a liquid typically results in a moderately disordered situation where different regions have their crystal axes aligned in different directions: domains are formed wherein the rotational symmetry is broken in different ways. We may similarly expect that domains of broken-symmetry phase were formed as the universe cooled below 10^{29} K, with the symmetry broken in different ways in neighboring domains. In this case the horizon distance would be expected to govern the domain size.

There is one further aspect of phase transitions that is of central importance for cosmic inflation, namely, the phenomenon of *supercooling*. For a familiar example, consider water, which normally freezes at 0°C. If a sample of very pure (distilled) water is cooled in an environment free of vibrations, however, it is possible to cool the sample more than 20°C below the normal freezing point while still in the liquid phase. Liquid water below 0°C is obviously in an unstable state: any vibration is likely to trigger the onset of crystal formation (i.e., freezing), with consequent release of the latent heat of fusion from whatever fraction of the sample has frozen.

When supercooling of water is finally terminated by the formation of ice crystals, *reheating* occurs: release of the crystals' heat of fusion warms the partially solid, partially liquid sample (to 0°C, at which temperature the remaining liquid can be frozen by removal of its heat of fusion). Such reheating of a sample by released heat of fusion is to be expected whenever supercooling is ended by a phase transition (though not all systems will have their temperature rise as high as the normal freezing point).

Some degree of supercooling should be possible in any system that undergoes phase transitions, including a GUT with Higgs fields. A conjecture that supercooling (and reheating) occurred with respect to a GUT phase transition in the early universe plays a crucial role in theories of cosmic inflation, which we are now prepared to describe in detail.

VIII. SCENARIOS FOR COSMIC INFLATION

A. The Original Model

We now assume that some GUTs with the features described earlier (in Section VII.C) provide a correct description of particle physics, and consider the potential

consequences for the evolution of the cosmos. According to the standard big bang model, the temperature exceeded 10^{29} K for times earlier than 10^{-39} sec. Hence the operative GUT would initially have been in its symmetric phase wherein all Higgs fields were undergoing thermal fluctuations about a common value of zero.

As the temperature dropped below the critical value of $\sim 10^{29}$ K, there are two distinct possibilities (with gradations between). One possibility is that the GUT phase transition occurred promptly, in which case the expansion, cooling, and decrease in mass-energy density of the universe would have progressed in virtually the same way as in the standard model. The GUT symmetry would have been broken in different ways in domains outside each others' horizons, and one can estimate how many domain walls and magnetic monopoles would have resulted. A standard solution to the field equations (Section VI.A) tells one how much the universe has expanded since then, so it is a straightforward matter to estimate their present abundances.

No domain walls have yet been observed, nor magnetic monopoles with any confidence (there have been a few candidates, but too few for other explanations to be ruled out). The abundances predicted by the preceding line of reasoning vastly exceed the levels consistent with observations. For example, work by John Preskill and others indicated in 1979 that monopoles should be $\sim$1% as common as protons. Their magnetic properties would make them easy to detect, but no search for them has succeeded. Furthermore, monopoles are predicted to be $\sim 10^{16}$ times as massive as protons. If they were $\sim$1% as common as protons, their average density in the universe would be $\sim 10^{12}$ times the critical density. In reality, the observed density of monopoles is vastly less than predicted by GUTs with a prompt phase transition. This was called the "monopole problem."

The monopole problem was discovered by elementary particle theorists soon after the first GUT was introduced, and appeared for several years to constitute evidence against GUTs. In 1980, however, Alan H. Guth proposed a remarkable scenario based on GUTs that held promise for resolving the horizon, flatness, monopole, and smoothness problems, and perhaps others as well. Guth conjectured that as the universe expanded and cooled below the GUT critical temperature, supercooling occurred with respect to the GUT phase transition. All Higgs fields retained values near zero for a period of time after it became thermodynamically favorable for one (or a linear combination) of them to assume a nonzero value in the symmetry-breaking phase.

Supercooling would have commenced when the universe was $\sim 10^{-39}$ sec old. The mass-energy density of all forms of matter and radiation other than Higgs fields would have continued to decrease as the universe expanded, while that of the Higgs fields retained the constant value ρ_{H} corresponding to zero values for all Higgs fields. Within $\sim 10^{-38}$ sec, ρ_{H} and the corresponding negative pressure $P_{\mathrm{H}} = -\rho_{\mathrm{H}} c^2$ would have dominated the stress-energy tensor, which would then have mimicked a large, positive cosmological constant. The cosmic scale factor would have experienced a period of exponential growth, with a doubling time of $\sim 10^{-37}$ sec. Guth proposed that all this had happened, and called the period of exponential growth the *inflationary era*. (The numbers estimated by Guth were somewhat different in 1980. Current values are used here, with no change in the observable results.)

The Higgs fields were in a metastable or unstable state during the inflationary era, so the era was destined to end. Knowledge of its duration is obviously important: How many doublings of the scale factor occurred during this anomalous period of cosmic growth? The answer to this question depends on the precise form of the Higgs potential, which varies from one GUT to another and unfortunately depends, in any GUT, on the values of several parameters that can only be estimated.

Guth originally assumed that the Higgs potential had the form suggested by Fig. 4a. For such a potential, a zero Higgs field corresponds to a local minimum of the energy density, and the corresponding state is called a *false vacuum*. (The *true vacuum* is of course the state wherein the Higgs potential is an absolute minimum; that is the broken-symmetry phase where one of the Higgs fields has a nonzero constant value.)

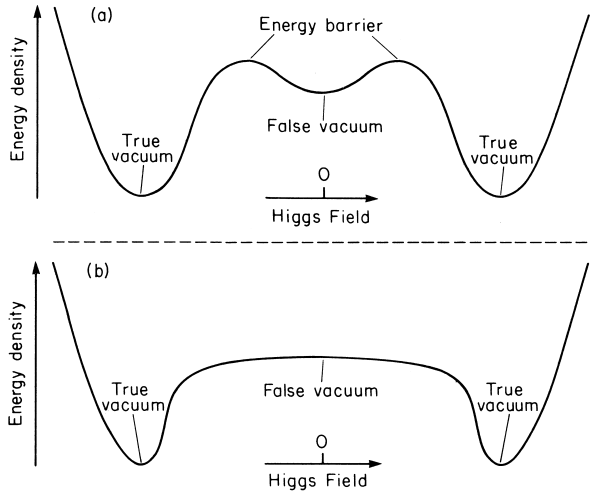


FIGURE 4 Higgs potential energy density for (a) original inflationary model, and (b) new inflationary model. For a theory with two Higgs fields, imagine each curve to be rotated about its vertical axis of symmetry, forming a two-dimensional surface. A true vacuum would then correspond to any point on a circle in the horizontal plane.

In the classical (i.e., nonquantum) form of the theory, a false vacuum would persist forever. The physics is analogous to that of a ball resting in the crater of a volcano. It would be energetically favorable for the ball to be at a lower elevation, hence outside the volcano at its base. In classical physics there is no way for the ball to rise spontaneously over the rim and reach the base, so the ball would remain in the crater for all time.

In quantum physics, the story changes: a ball in the crater is represented by a wave packet, which inevitably develops a small extension or "tail" outside the crater. This feature of quantum mechanics implies a finite probability per unit time that the ball will "tunnel through" the potential energy barrier posed by the crater's rim and materialize outside at a lower elevation, with kinetic energy equal to the decrease in potential energy. Such quantum-mechanical tunneling through energy barriers explains the decay of certain species of heavy nuclei via the spontaneous emission of alpha particles, and is a well-established phenomenon.

If there exists a region of lower potential energy, for either alpha particles or the value of a Higgs field, then "decay" of the initial state via tunneling to that region is bound to occur sooner or later. The timing of individual decays is governed by rules of probability, in accordance with the statistical nature of quantum-mechanical predictions.

The decay of the false vacuum was first studied and described by Sidney R. Coleman. Initially all Higgs fields are zero, but one or another of them tunnels through the energy barrier in a small region of space and acquires the nonzero value corresponding to the true vacuum, thereby creating a "bubble" of true vacuum in the broken symmetry phase.

Once a bubble of true vacuum has formed, it grows at a speed that rapidly approaches the speed of light, converting the surrounding region of false vacuum into true vacuum as the bubble expands. As the phase transition proceeds, energy conservation implies that the mass-energy density $\rho_{\rm H}$ of the preexisting false vacuum is converted into an equal density of other forms of mass-energy, which Guth hoped would become hot matter and radiation that evolved into our present universe.

A low rate of bubble formation corresponds to a high probability that any particular region of false vacuum will undergo many doublings in size before decay of the false vacuum ends inflation for that region. Guth noted that if the rate of bubble formation were low enough for inflation to have occurred by a factor of $\sim 10^{30}$ or more, then the observable portion would have emerged from small enough a region for all parts to have been in causal contact before inflation. Such causal contact would have resulted in a smoothing of any earlier irregularities via processes leading toward thermal equilibrium, thereby explaining (one hopes) the homogeneity and isotropy now observed: the horizon problem would be solved (and perhaps the smoothness problem as well).

It was furthermore noted by Guth that comparable growth of the cosmic scale factor could solve the flatness problem. Recall the definition of Ω in Section VI.B: $\Omega \equiv \rho/\rho_c$, where ρ and ρ_c denote the actual and critical densities of mass-energy, respectively. If the scale factor had grown by a factor of 10^{30} or more during the period of inflation, any difference between Ω and unity *prior* to inflation would have been reduced by a factor of 10^{60} or more by its *end*. [Equation (48) holds at all times]. R. Dicke and P. J. E. Peebles have shown that Ω could not have differed from unity by more than one part in 10^{15} when the universe was one second old, if the present density of matter lies between 20 and 200% of the critical density (as it does, with $\rho_M \approx 0.3\rho_c$). Perhaps of even greater interest, they showed that if Ω had differed from unity by more that one part in 10^{14} at one second, then *stars would never have formed*. (For $\Omega > 1$, the universe would have collapsed too soon. For $\Omega < 1$, matter would have been diluted too soon by the expansion.) This need for extraordinary "fine-tuning" at early times is the "flatness problem." Inflation by a factor $\geq 10^{30}$ could explain why Ω was so near unity at 1 sec.

If such inflation has occurred, the cosmic scale factor a_0 would be 10^{30} (or more) times greater than in the standard model. Even if space *is* curved, the curvature only becomes significant over distances comparable to the value of a_0 (as explained in Section III.C). Inflation could render a_0 so much larger than the radius of the observable universe that no curvature would be apparent.

The rate at which bubbles form depends on the particular GUT and, within any GUT, on parameters including some whose values can only be estimated. Guth assumed that inflation *is* the reason why our universe is nearly flat, hence that bubble formation was slow enough for $\sim$100 or more doublings to occur ($2^{100} \sim 10^{30}$). With a doubling time $t_d \sim 10^{-37}$ sec, this would have required only $\sim 10^{-35}$ sec.

It would seem coincidental if 100 doublings had occurred but not, for example, 105 or more, which would have reduced $|\Omega - 1|$ by another factor of $2^{10} \approx 1000$ (or more). Standard inflation predicts that considerably more doublings have occurred than are required to explain the obvious features of our universe (e.g., the presence of stars and us). Standard inflation therefore predicts enough doublings to render the observed universe indistinguishable from flat. Even 10^7 doublings would require only $\sim 10^{-30}$ sec, and this would be inflation by a factor of $\sim 10^{3,000,000}$.

Guth sought to understand whether a region now the size of the observable universe was more likely to have evolved from few or many bubbles, and he encountered a major shortcoming of his model. The role played by bubbles in the phase transition led to a bizarre distribution of mass-energy, quite unlike what astronomers observe. The problem may be described as follows.

We have remarked that the ρ_{H} of the false vacuum is converted into an equal amount of mass-energy in other forms as the phase transition proceeds. It can be shown, however, that initially the mass-energy is concentrated in the expanding walls of the bubbles of true vacuum, in the form of moving "kinks" in the values of the Higgs fields. Conversion of this energy into a homogeneous gas of particles and radiation would require a large number of collisions between bubbles of comparable size. Bubbles are forming at a constant rate, however, and grow at nearly the speed of light. If enough time has passed for there to be many bubbles, there will inevitably be a wide range in the ages and therefore sizes of the bubbles: collisions among them will convert only a small fraction of the energy in their walls into particles and radiation. In fact it can be shown that the bubbles would form finite clusters dominated by a single largest bubble, whose energy would remain concentrated in its walls.

One could explain the homogeneity of the observable universe if it were contained within a single large bubble. There remains, however, the problem that the mass-energy would be concentrated in the bubble's walls: the interior of a single large bubble would be too empty (and too cold at early times) to resemble our universe. In the best spirit of science, Guth acknowledged these problems in his initial publication, and closed it with an invitation to readers to find a plausible modification of his inflationary model, one that preserved its strengths while overcoming its initial failure to produce a homogeneous, hot gas of particles and radiation that might have evolved into our present universe.

We note in passing that Andre D. Linde, working at the Lebedev Physical Institute in Moscow, had discovered the principal ingredients of inflation prior to Guth. Furthermore, Linde and Gennady Chibisov realized in the late 1970s that supercooling might occur at the critical temperature, resulting in exponential growth. To their misfortune, however, they were not aware of the horizon and flatness problems of standard cosmology. When they realized that bubble collisions in their model would lead to gross inhomogeneities in the universe, they saw no reason to publicize their work.

Alexei Starobinsky, of the Landau Institute for Theoretical Physics in Moscow, proposed a form of inflation in a talk at Cambridge University in 1979. His focus was on avoidance of an initial singularity at time zero, however, and no mention was made of the horizon, flatness, or monopole problems. Although well received, his talk did not generate enough excitement for word of it to reach the United States. By the time it was published in 1980, Guth had already given many lectures emphasizing how inflation might solve problems of standard cosmology. It was this emphasis that made the potential significance of Starobinsky's and Linde's work apparent to a wide audience.

B. The New Inflationary Model

A major improvement over Guth's original model was developed in 1981 by A. D. Linde and, independently, by Andreas Albrecht with Paul J. Steinhardt. Success hinges on obtaining a "graceful exit" from the inflationary era, i.e., an end to inflation that preserves homogeneity and isotropy and furthermore results in the efficient conversion of ρ_{H} into the mass-energy of hot particles and radiation.

The new inflationary model achieves a graceful exit by postulating a modified form for the potential energy function of Higgs fields: the form depicted in Fig. 4b, which was first studied by Sidney Coleman in collaboration with Erick J. Weinberg. A Coleman–Weinberg potential has no energy barrier separating the false vacuum from true vacuum; instead of resembling the crater of a volcano, the false vacuum corresponds to the center of a rather flat plateau. (Though not a local minimum of energy, the plateau's center is nevertheless called a false vacuum for historical reasons.)

Supercooling of a Coleman–Weinberg false vacuum results in the formation of contiguous, causually connected domains throughout each of which the Higgs fields gradually evolve in a simultaneous, uniform way toward a single phase of true vacuum. The initial sizes of these domains are comparable to the horizon distance at the onset of supercooling, $\sim 10^{-28}$ cm. The mass-energy density of the false vacuum was that of null Higgs fields, $\rho_H \sim 10^{80}$ g/cm^3. Despite this enormous density, the initial mass of the domain in which our entire universe lies was $\sim 10^{-4}$ g.

Within each domain the evolution of a Higgs field away from its initial value of zero is similar to the horizontal motion of a ball that, given a slight nudge, rolls down the initially flat but gradually increasing slope of such a plateau. The equations describing evolution of the Higgs fields have a term corresponding to a frictional drag force for a rolling ball; in addition to the initial flatness of the plateau, this "drag term" serves to retard the evolution of a Higgs field away from zero toward its true-vacuum, broken-symmetry value. The result is called a "slow-rollover" transition to the true vacuum.

As always, there are different combinations of values of Higgs fields that have zero energy and qualify as true vacua; different domains undergo transitions to different true vacua, corresponding to the possibility of a ball rolling off a plateau in any one of many different directions.

As a Higgs field (or linear combination thereof) "rolls along" the top of the potential energy plateau, the mass-energy density remains almost constant despite any change in the volume of space. The corresponding pressure is large and negative, so inflation occurs. The doubling time for the cosmic scale factor is again expected to be $\sim 10^{-37}$ sec. The period of time required for a Higgs field to reach the edge of the plateau (thereby terminating inflation) can only be estimated, but again inflation by a factor of 10^{30}, or by very much more, is quite plausible. As in the "old inflationary model" (Guth's original scenario), new "standard" inflation assumed that the actual inflation greatly exceeded the amount required to resolve the horizon and flatness problems.

When a Higgs field "rolls off the edge" of the energy plateau, it enters a brief period of rapid oscillations about its true-vacuum value. In accordance with the field-particle duality of QFT, these rapid oscillations of the field correspond to a dense distribution of Higgs particles. The Higgs particles are intrinsically unstable, and their decay results in a rapid conversion of their mass-energy into a wide spectrum of less massive particles, antiparticles, and radiation. The energy released in this process results in a high temperature that is less than the GUT critical temperature of $\sim 10^{29}$ K only by a modest factor ranging between 2 and 10, depending on details of the model. This is called the "reheating" of the universe.

The reheating temperature is high enough for GUT baryon-number–violating processes to be common, which leads to matter becoming dominant over antimatter. Reheating occurs when the universe is very young, probably less than 10^{-30} seconds old, depending on precisely how long the inflationary era lasts. From this stage onward, the observable portion evolves in the same way as in the standard big bang model. (A slow-rollover transition is slow relative to the rate of cooling at that early time, so that supercooling occurs, and slow relative to the doubling time of $\sim 10^{-37}$ sec; but it is extremely rapid by any macroscopic standard.)

The building blocks of the new inflationary model have now been displayed. As with the old inflationary model, there is no need to assume that any portion of the universe was initially homogeneous and isotropic. It is sufficient to assume that at least one small region was initially expanding, with a temperature above the GUT critical temperature. The following chain of events then unfolded.

As long as the temperature exceeded the critical temperature of $\sim 10^{29}$ K, all Higgs fields would have undergone thermal fluctuations about a common value of zero, and the full symmetry of the GUT would have been manifest. The assumed expansion caused cooling, however, and the temperature fell below 10^{29} K when the universe was $\sim 10^{-39}$ sec old. Supercooling began, and domains formed with diameters comparable to the horizon distance at that time, $\sim 10^{-28}$ cm, with initial masses of $\sim 10^{-4}$ g.

As supercooling proceeded, the continuing expansion caused dilution and further cooling of any preexisting particles and radiation, and the stress-energy tensor became dominated at a time of roughly 10^{-37} sec by the virtually constant contribution of Higgs fields in the false vacuum state. With the dominance of false-vacuum energy and negative pressure came inflation: the expansion entered a period of exponential growth, with a doubling time of $\sim 10^{-37}$ sec. The onset of inflation accelerated the rate of cooling; the temperature quickly dropped to $\sim 10^{22}$ K (the *Hawking temperature*), where it remained throughout the latter part of the inflationary era because of quantum effects that arise in the context of GTR.

For the sake of definiteness, assume that the domain in which we reside inflated by a factor of 10^{50}, which would have occurred in $\sim 2 \times 10^{-35}$ sec. With an initial diameter of $\sim 10^{-28}$ cm and mass of $\sim 10^{-4}$ g, such a domain would have spanned $\sim 10^{22}$ cm when inflation ended (as in Fig. 2), with a total mass of $\sim 10^{146}$ g. *All but one part in* $\sim 10^{150}$ *of this final mass was produced during inflation*, *as the mass-energy of the false vacuum grew in proportion to the exponentially increasing volume*. (Guth has called this "the ultimate free lunch.") The radius of our presently observable universe would have been ~60 cm when inflation ended (based on the methods of Section VI.A, assuming that $t_{EQ} \geq t_{LS}$).

Inflation ended with a rapid and chaotic conversion of ρ_H into a hot ($T \geq 10^{26}$ K) gas of particles, antiparticles, and radiation. Note that if baryon number were strictly conserved, inflation would be followed for all time by equal amounts of matter and antimatter (which would have annihilated each other almost completely, with conversion of their mass-energy into radiation). The observable universe, however, is strongly dominated by matter. Hence inflation requires a GUT (or a theory sharing key features with GUTs) not only to explain an accelerating expansion, but also to explain how the symmetry between matter and antimatter was broken: through baryon-number–violating processes at the high temperature following reheating.

C. New, Improved Inflation

In the standard big bang model, thermal fluctuations in density at the Planck time would gradually have been amplified, by the concentrating effect of gravity, into present clumps of matter much larger than anything we see. This

is the *smoothness problem*, first noted by P. J. E. Peebles in 1968. When inflation was proposed in 1980, Guth and others hoped that it would solve this problem. The end of inflation creates the initial conditions for the ensuing big bang, so the density perturbations at the end of inflation were the basic issue. During the Nuffield Workshop on the Very Early Universe in the summer of 1982, however, it became clear that the simplest GUT, the SU(5) model of Georgi and Glashow, was unacceptable: it gave rise to values for $\delta\rho/\rho$ that were too large by a factor of $\sim 10^6$.

As other GUTs were analyzed with similar results, a consensus gradually emerged that *no* symmetry-breaking Higgs field in a GUT could give rise to acceptable inflation. The size of $\delta\rho/\rho$ at the end of inflation depends critically on details of the Higgs potential. The potential of a GUT Higgs field, like that in Figure 4b, would have to be wider and flatter by factors $\sim 10^{13}$ in order for the resulting $\delta\rho/\rho$ to have given rise to the stars and galaxies that we see.

Another problem requiring a much wider and flatter potential became apparent. As the early universe was cooling toward the GUT critical temperature, all Higgs fields would indeed have undergone thermal fluctuations about *average* values of zero, because of cancellations between positive and negative values. For Higgs potentials that meet the requirements of GUTs, however, the average *magnitudes* would have been comparable to those in a *true* vacuum. The creation of a false vacuum by cooling to the critical temperature would have been extremely unlikely. "Unlikely" is not "impossible," but plausibility of the theory would be undermined by the need for a special initial condition.

In contrast, if the potential were 10^{13} times wider and flatter, cooling below the critical temperature should have resulted in a false vacuum over *some* tiny region of space, which might then have inflated into a universe like that in which we find ourselves. We are therefore led to consider theories of this kind.

The term *inflaton* (*in*-fluh-*tonn*) refers to any hypothetical field that could cause inflation. We are interested, of course, in inflatons that might have produced *our* universe. A scalar field (having no spatial direction) ϕ seems best suited (perhaps two of them). The underlying particle theory must convert false-vacuum energy into a hot soup of appropriate particles at the end of inflation, in a graceful exit producing density perturbations of the required kind. Finally, the particle theory must contain a mechanism for baryogenesis, making matter dominant over antimatter at an early time.

The potential $V(\phi)$ could have an extremely broad, almost flat region surrounding $\phi = 0$, with $V(0) > 0$. An attractive alternative was proposed by Andrei Linde, however, in 1983. The inflaton potential of Linde's *chaotic inflation* has the shape of a very wide, upright bowl, with the true vacuum at its center ($\phi = 0$). Space-time could have been in a highly disordered, chaotic state at the Planck time. The theory only requires that a single, infinitesimal region had a value for ϕ far enough up the wall to have produced the desired inflation while "rolling" slowly toward the center. Like chaotic inflation, most inflaton models assume that inflation began very near the Planck time, in part to explain why the initial microcosm did not collapse before inflation began.

A broad range of particle theories can be constructed with suitable inflaton fields, including supersymmetric theories, supergravity theories, and superstring theory. Further discussion would take us beyond the range of this article. Improved models employing a wide range of inflatons share all the successes of the "new inflationary model" of 1982, however, and those predicting a flat universe will be included when we speak of "standard inflation."

D. Successes of Inflation

Our universe lies inside a domain with a uniform vacuum state, corresponding to a particular way in which any (now) hidden symmetry of the underlying particle theory was broken. Standard inflation predicts that this domain inflated by a factor much greater than 10^{30}, in which case our universe (almost certainly) lies *deeply* inside. Neither domain walls nor magnetic monopoles would then be found within the observable universe (except for a few that might have resulted from extreme thermal fluctuations after inflation ended). Standard inflation therefore solves the monopole problem. The horizon problem is solved as well, as illustrated in Fig. 2.

The flatness problem has an interesting history. When inflation was proposed in 1980, the matter yet found by astronomers corresponded to a density of matter ρ_M that was $\sim$10% of the critical density ρ_c. Astronomers had by no means completed their search for matter, however. When inflation was proposed, it seemed possible that future discoveries would bring ρ_M up to ρ_c, as required by standard inflation's prediction of flatness. With the passage of time, the presence of additional (dark) matter was indicated by observations, but there has never been evidence that ρ_M is even half of ρ_c. The currently observed value is $\rho_M = (0.31 \pm 0.06)\rho_c$, and it seems virtually certain that $\rho_M \leq \rho_c/2$ (Section V).

By the early 1990s, the apparent discrepancy between ρ_M and ρ_c led to a serious consideration of "open" inflationary models, wherein $\rho_M < \rho_c$ and space has negative curvature. Such models are ingenious and technically viable, but require special assumptions that seem rather artificial.

As evidence mounted that $\rho_M < \rho_c$, however, standard inflation acquired a bizarre feature of its own. Its

prediction of flatness implied the existence of *positive vacuum energy*, in order to bring the total density up to ρ_c. This would be equivalent to a positive cosmological constant (Section IV.F and G), a concept that Einstein abandoned long ago and few had been tempted to resuscitate. Such vacuum energy would have caused an *acceleration* of the cosmic expansion, beginning when the universe was roughly half its present age (Section V.F).

In 1998 two teams of astronomers, one led by Saul Perlmutter and the other by Brian P. Schmidt, reported an astonishing discovery. They had observed supernovae of type Ia (SNe Ia) over a wide range of distances out to ~12 billion light-years (redshift parameter $z \approx 1$, see Section VI.A). Their data strongly indicated that the cosmic expansion is, indeed, accelerating. This conclusion has been confirmed by subsequent studies of SNe Ia. Furthermore, the *amount* of vacuum energy required to explain the acceleration appears to be just the amount required (±10%) to make the universe flat: $\rho_V \approx 0.7\rho_c$, hence $(\rho_M + \rho_V) \approx \rho_c$ (details in Section V.E).

The vacuum energy implied by supernovae data has been strikingly confirmed by studies of the CMB, which has been traveling toward us since the time of last scattering: $t_{LS} \approx 300{,}000$ yr. The CMB is extremely uniform, but contains small regions with deviations from the average temperature on the order of $\delta T/T \sim 10^{-5}$. These provide a snapshot of temperature perturbations at t_{LS}, which were strongly correlated with density perturbations.

Prior to t_{LS}, baryonic matter was strongly ionized. The free electrons and nuclei were frequently scattered by abundant photons, which pressurized the medium. Density waves (often called *acoustic waves*) will be excited by perturbations in any pressurized medium, and the speed of such waves is determined (in a known way) by the medium's properties. At any given time before t_{LS}, there was a maximum distance that such waves could yet have traveled, called the *acoustic horizon distance*.

As the acoustic horizon expanded with the passage of time, it encompassed a growing number of density perturbations, which excited standing waves in the medium. The fundamental mode acquired the greatest amplitude, spanning a distance comparable to the acoustic horizon distance. This is called the *first acoustic peak*. Perturbations in the CMB should contain this peak, with an angular diameter reflecting the acoustic horizon distance at t_{LS}.

The angular diameters of perturbed regions in the CMB depend not only on the linear dimensions of their sources, but also on the curvature (if any) of space. Sources of the CMB are presently a distance $r_p(z_{CMB}) \approx 47 \times 10^9$ ly from us (Section VI.A). If space is curved, angular diameters of perturbations in the CMB would be noticeably affected unless the cosmic scale factor a_0 satisfied (roughly) $a_0 \geq r_p(z_{CMB})$ (Section III.B). From Eqs. (48) and (69), we see that this would require $|\Omega_0 - 1| \leq 0.11$.

BOOMERanG (Balloon Observations of Millimetric Extragalactic Radiation and Geomagnetics) results published in 2000 are portrayed in Fig. 5. The index ℓ refers to a multipole expansion (spherical harmonics, averaged over m for each ℓ). If space were flat, the angular diameter of the first acoustic peak would be $\approx 0.9^\circ$, corresponding to $\ell_{peak} \approx 180^\circ/0.9^\circ \approx 200$.

With $\rho_M \approx 0.3\rho_c$ but no vacuum energy, space would be negatively curved, giving rise to $\ell_{peak} \approx 500$. This is clearly ruled out by the data. For Ω_0 near unity, the prediction can be expressed as $\ell_{peak} \approx 200/\sqrt{\Omega_0}$, and the data in Fig. 5 indicate $\ell_{peak} = (197 \pm 6)$. At the 95% confidence level, the authors concluded that $0.88 \leq \Omega_0 \leq 1.12$. The best-fitting Friedmann solution for spacetime (Section IV.A) has $\rho_M = 0.31\rho_c$ and $\rho_V = 0.75$, reproducing the SNe Ia result for vacuum energy within the margin of error. A more resounding success for standard inflation is not easily imagined.

A sweeping, semiquantitative prediction about density perturbations is made by inflation. In the absence of quantum fluctuations, inflation would rapidly drive the false vacuum to an extremely uniform state. Quantum fluctuations are ordinarily microscopic in size and very short-lived, but rapid inflation would negate both of these usual properties.

Inflation would multiply the diameters of quantum fluctuations, and furthermore, do this so rapidly that opposite sides would lose causal contact with each other before the

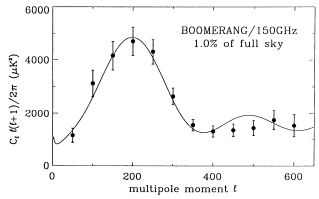


FIGURE 5 Evidence for flatness and vacuum energy (cosmological constant). Angular distribution of $(\delta T/T)^2$ of the cosmic microwave background, measured by instruments in balloons launched from McMurdo Station, Antarctica, by the BOOMERanG Project. Curve indicates best-fitting Friedmann solution for spacetime, with cosmological parameters $\Omega_M = 0.31$, $\Omega_V = 0.75$, and Hubble parameter $h_0 = 0.70$. At the 95% confidence level, the study concluded that $0.88 \leq \Omega_0 \leq 1.12$, where $\Omega_0 \equiv \Omega_M + \Omega_V = 1$ for flat space. The positive Ω_V means positive vacuum energy, corresponding to a cosmological constant that would accelerate the cosmic expansion. The index ℓ refers to a multipole expansion in terms of spherical harmonics $Y_{\ell m}(\theta, \varphi)$, where m has been summed over for each ℓ. [Adapted from P. de Bernardis, *et al.* Reprinted by permission from *Nature* **404,** 955 (2000) MacMillan Magazines Ltd.]

fluctuation vanished. The resulting density perturbations would be "frozen" into the expanding medium, thereby achieving permanency. By inflation's end, the sizes of perturbations would span an *extremely* broad range. The earliest fluctuations would have been stretched enormously, the latest very little, and all those from intermediate times by intermediate amounts. The magnitudes of the resulting $\delta\rho/\rho$ are very sensitive to details of the inflaton potential, but the mechanism described above results in a distribution of sizes that depends rather little on details of the potential.

The kind of probability distribution described above should be apparent in the CMB, since density perturbations correspond closely to perturbations in the radiation's temperature. The Cosmic Background Explorer satellite, better known as COBE (*co*-bee), was launched in 1989 to study the CMB with unprecedented precision. Its mission included measuring deviations from uniformity in the CMB's temperature. Over two years were required to gather and analyze the data displayed in Fig. 6.

These data were first presented at a cosmology conference in Irvine, CA, in March of 1992. The long awaited results struck many attendees as a historic confirmation of inflation, and none have since had reason to feel otherwise. Many subsequent studies have confirmed and refined the data in Fig. 6, but we present the earliest results because of their historical significance.

The density perturbations predicted by inflation are similar to those required to explain the evolution of structure formation in our universe. Indeed, one may reasonably hope that some *particular* inflaton potential will be wholly successful in describing further details of the CMB soon to be measured, and also in explaining details of the structures we observe.

Inflation provides the only known explanations for several puzzling, quite special features of our universe. Standard inflation has also made two quantitive predictions, both of which have been strikingly confirmed (at least within present uncertainties). Only time can tell, but the prognosis for inflation seems excellent, in some form broadly similar to that described here.

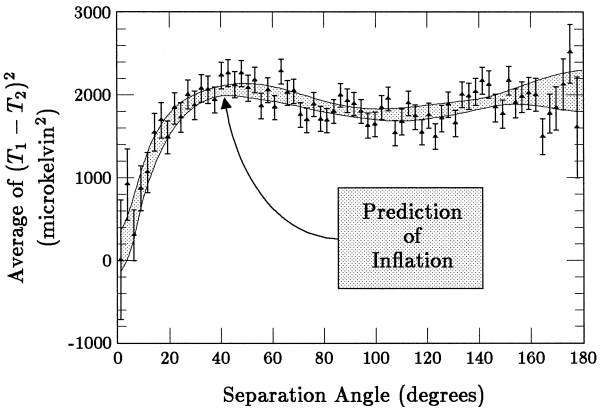


FIGURE 6 COBE nonuniformity data. The data from the COBE satellite gives the temperature T of the background radiation for any direction in the sky. The experimental uncertainties for any one direction are large, but statistically meaningful quantities can be obtained by averaging. The COBE team considered two directions separated by some specific angle, say 15°, and computed the square of the temperature difference, $(T_1 - T_2)^2$, measured in microkelvins (10^{-6} K). By averaging this quantity over all pairs of directions separated by 15°, they obtained a statistically reliable number for that angle. The process was repeated for angles between 0 and 180°. The computed points are shown as small triangles, with the estimated uncertainty shown as a vertical line extending above and below the point. The gray band shows the theoretically predicted range of values corresponding to the scale-invariant spectrum of density perturbations arising from inflation. The effect of the earth's motion through the background radiation has been subtracted from both the data and the prediction, as has a specific angular pattern (called a quadrupole) which is believed to be contaminated by interference from our own galaxy. Since inflation determines the shape of the spectrum but not the magnitude of density perturbations, the magnitude of the predicted gray band was adjusted to fit the data. [From "The Inflationary Universe" by Alan Guth. Copyright © 1997 by Alan Guth. Reprinted by permission of Perseus Books Publishers, a member of Perseus Books, L.L.C.]

E. Eternal Inflation, Endless Creation

Throughout inflation, the inflation field ϕ is on a downhill slope of the potential $V(\phi)$. The false-vacuum density ρ_{FV} is therefore not strictly constant, but gradually decreases as ϕ approaches the end of inflation. Since $t_d \propto 1/\sqrt{\rho_{FV}}$ [Eqs. (31) and (42)], the rate of exponential growth also tends to decrease, all else being equal. In our discussion of density perturbations, however, we noted that quantum fluctuations in ϕ result in perturbations $\delta\rho$ that are frozen into permanency by the rapid expansion. A fascinating question therefore arises.

For simplicity, suppose that ϕ is precisely uniform throughout space at the beginning of inflation. Quantum fluctuations will immediately result in perturbations that become frozen into space. In regions where $\delta\rho > 0$, the expansion rate will be slightly greater than average, and such regions will fill an increasing fraction of space. Within these growing regions, later fluctuations will produce *new* regions where $\delta\rho > 0$, compounding the original deviation from average density and rate of expansion. One expects negative $\delta\rho$ as often as positive, but a string of positives causes that region of space to grow more rapidly, increasing the volume of space where a randomly positive $\delta\rho$ would lengthen the string. This can continue indefinitely.

Of course the *average* ρ is decreasing while the region described above continues rising above the average, so

the outcome is unclear. Will inflation come to an end in such a region, sometime later than average, or will there be regions where ρ_{FV} never shrinks to zero and inflation never ends?

The answer clearly depends on the size, frequency, and magnitude of quantum fluctuations, and also on details of the inflaton potential. Alexander Vilenkin answered this question for new inflationary models (Section VIII.B) in 1973. His extraordinary conclusion was that once "new inflation" has begun, there will *always* be regions where it continues. Andrei Linde established this result for a wide class of "improved" inflationary models, including his own "chaotic inflation," in 1986. This phenomenon is called *eternal inflation*, and appears to be characteristic of virtually all inflationary models that might be relevant to our universe.

Eternal inflation is accompanied by endless formulation of "pockets" of true vacuum. Each is surrounded by eternally inflating false vacuum, containing within it all the other pockets yet created. A detailed analysis reveals that the *number* of pockets of true vacuum *grows at an exponential rate*, with a doubling time comparable to that of the surrounding false vacuum (typically *much* less than 10^{-30} sec).

If standard inflation is correct, *our observable universe lies deep within a single such pocket*. The number of such pockets would double a minimum of 10^{30} times each second. This corresponds to the number of pockets (and universes contained therein) increasing by a *minimum* factor of $(10^{1,000,000,000})^3$, every second, for eternity. The possibility of truth being stranger than fiction has risen to a new level.

Given that standard inflation implies endless creation, human curiosity naturally wonders whether such inflation had any *beginning*. A definitive answer remains elusive. Under certain very technical but plausible assumptions, however, Arvind Borde and Alexander Vilenkin proved in 1994 that inflation cannot extend into the infinite *past*. A beginning is required, which renews an ancient question.

F. Creation *ex nihilo*

We have seen that the observable universe may have emerged from an extremely tiny region that experienced inflation and then populated the resulting cosmos with particles and radiation created from the mass-energy of the false vacuum. An ancient question arises in a new context: How did that tiny region come into being, from which the observable universe emerged? Is it possible to understand the creation of a universe *ex nihilo* (from nothing)?

Scientific speculation about the ultimate origin of the universe appears to have begun in 1973, with a proposal by Edward P. Tryon that the universe arose as a *vacuum fluctuation*, i.e., a spontaneous quantum transition from an empty state to a state containing tangible matter and energy. The discovery of this theoretical possibility shares two features with Guth's discovery of inflation. The first is that neither investigator was considering the question to which an answer was found. Guth was a particle theorist, seeking to determine whether GUTs failed the test of reality by predicting too many monopoles. He had not set out to study cosmology, let alone to revolutionize it.

In Tryon's case, he had never imagined it *possible* to explain how and why the big bang occurred. Universes with no beginning had been quite popular, in large measure because they nullified (by fiat) the issue of creation. When the steady-state model was undermined by discovery of the CMB in 1965, the eternally oscillating model replaced it as the favored candidate, despite the mystery of how a universe could have bounced back from each collapse and furthermore retained its uniformity through infinitely many bounces. The only alternative was a universe of finite age, which seemed incomprehensible from a scientific point of view.

Interests in cosmology, quantum theory, and particle physics arose in Tryon's undergraduate years. Graduate studies commenced in 1962 at the University of California, Berkeley, where his courses included quantum field theory and general relativity. Classical and quantum theories of gravitation were the subjects of his dissertation.

It had long been a mystery why mass plays two quite different roles. It describes resistance to acceleration, or *inertia*, in Newton's laws of motion. In Newton's law of gravitation, $F = GmM/r^2$, mass plays the role of gravitational "charge" (analogous to electric charge in Coulomb's law). This puzzle had led Ernst Mach in the 1880s to conjecture that inertia might somehow be a result of interaction with the rest of the universe (*Mach's Principle*). Tryon found this possibility tantalizing, and took a break from his thesis project one afternoon to see if he could invent a quantitative theory.

Denoting inertial mass by m_I and gravitational mass (or "charge") by m_G, Newton's gravity becomes $F = m_G M_G/r^2$ in suitably chosen units. The two types of mass are presumably related by

$$m_I = f(\vec{r}, t) m_G, \tag{92}$$

where f is a scalar field expressing the inertial effect of all matter in the universe. An obvious constraint on any such theory is the empirical fact that $f \simeq 1/\sqrt{G}$ near earth.

Guided by dimensional analysis and simplicity, Tryon guessed that f might be the magnitude of a relativistic analog to gravitational potential, divided by c^2:

$$f = |V_G|/c^2, \tag{93}$$

where the gravitational potential energy (GPE) of a mass m would be

$$\text{GPE} = -m_G|V_G|, \tag{94}$$

since $V_G < 0$. In the nonrelativistic (Newtonian) limit, one would then have $\nabla^2 f = -4\pi\rho_G/c^2$, where ρ_G denotes the density of gravitational mass or "charge," and ∇^2 is the Laplacian differential operator. The desired analog is

$$f\Box^2 f = \frac{4\pi}{c^2}T_\mu^\mu, \tag{95}$$

where $\Box^2$ is the invariant d'Alembertian of general relativity, $T_\mu^\mu \equiv g^{\mu\nu}T_{\mu\nu}$, and $T_{\mu\nu}$ is the *inertial* stress-energy tensor (so that ρ_G appears in $T_{\mu\nu}/f$).

For a simple test of the theory, Tryon used the $T_{\mu\nu}$ of Eq. (21) and the critical density ρ_c given by Eq. (45), assuming zero pressure. (The actual density was known to be comparable to ρ_c, perhaps even equal, and calculations were simpler in a flat space). In a rough estimate for f, the universe was treated as static. Since no causal influence could have reached us from beyond the horizon, the volume integral over ρ_c was cut off at the horizon distance $d_{SH} = 2c/H_0$ [Eqs. (62) and (64)]. The result was a cut-off version of Newton's gravity:

$$f^2 \approx \left(M_I/c^2\right) \times \langle 1/r\rangle, \tag{96}$$

where M_I denotes the total (inertial) mass within the horizon, and $\langle 1/r\rangle = 3/2d_{SH}$ is the average value of $1/r$.

With M_I given by ρ_c, Eq. (96) yields $f \approx \sqrt{3/G}$, which is the result desired within a factor of $\sqrt{3}$. It would be difficult to exaggerate the adrenaline that was produced by this discovery. Given the crudity of the calculation, it seemed likely that a more careful analysis would yield the *exact* result desired for *some* density near ρ_c, perhaps the density of our own universe.

When Tryon met with his mentor the next day, he was disappointed to learn that this approximate relation between ρ_c and G was already known. In fact Carl Brans and Robert Dicke had published a similar (but far more sophisticated) theory a few years earlier (1961). One difference was that unlike Tryon's crude theory, theirs contained an adjustable parameter ω, and became identical with GTR in the limit as $\omega \to \infty$.

Active interest in the Brans-Dicke theory was soon undermined by observations, all of which were consistent with standard GTR. The Brans-Dicke theory could not be ruled out by such data, since ω could always be chosen large enough to obtain agreement. In the absence of any discrepancy between observations and GTR, however, GTR remained the favored candidate. Tryon followed these developments closely, and gradually abandoned the Machian interpretation of the relation ρ_c between and G. He retained a deep conviction, however, that the nearly critical density of our universe was no coincidence. He believed it to be a cryptic manifestation of some truth of fundamental importance, and he resolved to revisit the issue from time to time, hoping eventually to decode the message.

Tryon's research centered on the strong interactions for several years, but he occasionally spent a few hours wrestling with the significance of a density near ρ_c. One afternoon in 1973, an epiphany occurred. To his utter surprise and astonishment, his "mind's eye" saw a brilliant flash of light, and he realized instantly that he was witnessing the birth of the universe. In that same instant, he realized how and why it happened: the universe is a vacuum fluctuation, with its cosmic size and duration explained by the relation between ρ_c and G that he had struggled for so long to understand.

For pedagogical purposes, textbooks on quantum field theory discuss a few simple models which, for a variety of reasons, fail to describe physical reality. One such model contains a field ψ with a potential $V(\psi) = \lambda\psi^3$, where λ is a constant. This model is *immediately* rejected, because it has no state of minimum energy. (For $\lambda = \pm|\lambda|$, $V(\psi) \to -\infty$ as $\psi \to \mp\infty$). The existence of a state of minimum energy, called the *vacuum*, is traditionally regarded as essential to any realistic quantum field theory of elementary particles.

Consider now a distribution of particles with total mass M. Adding Einstein's Mc^2 to Newtonian approximations for kinetic and potential energies, one obtains a total energy

$$E = M\left(c^2 + \tfrac{1}{2}\langle v^2\rangle\right) - \tfrac{1}{2}GM^2 \times \langle 1/r\rangle, \tag{97}$$

where angled brackets denote average values (of velocity-squared and inverse separation, respectively). Note that E has no minimum value: $E \to -\infty$ as $M \to \infty$. *Gravitational potential energy is precisely the kind that has traditionally been rejected in quantum field theory*, because such a theory has no state of minimum energy. An initially empty state with $E = M = 0$ would spontaneously undergo quantum transitions to states with $E = 0$ but *arbitrary* $M > 0$, subject only to the constraint that

$$\langle 1/r\rangle = \frac{2\left(c^2 + \tfrac{1}{2}\langle v^2\rangle\right)}{GM} \tag{98}$$

Quantum field theory has no stable vacuum when gravity is included.

A vacuum fluctuation from $M = 0$ to a state with $M > 0$ would require a form of quantum tunneling, since $E > 0$ for values of M between 0 and the final value. Such tunneling is a completely standard feature of quantum theory, however. We have already considered one example, the inevitable tunneling of a Higgs field from the false vacuum of Fig. 4a to the true vacuum.

In quantum field theory, pairs of particles (e.g. electrons and positrons) are continuously appearing in the most empty space possible (Section IV.G). Energy conservation is violated by an amount ΔE for a duration Δt. This is permitted (actually *implied*) by quantum mechanical uncertainties, however, subject to the constraint that $\Delta E \Delta t \approx \hbar$. Eventually, by chance, enough particles would spontaneously appear close enough together for them to have zero net energy, because of cancellation between positive mass-energy and negative gravitational energy. Such a state would have $\Delta E = 0$, and could last forever.

Now recall Tryon's guess that the f in Eq. (92) might be the magnitude of the cosmic gravitational potential. This would require that $f \simeq 1/\sqrt{G}$ but, in a crude calculation assuming the density to be ρ_c, he had obtained $f \simeq \sqrt{3/G}$. Equations (92)–(94) then imply that $m_I c^2 + \mathrm{GPE} \approx 0$ for every piece of matter, i.e., *a universe with density near ρ_c could have zero energy*. Such a universe could have originated as a spontaneous vacuum fluctuation. Conversely, such an origin would explain *why* our universe has a density near ρ_c.

More generally, creation *ex nihilo* (from nothing) would imply that our universe has zero net values for all conserved quantities ("the quantum numbers of the vacuum"). The only apparent challenge to this prediction arises from the fact that our universe is strongly dominated by matter (i.e., antimatter is quite rare). Laboratory experiments have revealed that the symmetry between matter and antimatter is not exact, however, and most theories of elementary particles going beyond the standard model contain mechanisms for matter to have become dominant over antimatter at a very early time (*baryogenesis*, see Section VII.C).

The preceding discussion has used Newtonian approximations for energy, and these are only suggestive. We note, however, that Newtonian theory resembles GTR far more than one might guess. For a matter-dominated universe, Newton's and Einstein's physics predict the same rate of slowing for the expansion, as was shown in Section IV.D. As another example, the escape velocity from the surface of a planet of mass M and radius R is given by Newtonian mechanics as $v_{esc} = \sqrt{2GM/R}$. For $v_{esc} = c$, the relation between R and M is precisely that of GTR describing a black hole (where R is called the *Schwarzschild radius*).

The field equation (22) is of particular interest. With $a(t)$ denoting the cosmic scale factor and $\dot{a} \equiv da/dt$, Eq. (22) implies

$$(kc^2 + \dot{a}^2) - \frac{8\pi G(\rho a^3)}{3a} = 0 \qquad (99)$$

Now consider a closed, finite universe ($k = +1$). The proper volume of such a universe (Section III.C) is $V(t) = 2\pi^2 a^3(t)$, and the total mass-energy is $M(t) = \rho(t)V(t)$. [M is constant for zero pressure, but otherwise not; see Eq. (24).] With $r_p(t)$ denoting proper distance (Section III.C) and $\dot{r}_p \equiv dr_p/dt$ corresponding to velocity, Eq. (97) multiplied by M yields

$$M\left(c^2 + 0.36\langle \dot{r}_p^2 \rangle\right) - 0.55GM^2 \times \langle 1/r_p \rangle = 0, \qquad (100)$$

where $\dot{a}^2$ and $1/a$ have been expressed in terms of

$$\langle \dot{r}_p^2 \rangle = \frac{1}{V}\int dV\left(\dot{r}_p^2\right) = \frac{2\dot{a}^2}{\pi}\int_0^\pi du(u\sin u)^2, \qquad (101)$$

$$\langle 1/r_p \rangle = \frac{1}{V}\int dV\frac{1}{r_p} = \frac{2}{\pi a}\int_0^\pi du\frac{\sin^2 u}{u} \qquad (102)$$

(see Section III.C). Note that the density, pressure, and total mass M are arbitrary in the above analysis. [In particular, Eqs. (99)–(102) are valid in both standard and inflationary models.]

Apart from slight differences in numerical coefficients, the left side of Eq. (100) is identical in form to the right side of Eq. (97). This suggests that, in some sense, all closed universes have zero energy. This interpretation is buttressed by the fact that such universes are finite and self-contained, so that no gravitational flux lines could extend outward from them into any surrounding space in which they might be embedded. The energy of a universe has no rigorous definition within the context of GTR (Section IV.B), but the preceding observations are nevertheless quite suggestive.

We also note that if dominated by matter, a closed universe expands to a maximum size in a finite time and then collapses back to zero size (Section IV.C). This is the typical behavior of a vacuum fluctuation. For the above reasons, and also because physical processes are expected to give rise to finite systems, Tryon's conjecture of 1973 was accompanied by a prediction that our universe is closed and finite.

It is a remarkable fact that creation *ex nihilo* could have been proposed at any time since the early 1930s. All the theoretical ingredients had been in place. There had been a long period when the magnitude and grandeur of the cosmos had befogged the normally clear vision of physicists, however. A few examples will illustrate this point.

Before Hubble's discovery of the cosmic expansion in 1929, Einstein had believed the universe to be infinitely old and static. This view was incompatible with the physics of even his time, however, in at least three respects.

1. In 1823, Heinrich Olbers had pointed out that the entire night sky should be as bright as the surface of a typical star in such a model. (*Olbers' paradox*: His reasoning was that every line of sight would end on a star, a valid argument. In more modern terms, the

entire universe would have reached a state of thermal equilibrium, in which case all of space would be filled by light as intense as that emitted from the surfaces of stars.)
2. Conservation of energy would be grossly violated by stars that shone forever.
3. Over an infinite span of time, gravitational instabilities would have caused all stars to merge into enormous, dense concentrations of matter, as was evident from work published in 1902 by Sir James Jeans (the *Jeans instability*).

Even Einstein, with few equals in the history of science, felt these considerations to be outweighed by the philosophical appeal of an infinitely old and static universe.

Hubble's discovery of the cosmic expansion resulted in a period of intense speculation about the nature of the universe. In the 1930s, several inventors of cosmological models based their conjectures on a so-called "cosmological principle," namely that the universe must (or *should*) be homogeneous and isotropic. This was extended in the late 1940s to the "perfect cosmological principle," meaning that the universe must (or *should*) also be eternal and unchanging. The steady-state model of cosmology was based on this "principle," and was widely favored in professional circles until empirical evidence undermined it in the 1960s. We remark here (as occasional skeptics did at the time) that both of the aforementioned "principles" were of a highly dubious nature, because (1) nothing supported them except observations which they pretended to explain, and (2) they had no *generality*, referring only to a single system (the entire universe).

The preceding examples are not meant as criticisms of the brilliant scientists who held such views, but rather as illustrations of how the mystique of the cosmos exempted it from the objective standards routinely applied to lesser systems.

Early in his graduate studies (about 1963), Tryon attended a colloquium given by a 30-year-old professor then unknown to him. In straightforward and compelling terms, this professor elucidated several ways in which neutrinos might play critical roles in astrophysics. The lecture inspired Tryon, and transformed his scientific worldview. For the first time, he realized that enormous structures, perhaps even the entire universe, could (and *should*) be analyzed in terms of the same principles that govern phenomena on earth, *including the principles of microscopic physics*. The professor's name was Steven Weinberg. It was Tryon's great good fortune that Weinberg later accepted him as a thesis student. In retrospect, Tryon has long felt deeply indebted to Weinberg for opening his mind to the possibility that the entire universe might have arisen as a spontaneous vacuum fluctuation.

We come now to the second striking feature shared by the psychological and theoretical breakthroughs of Guth and Tryon. In his highly engaging history and survey of inflation (see bibliography), Guth describes how he became involved in the research that culminated in his discovery.

In early 1979, Guth was a young research physicist at Cornell University, preoccupied with issues unrelated to his later work. He has written that in April of 1979,

> "my attitude was changed by a visit to Cornell by Steven Weinberg . . . . Steve gave two lectures on how grand unified theories can perhaps explain why the universe contains an excess of matter over antimatter . . . . For me, Weinberg's visit had a tremendous impact. I was shown that a respectable (and even a highly respected) scientist could think about things as crazy as grand unified theories and the universe at 10^{-39} sec. And I was shown that really exciting questions could be asked."

The day after Weinberg's visit ended, Guth immersed himself in a study of GUTs in the early universe. By year's end, he had discovered the possibility of cosmic inflation, and he knew immediately that he should take the possibility seriously.

In truth, Weinberg influenced and inspired a generation of young physicists, the first generation entirely liberated from the disorienting mystique of the cosmos. Through widespread lectures, writing, and the extraordinary power and clarity of his vision, Weinberg emboldened a generation to apply the principles of microscopic physics to the early universe.

We have now reached the extraordinary position where the following are regarded as plausible, even likely in the minds of many: (1) the universe originated as a spontaneous quantum fluctuation, (2) quantum fluctuations in the false vacuum lead to eternal inflation and endless creation, and (3) quantum fluctuations in the false vacuum were the primordial deviations from homogeneity that evolved into galaxies and the largest structures in the universe. These and related ideas are sometimes referred to as "quantum cosmology," which is surely among the most striking paradigm shifts in the history of science. Steven Weinberg may fairly be regarded as the spiritual father of quantum cosmology.

One can only speculate about the stage upon which our universe was born. If it was infinite in extent (in space and/or "time"), however, vacuum fluctuations giving rise to universes like ours would be inevitable if the probability were different from zero (no matter how small) on a finite stage. Indeed, such creations would be more numerous than any finite number—such would be the nature of an infinite stage. It is nevertheless true that this quantum theory of creation became much more widely accepted

after Guth's proposal of inflation, because inflation explains how even a microscopic fluctuation could balloon into a cosmos.

A mathematically detailed scenario for creation *ex nihilo* was published in 1978 by R. Brout, F. Englert, and E. Gunzig. This model assumed a large, negative pressure for the primordial state of matter, giving rise to exponential growth that converted an initial microsciopic quantum fluctuation into an open universe. This model was a significant precursor to inflationary scenarios based on GUTs, as well as being a model for the origin of the universe.

During the years since Guth's proposal of inflation, numerous creation scenarios have been published, differing primarily in their details. An especially interesting and simple one is that of Alexander Vilenkin (1982). Vilenkin sought to eliminate the need for *any* kind of stage upon which creation occurred, i.e., he sought a more literal creation from *nothing*.

Vilenkin noted that a closed universe has finite volume: $V(t) = 2\pi^2 a^3(t)$. For $a = 0$, the volume is zero and, if such a universe is completely empty, it is then a candidate for true "nothingness"—a mere abstraction. He proposed that such an abstraction could experience quantum tunneling to a microscopic volume filled with false vacuum, which would then inflate to cosmic proportions. We do not (and may never) know whether our universe was born within some larger physical reality or emerged from a pure abstraction, but both scenarios seem plausible.

Creation *ex nihilo* suggests (without implying) a radical idealism, in the sense of ideals being abstract principles. To illustrate the point, it may be helpful to consider the alternative (and usual) view.

Well before a child acquires language and the ability to think abstractly, it has become aware of tangible matter in its surroundings. As the child develops, it furthermore discerns patterns of regularity—splinters hurt, chocolate tastes good, and day alternates with night. Given that matter exists, it must behave in *some* way, and one learns from experience and perhaps the sciences how very regular that behavior is. The unarticulated hierarchy in one's mind is (1) matter exists, (2) it must do *something*, and (3) it seems to follow some set of rules. Matter is the *fundamental* reality, without which it would make no sense for there to be principles governing it.

The maximal version of creation *ex nihilo* inverts this hierarchy: abstract principles (*ideals*) are the fundamental reality. Tangible matter and energy are *manifestations of a deeper reality*, the principles that gave rise to their existence. This remains, of course, an open question, but it is one of the many issues that quantum cosmology has given us to contemplate.

SEE ALSO THE FOLLOWING ARTICLES

CELESTIAL MECHANICS • COSMIC RADIATION • COSMOLOGY • DARK MATTER IN THE UNIVERSE • GALACTIC STRUCTURE AND EVOLUTION • HEAT TRANSFER • QUANTUM THEORY • RELATIVITY, GENERAL • RELATIVITY, SPECIAL • STELLAR STRUCTURE AND EVOLUTION • UNIFIED FIELD THEORIES

BIBLIOGRAPHY

Guth, A. (1997). "The Inflationary Universe," Perseus Books, New York.
Krauss, L. M. (2000). "Quintessence: The Mystery of the Missing Mass in the Universe," Basic Books, New York.
Liddle, A. R. (1999). "An Introduction to Modern Cosmology," Wiley, New York.
Liddle, A. R., and Lyth, D. H. (2000). "Cosmological Inflation and Large-Scale Structure," Cambridge Univ. Press, Cambridge, U.K.
Linde, A. D. (1990). "Inflation and Quantum Cosmology," Academic Press, San Diego.
Peacock, J. A. (1999). "Cosmological Physics," Cambridge Univ. Press, Cambridge, U.K.
Peebles, P. J. E. (1993). "Principles of Physical Cosmology," Princeton Univ. Press, Princeton, New Jersey.
Weinberg, S. (1972). "Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity," Wiley, New York.
Weinberg, S. (1977). "The First Three Minutes," Basic Books, New York.
Weinberg, S. (1992). "Dreams of a Final Theory," Pantheon Books, New York.

Dark Matter in the Universe

Steven N. Shore

Indiana University

Virginia Trimble

University of California, Irvine, and University of Maryland, College Park

GLOSSARY

Baryons Protons and neutrons that are the basic constituents of luminous objects and which take part in nuclear reactions. These are strongly interacting particles, which feel the electromagnetic and nuclear, or strong, forces.

Cosmic background radiation (CBR) Relic radiation from the Big Bang, currently having a temperature of 2.74 K. It separated out from the matter at the epoch at which the opacity to scattering of the universe, because of cooling and expansion, fell to small enough values that the probability of interaction between the matter and the primordial photons became small compared with unity. It is essentially an isotropic background and its temperature (intensity) fluctuations serve to place limits on the scale of the density fluctuations in the universe at a redshift of about 1000. The redshift is defined as $\lambda/\lambda_0 = 1 + z$, where λ is the currently observed wavelength of a photon and λ_0 is the wavelength at which it was emitted from a distant object.

Deceleration parameter Constant that measures both the density of the universe, compared with the critical density necessary for a flat space–time metric (called $\Omega = \rho/\rho_c$), and the departure of the velocity–redshift relation from linearity (q_0).

Faber–Jackson relation Relation between the bolometric luminosity of a galaxy and its core velocity

dispersion. This permits the determination of the intrinsic luminosity of elliptical galaxies without the need to assume a population model for the galaxy. It is used in the measurement of distances to galaxies independent of their Hubble types.

Hubble constant (H_0) Constant of proportionality for the rate of velocity of recession of galaxies as a function of redshift. Its current value is 71 ± 8 km sec^{-1} Mpc^{-1}. Often, in order to scale results to this empirically determined constant, it is quoted as $h = H_0/(100$ km sec^{-1} Mpc$^{-1})$. The inverse of the Hubble constant is a measure of the age of the universe, called the *Hubble time*, and is about 1.5×10^{10} years.

Leptons Lightest particles, especially the electron, muon, and tau and their associated neutrinos. These particles interact via the weak and electromagnetic forces (electroweak).

Tully–Fisher relation Relation between the maximum orbital velocity observed for a galaxy's 21-cm neutral hydrogen (the line width) and the bolometric luminosity of the galaxy.

Units Solar mass ($M_\odot$), 2×10^{33} g; solar luminosity ($L_\odot$), 4×10^{33} erg sec^{-1}; parsec (pc), 3×10^{18} cm (3.26 light years).

DARK MATTER is the subluminous matter currently required to explain the mass defect observed from visible material on scales ranging from galaxies to superclusters of galaxies. The evidence shows that the need for some form of invisible but gravitationally influential matter is present on length scales larger than the distance between stars in a galaxy. This article examines the methods used for determining galaxies and clusters of galaxies and discusses the cosmological implications of various explanations of dark matter.

I. HISTORICAL PROLOGUE

The existence of a species of matter that can not be seen but merely serves some mechanical function in the universe can be traced to Aristotelian physics and was clerly supported throughout the development of physical models-prior to relativity. The concept of the *aether*, at first something apart from normal material in being imponderable eventually metamorphosed into that of a fluid medium capable of supporting gravitation and electromagnetic radiation. However, since it was the medium that was responsible both for generating and transmitting such forces, its nature was intimately tied to the overall structure of the universe and could not be separated from it. It had to be assumed and could not easily be studied. Nineteenth century attempts, using high-precision measurements, to observe the anisotropy of the propagation of light because of the motion of the earth, the Michelson–Morley experiment being the best known, showed that the aether's mechanical properties could not conform to those normally associated with terrestrial fluids.

A version of the search for some cosmic form of dark matter began in the late 17th century with Halley's question of the origin of the darkness of the night sky. Later enunciated in the 19th century as *Olber's paradox,* it required an understanding of the flux law for luminous matter and was argued as follows: in an infinite universe, with an infinite number of luminous bodies, the night sky should be at least as bright as the surface of a star, if not infinitely bright. Much as in current work on dark matter, the argument was based on the assumption of the premise of a cosmological model and of the dynamical (for in this case phenomenological), character of a particular observable.

In order to circumvent the ramifications of the paradox without questioning its basic premises, F. Struve, in the 1840s, argued for the existence of a new kind of matter in the cosmos, one capable of extinguishing the light of the stars as a function of their distance, hence path length through the medium. The discovery of dark nebulae by Barnard and Wolf at the turn of the century, of stellar extinction by Kapteyn about a decade later, and of the reddening of distant globular clusters by Trumpler in the 1930s served to support the contention that this new class of matter was an effective solution to the *missing light* problem. However, as J Hershel realized very soon after Struve's original suggestion of interstellar dust, this cannot be a viable solution for the paradox. In an infinite universe, the initially cold, dark matter would eventually come into thermal equilibrium. The resulting glow would eventually reach, if not the intensity of the light of a stellar surface, an intensity still substantially above the levels required to reproduce the darkness of the sky between the stars.

An additional thread in this history is the explanation of the distribution of the nebulae, those objects we now know to be extragalactic. In detailed statistical studies. Charlier, Lundmark, and later Hubble, among others, noted a *zone of avoidance,* which was located within some 20° of the galactic plane. Several new forces were postulated to explain this behavior, all of which were removed by the discovery of the particulate nature of the dust of the interstellar medium and the expansion of the universe. However, the explanatory power of the concepts of dark matter and unknown forces of nature serves as a prototype for many, of the current questions, although the circumstances and nature of the evidence have dramatically altered in time.

The origin of the current problem of dark matter (hereafter called *DM*) really starts in the late 1930s with the discovery of galaxy clusters by Shapley. The fact that the

universe should be filled with galaxies that are riding on the expanding space–time was not seen as a serious problem for early cosmology, but the fact that these objects cluster seems quite difficult to understand. Their distribution appears to be very anisotropic, and was seen at the time to be the evidence that the initial conditions of the expansion may be drastically altered by later gravitational developments of the universe.

About the same time, Zwicky, analyzing the distribution of velocities in clusters of galaxies, argued that the structures on scales of megaparsecs (Mpc) cannot be bound without the assumption of substantial amounts of nonluminous material being bound along with the galaxies. His was also the first attempt to apply the virial theorem to the analysis of the mass of such structures. Oort, in his analysis of the gravitational acceleration perpendicular to the galactic plane, also concluded that only about half of the total mass was in the form of visible stars.

With the advent of large-scale surveys of the velocities of galaxies in distant clusters, the problem of DM has become central to cosmology. It is the purpose of this article to examine some of the issues connected with current cosmological requirements for some form of nonluminous, gravitationally interacting matter and the evidence of its presence on scales from galactic to supercluster.

II. KINEMATIC STUDIES OF THE GALAXY

The Milky Way Galaxy (the Galaxy) is a spiral, consisting of a central spheroid and a disk of stars extending to several tens of kiloparsecs. A halo is also present and appears to envelope the entire system extending to distances of over 50 kpc from the center. It is this halo that is assumed to be the seat of most of the DM. Since it is essentially spherical, it contributes to the mass that lies interior to any radius and also to the vertical acceleration. The stars of which it is composed have high-velocity dispersions, of order 150 km sec^{-1}, and are therefore distributed over a larger volume while still bound to the Milky Way.

Galactic studies of dark matter come in two varieties: local, meaning the solar neighborhood (a region of several hundred parsecs radius), and large-scale, on distances of many kiloparsecs. The study of the kinematics of stars in the Galaxy produced the first evidence for "missing mass," and there is more detailed information available for the Milky Way than for any other system. Further, many of the methods that have been applied, or will be applied in future years, to extragalactic systems have been developed for the Galaxy. So that the bases of the evidence can be better understood, this section will be a more detailed discussion of some of the methods used for the galactic mass determination.

A. Local Evidence

The local evidence for DM in the Galaxy comes from the study of the vertical acceleration relative to the galactic plane, the so-called "Oort criterion." This uses the fact that the gravitational acceleration perpendicular to the plane of the Galaxy structures the stellar distribution in that direction and is dependent on the surface density of material in the plane. The basic assumption is that the stars, which are collisionless, form a sort of "gas" whose vertical extent is dependent on the velocity dispersion and the gravitational acceleration. For instance, taking the pressure of the stars to be $\rho\sigma^2$, where ρ is the stellar density and σ is their velocity dispersion in the z direction, then vertical hydrostatic equilibrium in the presence of an acceleration K_z gives the scale height of an "isothermal" distribution:

$$z_0 \approx \sigma^2 \big/ K_z. \tag{1}$$

This represents the mean height that stars, in their oscillations through the plane, will reach. It is sometimes called the *scale height* of the stellar distribution.

The gravitational potential is determined by the Poisson equation:

$$\nabla^2 \Phi = -4\pi G\rho = \nabla \cdot \mathbf{K}, \tag{2}$$

where G is the gravitational constant and $\mathbf{K}$ is the total gravitational acceleration. Separation of this equation into radial and vertical components, assuming that the system is planar and axisymmetric, permits the determination of the vertical acceleration. It can be assumed that the stars in the plane are orbiting the central portions of the galaxy, and that

$$K_r = \frac{\Theta^2(r)}{r}. \tag{3}$$

Here, Θ is the circular velocity at distance r from the galactic center. Under this assumption, the Poisson equation becomes

$$\frac{\partial}{\partial z} K_z = -4\pi G\rho + \frac{2\Theta}{r}\,\frac{d\Theta}{dr}, \tag{4}$$

where the radial gradient of the circular velocity is determined from observation. From the observation of the radial gradient in the rotation curve, it is possible to determine the vertical acceleration from dynamics of stars above the plane and thus determine the space mass density in the solar neighborhood.

The argument proceeds using the fact that the vertical motion of a star through the plane is that of a harmonic oscillator. The maximum z velocity occurs when the star passes through the plane. From the vertical gradient in the z component of the motion, V_z, one obtains the acceleration, and the mass density follows from this. Assuming that stars start from free fall through the plane, their total energy is given by

$$\Phi_0 + \tfrac{1}{2}V_{z,0}^2 = \tfrac{1}{2}V_z^2, \tag{5}$$

so that assuming that $V_{z,0} = 0$, in other words, that the star falls from a large distance having started at rest, it follows that the maximum velocity through the plane of a galaxy is determined by the potential, thus the acceleration, at the extremum of the orbit (apogalactic distance). The vertical acceleration is a function of the surface mass density. Thus, star counts can be used to constrain, from luminous material, the mass in the solar neighborhood. Note that $K_z \sim -4\pi G\Sigma$, where Σ is the surface density, given by

$$\Sigma = \int_0^z \rho(r, z')\, dz'. \tag{6}$$

It should therefore be possible to determine the gravitational acceleration from observations of the vertical distribution of stars, obtain the required surface density, and compare this with the observations of stars in the plane.

The mass required to explain K_z observed in the solar neighborhood, about $0.15 M_\odot$ pc^{-3}, exceeds the observed mass by a factor of about two. This is not likely made up by the presence of numerous black holes or more exotic objects in the plane since, for instance, the stability of the solar system is such as to preclude the numerous encounters with such gravitational objects in its lifetime. It appears that there must be a considerable amount of uniformly distributed, low-mass objects or some other explanation for the mass.

Recent observations with the FUSE satellite in the ultraviolet and with ISO in the infrared reveal significant abundances of H_2 in the interstellar medium. ISO data for the edge-on galaxy NGC 891 show enough in this cold and shocked gas to account for the *disk* (only) component of the missing mass. It thus appears plausible that for the relatively low mass component of disk galaxies—the disk—there is no evidence for a low-velocity exotic source of dark matter. The same, however, is not true for the halo, as we will soon discuss.

The velocity dispersion of stars is given by the gravitational potential in the absence of collisions. Thus, from the determination of the mass density in the solar neighborhood with distance off the plane, the vertical component of the gravitational acceleration can be found, leading to a comparison with the velocity dispersion of stars in the z direction. The resulting comparison of the dynamic and static distributions determines the amount of mass required for the acceleration versus the amount of mass observed.

Most determinations of K_z find that only one-third to one-half of the mass required for the dynamics is observed in the form of luminous matter. In spite of several attempts to account for this mass in extremely low luminosity M dwarf stars, the lowest end of the mass distribution of stellar masses, the problem remains. The mass of the observed halo is not sufficient to explain the vertical acceleration, and the mass function does not appear to extend to sufficiently low masses to solve this problem.

Another local determination of the mass of the Galaxy, this time in the plane, can be made from the observation of the maximum velocity, relative to the so-called local standard of rest, of stars orbiting the galactic center. The argument is best illustrated for a point mass. In the case of a circular orbit about a point, the orbital velocity, Θ, is given by centrifugal acceleration being balanced by the gravitational attraction of the central mass,

$$\Theta(r) = \left(\frac{GM}{r}\right)^{1/2}, \tag{7}$$

and it is easily seen that the escape velocity is $v_{\rm esc}(r) = \sqrt{2}\Theta(r)$ for all radii. The coefficient is slightly lower in the case of an extended mass distribution, but the argument follows the same basic form. The local standard of rest (LSR) is found from the velocity of the sun relative to the most distant objects, galaxies and quasi-stellar objects (QSOs), and the globular cluster system of our galaxy. The maximum orbital velocity observed in the solar neighborhood should then be about $0.4 v_{\rm LSR}$, or in the case of our system about 70 km sec^{-1} in the direction of solar motion. For escape from the system, at a distance of 8.5 kpc with a $\Theta_0 = 250$ km sec^{-1}, the escape velocity from the solar circle should be only about 300 km sec^{-1}. The mass of the Galaxy can be determined from the knowledge of the distance of the sun from the galactic center (determined from mass distributions and from the brightness of variable stars in globular clusters) and from the shape of the rotation curve as a function of distance from the galactic center.

B. Large-Scale Determinations

From the stellar orbits, one knows the rate of galactic differential rotation in the vicinity of the sun, a region about 3 kpc in radius and located about 8.5 kpc from the galactic center. The measurement of galactic rotational motion can be greatly extended through the use of millimeter molecular line observations, such as OH masers, and the 21-cm line of neutral hydrogen. Most of the disk is accessible, but at the expense of a new assumption. It is assumed that the gas motions are good tracers of the gravitational field and that random motions are small and unimportant in comparison with the rotational motions. It is also assumed that the gas is coplanar with the stellar distribution and that the motion is largely circular, with no large-scale, noncircular flows being superimposed. These assumptions are not obviously violated in the Galaxy, but may be problematic in many extragalactic systems, especially barred spiral

galaxies. The maximum radial velocity along lines of sight interior to the solar radius occurs at the *tangent points* to the orbits. This method of tangent points presumes that the differential rotation of the Galaxy produces a gradient in the radial velocity along any line of sight through the disk for gas and stars viewed from the sun and that the maximum velocity occurs at a point at which the line of sight is tangent to some orbit. The distance can be determined from knowledge of the distance of the sun from the galactic center, about 8 kpc. Using $\Theta_{\rm HI}$ gives a measure of the mass interior to the point r and thus the cumulative mass of the Galaxy. Recent work has extended this to include large molecular clouds, which also orbit the galactic center.

The determination of the orbits and distances of stars and gas clouds outside of the solar orbit is difficult since the tangent point method cannot be applied for extrasolar orbits, but from the study of molecular clouds and 21-cm absorption and stellar velocities of standard stars it appears that the rotation curve outside of the solar circle is still flat, or perhaps rising. The argument that this implies a considerable amount of extended, dark mass follows from an extension of the argument for the orbital velocity about a point mass, $\Theta^2(r) \sim M(r)/r$. Since the observations support $\Theta = \text{const}$, it appears that $M(r) \sim r$. The scale length for the luminous matter is small, about 5 kpc, and this is substantially smaller than the radial distances where the rotation curve has been observed to still be flat (of order 15 kpc). The same behavior has been observed in external galaxies, as we will discuss shortly. Thus, there is a substantial need for the Galaxy to have a large amount of dark matter in order to account for dynamics on scales larger than 10 kpc.

As mentioned, mass measurements of the Galaxy from stellar and gas rotation curves beyond the distance of the sun from the center are subject to several serious problems, which serve as warnings for extragalactic studies. One is that there appear to be substantial departures of the distribution from planarity. That is, the outer parts of the disk appear to be warped. This means that much of the motion may not be circular and there may be sizable vertical motions, which means that the motions are not strictly indicative of the mass. For the inner galactic plane, there is some evidence of noncircular motion perhaps caused by barlike structures in the mass distribution, thus affecting the mass studies.

An extension of the Oort method for the vertical acceleration, now taken to the halo stars, can be used to measure the total mass of the Galaxy. One looks at progressively more distant objects. These give handles on several quantities. For instance, the maximum velocity of stars falling through the plane can be compared with the maximum distances to which stars are observed to be bound to the galactic system. Halo stars of known intrinsic luminosity, such as RR Lyrae variable stars, can be studied with some certainty in their intrinsic brightness. From the observed brightness, the distance can be calculated. Thus, the distance within the halo can be found. From an observation of the vertical component of the velocity, one can obtain the total mass of the galactic system lying interior to the star's orbit. The same can be done for the globular clusters, the most distant of which should be nearly radially infalling toward the galactic center. These methods give a wide range of masses, from about $4 \times 10^{11} M_\odot$ to as high as $10^{12} M_\odot$. to distances of 50–100 kpc.

The phenomenological solution to the problem of subluminous mass increasing in fraction of the galactic population with increasing length scale measured assumes that the mass-to-light ratio, M/L, is a function of distance from the galactic center. For low-mass stars, this number is of order 1–5 in solar units ($M_\odot/L_\odot$), but for the required velocity distribution in the galaxy, this must be as high as 100. There are few objects, except Jupiter-sized or smaller masses, which have this property. The reason is simple—nuclear reactions that are responsible for the luminosity of massive objects, greater than about $0.08 M_\odot$, cannot be so inefficient as to produce this high value. Even the fact that the flux distribution can be redistributed in wavelength because of the effects of opacity in the atmospheres of stars of very low mass will still leave them bright enough to observe in the infrared, if not the visible, and their total (bolometric) luminosities are still high enough to provide M/L ratios that are less than about 10. Black holes, neutron stars, or cold white dwarfs would appear good candidates within the Galaxy. But limits can be set on the space density of these objects from X-ray surveys since they would accerete material from the interstellar medium and emit X rays as a result of the infall. The observed X-ray background and the known rate of production of such relics of stellar evolution are both too low to allow these objects to serve as the sole solution to the disk DM problem. Some other explanation is required.

III. LARGE-SCALE STRUCTURE OF THE GALAXY

A. Multiwavelength Observations

1. Radio and Far-Infrared (FIR) Observations

Neutral and molecular hydrogen (H I and H_2), form only a small component of the total mass of the galactic system. Both 21-cm and CO (millimeter) observations can only account for about one-third of the total galactic mass being in the form of diffuse or cloud matter. The IRAS satellite, which performed an all-sky survey between 12 and

100 μm, did not find a significant population of optically unseen but FIR bright point sources in the galaxy. Instead, it showed that the emission from dust in the diffuse interstellar medium is consistent with the amount of neutral gas present in the plane and that there is not a very sizable component of the galaxy at high-galactic latitude. This severely constrains conventional explanations for the DM since any object would come into equilibrium at temperatures of the same order as an interstellar grain and would likely be seen by the IRAS satellite observations. As we mentioned, however, spectra obtained with ISO show diffuse molecular gas in unexpectedly large abundance in spirals.

2. X-Ray Observations

These show the interstellar medium to possess a very hot, supernova-heated phase with kinetic temperatures of order 10^6–10^7 K. But here again it is only a small fraction of the total mass of luminous matter. The spatial extent of this gas is greater than that revealed by the IR and radio searches but is still consistent with the galactic potential derived from the optical studies. As mentioned, X-ray observations also constrain the space density of collapsed but subluminous objects, such as black holes, through limits on the background produced by such objects accreting interstellar gas and dust.

3. Ultraviolet Observations

The stars that produce the UV radiation are the most massive or most evolved members of the disk, being either young stars or the central stars of planetary nebulae. They have very low M/L ratios, are easily seen in the UV, and would not be likely candidates for DM. There is no compelling evidence from other wavelength regions that the missing matter is explained by relatively ordinary matter. Baryonic matter at temperatures from 10^7 K to below the microwave background of 3 K can be ruled out by the currently available observations.

B. The Components of the Galactic Disk and Halo

Star counts as a function of magnitude in specific directions (that is, as a function of $l_{\rm II}$ and $b_{\rm II}$, the galactic longitude and latitude, respectively) provide the best information on the mass of the visible matter and the structure of the galaxy. This has been discussed most completely by Bahcall and Soniera (1981). It depends on the luminosity function for stars, their distribution in spectral type, and their evolution. However, in the end, it provides more of a constraint on the evolution of the stellar population than on the question of whether the system is supported by a substantial fraction of dark matter.

Although such studies concentrate on the luminous matter, the reckoning of scale heights for the different stellar populations provides an estimating method for the vertical and radial components of the acceleration—direct mass modeling. Such studies are also sensitive to the details of reddening from dust both in and off of the galactic plane, metallicity effects as a function of distance from the galactic center, and evolutionary effects connected with the processes of star formation in different parts of the galaxy. Supplementing the space density studies with space motions (proper motion, tangential to the line of sight, can be found for some classes of high-velocity stars and can be added to the radial velocities to give an overall picture of the stellar kinematics) permits both a kinematic and photometric determination of the galactic mass. These studies have extended the determination of the vertical acceleration to distances of more than 50 kpc into the halo.

Another constraint on the mass of the Galaxy comes from the tidal radii of the Magellanic Clouds and from the masses and mass distributions of globular clusters. The Magellanic Clouds (the Clouds) are fairly large compared with their distances from the galactic center. They therefore feel a differential gravitational acceleration that counteracts their intrinsic self-gravity and tends to rip them apart as they move around the Galaxy. The maximum distance that a star can be from the Clouds before it is more bound to the Galaxy than the Clouds, or conversely the minimum distance to which the Clouds can approach the Galaxy without suffering tidal disruption, provides a measure of the total mass of the Galaxy. The distribution of the globular clusters does not provide any strong support for a DM halo. In addition, the clusters appear to be bound, even though they are the highest mass separate components of the galactic system, without involking any nonluminous mass. In order to account for the dynamics and to be consistent with the known ages of the clusters, the typical M/L ratio is about 3, characteristic of low-mass stars. So it appears that objects of order $10^6-10^8 M_\odot$, and with sizes of up to 100 pc, are *not* composed of large quantities of dark mass.

C. Theoretical Studies

The spiral structure of disk galaxies has been a long-standing problem since the discovery of the extragalactic nature of these objects. The first suggestion that the patterns seen in spirals might be due to some form of intrinsic collective mode of a disk gravitational instability was from Lindblad and was elaborated by Lin and Shu (1965) as the "quasistationary density wave" model. The picture requires a collective mode in which the stars in the

disk behave like a self-gravitating but collisionless fluid (there is no real viscous dissipation), which clumps in a spatially periodic wave. This wave of density feeds back into the gravitational potential, which then serves as the means for supporting the disk wave structure.

It was soon realized that these waves are not stationary. Further, because of angular momentum transport by tidal coupling of the inner and outer parts of the disk, the waves will wind up and dissipate in the absence of a continuing forcing. Halos have been shown to help stabilize the disk, suggesting that perhaps the very existence of spiral galaxies is an indication of some DM being distributed over a large volume of the Galaxy. Ostriker and Peebles (1973), by analysis of the stability of disks against collapse into barlike configurations, came up with an independent argument supporting the need for an extensive halo having at least two-thirds the total mass of a disk galaxy. Their criterion was established by the stability analysis of a self-gravitating ellipsoid, showing that a disk is unstable to the formation of a bar if the ratio of the kinetic energy to the gravitational potential is large, that is, if $T/W > 0.14$. Here, T is the kinetic energy of the stars and W is the gravitational potential energy. This was perhaps the first paper to argue for the existence of a hot halo on the basis of the stability of a simple model system. Further modeling supports this conclusion, which now forms one of the cornerstones of galactic structure models: if the halo is not massive enough to stabilize the disk of a spiral galaxy, the system will collapse to a bar embedded in a more extended spheroid of stars.

The implication is that, since there are many disks that do not possess large-scale, barlike structures, there may be a halo associated with many of these systems. In addition, the fact that a disk is observed is indication enough that there should be a considerable quantity of mass associated with it. Thus, the search was stimulated for rotation curves that remain flat or non-Keplerian (not point-mass-like) at large distances from the galactic center.

IV. DARK MATTER IN THE LOCAL GROUP

A. Magellanic Clouds

The total mass of the Galaxy can be determined from the stability of the Magellanic Clouds. These two Irr galaxies orbit the Milky Way, being slowly tidally disrupted by the interaction with the disk of the Galaxy. The tidal radius of the Clouds and the fact that they have been stable and bound to each other for far longer than a single orbit (these satellites of the Galaxy appear as old as the Milky Way) provide an upper limit to the mass of the Galaxy and Clouds to a distance of about 60 kpc. This is not far from that obtained from the study of halo stars and clusters, and provides an upper limit of about $10^{12} M_{\odot}$. The required M/L is thus about 30. The total mass of the Magellanic Clouds may be individually underestimated, but this appears to be a good limit on the total mass of the system and is in qualitative agreement with the mass required to explain the dynamics of the Local Group (the galaxy cluster to which the Galaxy belongs) as well.

B. M 31 and the Rotation Curves of External Galaxies

The study of the rotation curve of the Andromeda Galaxy, M 31, the nearest spiral galaxy to ours and one which forms a sort of binary system with the Milky Way, shows that the rotation curve is flat to the distance at which the stars are too faint to determine the rotation curve. The M/L ratio is close to 30, about the same as that found for the Galaxy. In order to place this result in context, it is necessary to describe the process whereby the rotation curves are determined for extragalactic systems.

First, one can assume that the disk is radially supported by the revolution of collisionless stars about the mass lying interior to their radial distance from the galactic center. This assumption makes the mass within a distance r, $M = M(r)$, a function of radius, and then allows the determination of the orbital velocity assuming that the mean eccentricity of the orbits is small or zero. A slit is placed along, and at various angles to, the symmetry axes of the galaxy, and the radial velocity of the stars is determined. The disk is assumed to be circular, so the inclination of the galaxy can be determined from the ellipticity of the projected disk in the line of sight. There are several methods for fitting the mass model to this rotation curve, most of which assume a power-law form for the rotation curve with distance and fit the coefficients of the density distribution required to produce the observed velocities by the solution of the dynamical equations in the radial direction. This density profile is then integrated to give the mass interior to a given radius.

Several assumptions go into this method, most of which appear theoretically well justified. The first is that the system is collisionless. There is little evidence that the stars in a galaxy undergo close encounters—they are simply too small in comparison with the distances that separate them. However, there are very massive objects present in the disk, the Giant Molecular Clouds, which may have masses as great as $10^7 M_{\odot}$ and which appear to control the velocity dispersion of the constituent stars. These clouds are also the sites for star formation, and as such they form the basis for the interaction of the gas and star components of the Galaxy. They may also transfer angular momentum through the Galaxy, making the disk appear viscous. Such

interactions act much like encounters to change the velocity distribution to something capable of supporting a flat rotation curve. However, this remains speculation.

The observation of the rotation curve using neutral hydrogen line emission is a better way of determining the mass of a galaxy at a large distance. First, it appears that the H I distribution is considerably more extended than that of the bright stars; H I maps extend, at times, to four or five times the optical radius of the disk. Even at these large distances, studies show that the rotation curves remain flat! This is a very difficult observation to understand if the sole means of support for the rotation is the mechanism just described. Rather, it would seem that some form of extended mass distribution of very high M/L objects is required. The radial extent of such masses may be as great as 100 kpc, where the optical radii of these galaxies is smaller than 30 kpc.

The measurement of the rotation curve for the Local Group galaxies is complicated by the fact that the systems are quite close and therefore require low-resolution observations to determine the global rotation curves. This is also an aid in that the small-scale structure can be studied in detail in order to determine the effects of departures from circular orbits on the final mass determination.

The study of radial variations of velocity curves is essentially the same for external galaxies as for ours. One assumes that there is a density distribution of the form $\rho_*(r, z)$ of the stars and that this is the only contributor to the rotation curve for the Galaxy. One then obtains $\Theta(r)$, from which one can solve the equation of motion for radial support alone:

$$\Theta^2 = 4\pi(1 - e^2)^{1/2} \int_0^r \frac{\rho(\xi)\xi^2\, d\xi}{(r^2 - \xi^2 e^2)^{1/2}}, \qquad (8)$$

where e is the eccentricity of the spheroid assumed for the mass distribution. One can then expand ρ in a power series of the form

$$\rho(r) = \frac{a_{-2}}{r^2} + \frac{a_{-1}}{r} + a_0 + \cdots \qquad (9)$$

and perform a similar expansion for the velocity at a given distance. The solution is then mapped term by term. Once finished, one can then integrate the entire mass of the disk by assuming that the spheroid is homogeneous, so that

$$M(r) = 4\pi(1 - e^2)^{1/2} \int_0^r \rho(\xi)\xi^2\, d\xi. \qquad (10)$$

Such models are, of course, extremely simplified (the method was introduced by Schmidt in the 1950s and is still useful for exposition), but they can illustrate the problem of the various assumptions one must make in obtaining quantitative estimates of the mass from the rotation curve. Since the rotation curves remain flat, one can also assume that the M/L ratio is a function of distance from the galactic center. This is folded into the final result after the mass to a given radius has been determined.

Unlike our system, where one cannot be sure what is or is not a halo star, one can determine with some ease the velocity dispersions for face-on galaxies perpendicular to the galactic plane with the distance from the center of the system. In a procedure much like the Oort method described for the Galaxy, it is possible to place constraints on the amount of missing mass in the disk by using the determination of surface densities from the dynamical information. External galaxies appear to have values similar to our system for the dark matter in the plane, about one-half of the matter being observed in stars.

Many disk systems show warps in their peripheries; M 31, the Galaxy, and other systems show distortions of both their stellar and H I distributions in a bending of the plane on opposite parts of the disk. This argues that the halos of these galaxies cannot be too extensive, beyond the observable light, or one would not be able to observe such warps. Calculations show that these will damp, from dynamical friction (dissipation) and also viscous effects, if the halo is too extensive. However, in the case of some of the current models for the distribution of DM in clusters, it is not clear whether this can also constrain a more evenly distributed DM component (one less tied to the galaxies but more evenly distributed in the clusters than the halos).

V. BINARY GALAXIES

Binary galaxies may also be used to determine the M/L of the DM. The process is much like that employed for a binary star, with the exception that the galaxies do not actually execute an orbit within the time scale or our observations. Instead, one must perform the determination statistically. For a range of separations and velocities, one can determine a statistical distribution of the M/L values, which can then be compared with the mean separations and masses obtained from the rotation curves of the galaxies involved. Again, there is strong evidence for a large amount of unseen matter, often reaching M/L values greater than 100. The method has also been extended to small galactic groups, with similar results.

Using large samples of binary galaxies, so that the effects of orbital eccentricity, inclination, and mass transfer can be accounted for, recent studies have shown that about a factor of a three times the observed mass is required to explain binary galaxy dynamics. Summarized, the estimates of M/L range from about 10 to 50. The wide range reflects uncertainties in both the Hubble constant, which is required for the specification of the separation in physical units, and the projection of the velocities of the galaxies in the line of sight. Unlike clusters, binary galaxies are rather

sensitive to the projection factors, since one is dealing only with two objects at a time.

Measurement of the masses of the component galaxies can also add information to this picture, and there have been efforts to obtain the rotation curves for the galaxies in binaries and small groups. The interpretation of the velocity differences, the quantity which, along with the projected separation, gives the mass of the system, requires some idea of the potential in which the galaxies move. If there is very extended DM around each of the component galaxies, the use of point-mass or simple spheroidal models will fail to give meaningful masses. Also, since some of the binary galaxies may be barred spirals and others may show subtle and not yet observed tidal effects, it is likely that the mass estimates from these systems should be taken with considerable caution.

Small groups of galaxies, the so-called Hickson groups, can be used to provide an example of dynamical interaction which lies between the clusters and the binaries. Here, however, the picture is also complex. Interactions between the constituent galaxies can alter their mass distributions and affect their orbits in a manner that vitiates the determination of the orbital properties for the member galaxies. If these groups are bound and stable, the M/L must be of order 100. However, Rubin provided evidence, from the study of individual galaxies, that the rotation curves do not show the flat behavior of isolated galaxies. It is possible that the halos of the galaxies have been altered in the groups. Deep photographs of the groups provide some handle on this. Complicated mass distributions of the luminous matter are observed, indicating that the galaxies have been strongly interacting with one another. Mergers, collisions, and tidal interactions at a distance are all indicated in the images. Clearly, more detailed study is required of these objects before they can be used as demonstrations of the need for DM.

VI. THE LOCAL SUPERCLUSTER

A. Virgo and the Virgocentric Flow

The Local Group sits on the periphery of the Virgo supercluster, containing several hundred bright galaxies and the Virgo cluster centered on the giant elliptical galaxy M 87, which contains as much as $10^{14} M_{\odot}$. Because of its mass and proximity, its composition and dynamics have been well studied. Virial analyses indicate that the total cluster is bound by DM, which amounts to about 10 times the observed mass (a figure that is typical of clusters and which will be discussed in more detail later). Because of its closeness, however, the Virgo supercluster has an even more dramatic effect on motion of the Local Group, which can be used to independently estimate its total mass.

From the motion of the Local Group with respect to the Virgo cluster, it is possible to estimate the total mass of the cluster. The argument is a bit like that of Newton's apple analogy. In the free expansion of the universe, local regions may be massive enough to retard the flow of the galaxies away from each other. This is responsible for the stability of groups of galaxies in the first place. They are locally self-gravitating. In the case of the motion of the Galaxy toward Virgo, there are several ways of measuring this velocity.

One of the most elegant is the dipole anisotropy of the cosmic background radiation (CBR). The idea of using the CBR as a fixed light source relative to which the motion of the observer can be obtained is sometimes called the "new aether drift." The motion of the observer produces a variation in the effective temperature of the background that is directly proportional to the redshift of the observer relative to the local Hubble flow. For motions of the order of 300 km $\sec^{-1}$, for example, the variation in the temperature of the CBR should be about 1 mK, with a dipole (that is, consine) dependence on angle. The CBR shows a dipole temperature distribution, about one-third of which can be explained by the motion of the Local Group toward Virgo, about 250–300 km $\sec^{-1}$, called the *virgocentric velocity*. Since the Local Group is part of the supercluster system that contains Virgo, it is not surprising that we are at least partially bound to the cluster. The total mass implied by this motion is consistent with the virial mass determined for the cluster, which implies a considerable amount of dark matter (about 10 times the observable mass). The dipole anisotropy, however, is only partially, explained by the virgocentric flow. There must be some other larger scale motion, about a factor of two larger, responsible for most of the variation in the CBR. This seems to result from the bulk motion of the local supercluster toward the Hydra-Centaurus supercluster ($l_{\rm II} \approx 270^{\circ}$, $b_{\rm II} \approx 30^{\circ}$). This brings up the question of large-scale deviations, on the scale of superclusters, of the expansion from the Hubble law.

B. The Large-Scale Streaming in the Local Galaxies: Dark Matter and the Hubble Flow

Distance determination can be a problem for galaxies outside of the Local Group and Virgo cluster. This has to do with our inability to resolve individual stars and H II regions. Several methods have evolved that allow for determination of the intrinsic brightness of galaxies independent of their Hubble flow velocities. The *Tully–Fisher relation* correlates the absolute magnitude of a galaxy, either in the blue or in the infrared, with the width of the 21-cm line. Although this depends on the inclination of the galaxy, this can be taken into account from optical

imaging, and it provides a calibration for the intrinsic luminosity of the galaxy from which, using the apparent magnitude, the distance can be obtained directly, without cosmological assumptions. Another calibrator, especially useful for elliptical and gas-poor galaxies, is the *Faber–Jackson relation*. This uses the velocity dispersion observed for the core of elliptical galaxies and spheroids to determine the intrinsic luminosity of the parent galaxy. It is representative of a wide class of objects, indicative of dissipative formation of the systems, and is

$$L \sim \sigma^n, \tag{11}$$

where σ is the velocity dispersion of the nucleus and $n \approx 4$ from most studies.

Using these methods, it is possible to obtain the distance to a galaxy independent of the Hubble velocity; one can then look for systematic deviations from the isotropy of the expansion of the galaxies in the vicinity of the Local Group. Observations show that there is a large-scale deviation of galaxy motion in the vicinity of the Galaxy. Deviations at large displacement to the virgocentric velocity of order 600–1000 km sec^{-1} show that there is a large departure from the Hubble law relative to the uniform expansion that is usually assumed. The characteristic scale associated with this deviation is about 50–100 Mpc, much larger than the size usually associated with clusters of galaxies but on a scale of superclusters. No mass can clearly be identified with this gravitationally perturbing concentration, but it has been argued that it may be a group of galaxies of order $10^{14} M_\odot$, about the size associated with a very large cluster or small supercluster of galaxies. The mystery is its low luminosity, but it is located near the galactic plane, which would account for the difficulty in observing its constituent galaxies. This may be indicative of other large-scale deviations on a larger scale of gigaparsecs.

VII. CLUSTERS OF GALAXIES

From the first recognition of galaxies as stellar systems like the Milky Way, it has been clear that their distribution is markedly inhomogeneous. Early studies by Shapley indicated large (factor of two) density fluctuations in their distribution, while later work has expanded the complexity of this distribution to larger density contrasts. A variety of large-scale structures is observed in the galactic clustering, which gives rise to the idea of a hierarchy. First, there are clusters, megaparsec scale concentrations of luminous objects with total masses typically of $10^{13}-10^{14} M_\odot$. These contrast with voids, the most famous of which is the Bootes Void, in which the density of galaxies is about 1/10 that of the averages. Clusters of galaxies cluster themselves into structures called superclusters, and it appears that even these may have some clustering characteristics, although they are sufficiently rare that the statistical information is shaky at best. Clusters can be dynamically and morphologically separated from the background of galaxies and these can be studied more or less in isolation from one another.

Several catalogs are available for galaxy clusters, generally identified with the names Abell and Zwicky. In addition, large-scale sky counts of galaxy frequency per square degree have been made by Shane and Wirtanen (the Lick survey), Zwicky, and the Jagellonian surveys. These do not contain any dynamical information; they are number counts only. But from information about the spatial correlation on the two-dimensional surface of the sky, the three-dimensional properties of galaxies can be crudely determined.

This situation greatly improved with the Center for Astrophysics redshift survey, a complete radial velocity study for galaxies within 15,000 km sec^{-1} of the Local Group (covering scales $< 0.1c$). With velocity information available, it is possible to place the galaxies in question in three-dimensional space and to look at the detailed distribution of both the galaxies and clusters on the scale of 100 Mpc. This has given the first evidence for the nature of the large-scale structure of the universe and the need to consider dark matter in the context of both clusters and superclusters of galaxies. First, we examine the basis for the determination of mass within the clusters and the arguments on the reality of clustering. Then, we discuss the ways in which this information can be used to address the distribution of, and need for, the DM on scales comparable with the size of the visible universe.

The problem of cold versus hot dark matter has been addressed by a number of techniques. Large-scale structures have been detailed in the neighborhood of the Local Group. Two big structures, dubbed the Great Attractor (Lynden-Bell *et al.*, 1988) and the Great Wall (Geller and Huchra, 1990), have been discovered on size scales of order 100 Mpc. The presence of a large-scale deviation from the velocities of the Hubble flow was the signature of the Attactor, a comparatively local region possibly associated with a supercluster.

The Harvard Center for Astrophysics redshift survey of about 15,000 galaxies in the northern hemisphere has revealed several superstructures and, in general, a distinctly inhomogeneous distribution for local luminous matter. Its implication for dark matter is that, at least on the scale of clusters and individual galaxies, there is considerable power in the small-scale length for which CDM is likely responsible. A statistically more distant sample has, however, been analyzed using the IRAS database. These infrared-detected galaxies provide a sample with

very different selection criteria than the optically based, and optically based, surveys previously available. The procedure for analyzing such datasets is to follow up on the IR detections with random samples of redshifts of a comparatively large number of galaxies (of order several thousand) from which distances are obtained. The results, reported by Saunders *et al*. (1991), find a significant excess in the galaxy correlation function on scales in excess of 20 Mpc, a size large in comparison with that expected for ordinary CDM. Precisely how this will stand up in extended surveys based on deeper cuts of the existing catalogs, and how this compares with the source counts from COBE and other IR missions, remains to be seen.

A. The Virial Theorem

One of the first suggestions that there is a compelling need to include unseen matter in the universe came from the original determination of velocity dispersions in clusters of galaxies. The argument is quite similar to that used for a gas confined in a gravitational field. Galaxies in a cluster may collide, a problem to which we will later return, but for the moment we will assume that the galaxies are in independent orbits about the center of mass of the clusters. The velocity dispersion is the result of the distribution of orbits of the galaxies about the cluster center, and the radial extent of the galaxies in space results from the fact that they are bound to the cluster. A dynamical system in hydrostatic equilibrium obeys the virial theorem, which states that

$$2T + W = 0, \tag{12}$$

where T is the kinetic energy and W is the gravitational potential energy. The total energy of the system must be *at most zero*, since the system is assumed to be bound. Therefore, the total energy is given by $E = T + W = \frac{1}{2}W < 0$, and the velocity dispersion is given approximately by

$$\sigma^2 = \frac{GM_{\text{virial}}}{\langle R \rangle}. \tag{13}$$

The mass determined from the statistical distribution of orbits depends on the size of the system, and this requires some estimate of the density profile of the cluster. The rms value of the group radius usually suffices for the component of the radial velocity. It is assumed that the orbits are randomly distributed and that they have not been seriously altered since the formation of the cluster. Also, an assumption is embedded in this equation for the dispersion: the mass distribution has remained unaltered on a time scale comparable with the orbital time scale for the galaxies about the center of mass.

The argument that clusters must be bound and stable comes from the distribution of galaxy velocities within the clusters. The dynamical time scale, also called the crossing time, is the period of a galaxy orbit through the center of the cluster potential. The characteristic time scales are of order 10^9 years, much shorter than the Hubble time (the expansion time scale for the universe, taken as the inverse of the Hubble constant).

There have been numerous simulations of the evolution of galaxies in clusters, most of which show evidence for encounters of the cluster members. There is also strong evidence for interaction in the numerous disturbed galaxies in these clusters and the presence of cD galaxies, giant ellipticals with extended halos, in the cluster centers. The latter are assumed to grow by some form of accretion, also called *cannibalism*, of the neighboring galactic systems.

B. Interactions in Groups of Galaxies

Galaxies in clusters collide. This simple fact is responsible for many of the problems in interpreting the virial theorem because it assumes that the dynamical system is intrinsically dissipationless. Collisions redistribute angular momentum and mass and perturb orbits in stochastic and time-dependent ways. None of these effects can be included in the virial theorem formulation, but instead require a full Fokker–Planck treatment of the dynamics.

Observations of small groups of galaxies, the Hickson groups, reveal a complex array of interactions—shells, bridges, and tails are found for many of the members of these small groups. The extent to which this alters the mass distribution of the groups is not known nor is it known how this affects the determination of the M/L ratio for the group.

For clusters of galaxies, the situation is even more complicated. Many cD galaxies have been found to be embedded in shell systems, which are believed to arise from accretion of low-mass disk galaxies by the giant ellipticals. The shells have been used to argue for the mass distribution of the cDs, although the detailed mass distribution cannot be determined from the use of the shells. Instead, they serve as a warning about using simple collisionless arguments for the evolution of the cluster galaxy distribution.

C. The Distribution of Clusters and Superclustering

The two-point correlation function, $\xi(r)$, is the measure of the probability that there will be two galaxies within a fixed distance r of each other. The probability of finding a galaxy in volume element dV_1 within a distance r of another in volume dV_2 is given by

$$dP(1, 2) = n^2[1 + \xi(r_{1,2})]\, dV_1\, dV_2, \tag{14}$$

where n is the mean number density of galaxies. Defined in this way, it is a good measurement of the trend of galaxies to cluster. If there is a clustering, the correlation is asymptotically vanishing at large distance, very positive at small distance, and negative for some interval. It is only a requirement that the cumulant be positive definite (normalizable). The two-point function is not, however, the sole discriminant of clustering. The problem with the equations of motion is that, in an expanding background and under the influence of gravity, the mass points evolve in a highly nonlinear way. The equations of motion do not close at any order, and there is a hierarchy to the distribution of the correlations of different orders.

The correlation function seems to have a nearly universal dependence on separation, $\xi(r) \sim r^{-1.8}$, whether for galaxy–galaxy or cluster–cluster correlations. The coefficient is different, which indicates that the hierarchy is not quite exact and that there may be further differences lurking on the scale of superclusters. It is not likely, however, that this will be easily studied, considering the size of the samples and the extreme difficulty in determining the extent to which these largest scale structures can be separated from the clusters.

D. The Cosmic Microwave Background

The November 1989 launch of the Cosmic Background Explorer, COBE, revolutionized the study of large-scale structure. Designed to operate in the far-infrared and millimeter portions of the spectrum, COBE performed an all-sky survey of the diffuse cosmic background radiation (CBR).

The CBR is truly Planckian; that is, it is a uniform-temperature blackbody spectrum. The temperature is 2.735 K, with an error of less than 60 mK for the wavelength region between 400 μm and 1 cm. The dipole anisotropy, caused by motion of the local standard of rest toward the Virgo cluster, is easily detected in the data with an amplitude of 3.3 mK (with less than a 10% error) in the direction $\alpha = 11^h.1$ and $\delta = -6^o.3$. The lack of a strong quadrupole signature in the background is also very important. This places limits on the distribution of large mass concentrations at large redshift and also on strong mass concentrations at distances comparable to the Virgo cluster (about 10–100 Mpc).

An important limit provided by the shape of the spectrum is the variation of temperature over the sky due to scattering of the CBR by hot gas. Called the Sunyaev–Zeldovich effect, relativistic or near-relativistic electrons with temperature T_e. scatter the background photons, boosting their energies to $\nu \approx k_B T_e / h$, where h is Planck's constant and k_B is the Boltzmann constant. This scattering removes the photons from the long-wavelength part of the CBR, resulting in a change in the intensity with position. The limit provided by the COBE results is that y, the dimensionless Sunyaev–Zeldovich parameter, is $\leq 10^{-3}$, ruling out a substantial contribution by a hot intracluster medium to the temperature of the CBR.

The major result from COBE was the detection of large-scale ($>10^\circ$) temperature fluctuations, $\Delta T/T \approx (29 \pm 1) - (35 \pm 2)\,\mu$K (the smaller amplitude corresponds to the larger angular scales). These are the signature of the fluctuations in the matter distribution imposed during inflation that cross the horizon at the decoupling epoch and produce local perturbations in the gravitational field. Even without the further constraint of galaxy and cluster formation, these temperature variations require a large dark matter component. First, they are very low amplitude. There is little time for them to grow large enough to drive clustering on scales of tens of megaparsecs unless there is something massive but unseen. Second, they do make it out of the earliest stages of the expansion without damping. And most significantly, they are very large scale.

VIII. COSMOLOGICAL CONSTRAINTS AND EXPLANATIONS OF THE DARK MATTER

Several candidates have been suggested for the particular constituents of DM. Since this is a field of enormous variety and unusual richness of ideas, many of which change on short life times, only a generic summary will be provided. This should be sufficient to direct the interested reader to the literature.

A. Cosmological Preliminaries

In attempting to find a likely candidate for the DM, one must first look at some of the constraints placed on models by the cosmological expansion. First, the particles must survive the process of annihilation and creation that dominates their statistical population fluctuations during the early phases of the expansion. The cosmology is given by the Friedmann–Robertson–Walker metric for a homogenous, isotropically expanding universe with radial coordinate r and angular measures θ and ϕ:

$$ds^2 = dt^2 - \frac{R^2(t)\,dr^2}{1 - kr^2} - R^2(t)(d\theta^2 + \sin^2\theta\, d\phi^2), \quad (15)$$

where $R(t)$ is the scale factor, the rate of the expansion being given by

$$\frac{d}{dt}(\rho V) + p\,\frac{dV}{dt} = 0, \quad (16)$$

which is the entropy equation, with the volume $V \sim R^3$ and ρ and p being the energy density and pressure, respectively, and

$$H^2(t) = \left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{R^2}, \qquad (17)$$

which is essentially the Hubble law. The value of k is critical here to the argument, since it is the factor that determines the curvature of the universe. The critical solution, with $\Omega = 1$, has a deceleration parameter $q_0 = 0$. Here, $\Omega = \rho/\rho_c$, where ρ_c is the density needed to give a flat space–time, $\rho_c = 3H_0^2/8\pi G \approx 2 \times 10^{-29} h^2$ g cm^{-3}, where H_0 is the present value of the Hubble constant and $h = H_0/(100$ km sec^{-1} Mpc$^{-1})$. This solution is favored by inflationary cosmologies. During the radiation-dominated era, the energy density varies like R^{-4} and the pressure is given by $\frac{1}{3}\rho$, so that during the earliest epoch, $\rho \sim t^{-2.} \sim T^{-2}$.

The horizon is given by

$$l_H = \int_{t_0}^{t} \frac{dt'}{R(t')}. \qquad (18)$$

Perturbations in the expanding universe have a chance of becoming important when they grow to the scale of the horizon. Others will damp because they cannot be causally connected and will simply locally evolve and die away. The expansion then produces fluctuations in the background radiation on the scale of the horizon at the epoch at which regions begin to coalesce.

The entropy of the expanding universe is constant, so that the number density of particles at the time of their formation is related to the temperature of the background, $n_i T^{-3} =$ const, from which we can define the ratio of the photon to baryon number density as a measure of the entropy, $S = n_\gamma/n_b = 10^9$, where $n_\gamma \sim T^3$. For particle creation, the particles will freeze out of equilibrium when the energy in the background is small compared with their rest mass energy, or when $T \leq m_i c^2/k_B$, where k_B is the Boltzmann constant. Thus, for particles that are very massive, during the radiation-dominated phase, the temperature can be very high indeed. Put another way, the equilibrium temperature is defined as $T_{eq} = 10^{11} m_i$(GeV) K. Since the temperature during the radiation-dominated era varies like R^{-1}, this implies that $R/R_0 \sim 10^{-11} m_i$, where R_0 is the present radius. The point is that the particles which may constitute nonbaryonic explanations for the DM are created in a very early phase of the universe, one which presented a very different physical environment than anything we currently know directly. The cosmic background radiation has a temperature of 2.735 K at the present epoch.

B. The Cosmological Constant

The cosmological constant, Λ, was first introduced by Einstein in an attempt to maintain Mach's principle that local inertia should be determined by the distribution of matter on the largest scale. Although ignored whenever possible, the constant is easily included in the field equations through an added term $\Lambda/3$ on the right-hand side of Eq. (17). This behaves like a negative pressure, driving the expansion, and is presumed to arise from vacuum energy. Observations had, until recently, been consistent with a vanishingly small Λ. But recent data for type Ia supernovae strongly point to a value of $\Omega_\Lambda \approx 0.7$ (although still with $\Omega = \Omega_{matter} + \Omega_\Lambda = 1$), which in light of the nucleosynthesis limits would make this term the dominant contributor to the energy density. The data show an additional acceleration is required to explain why supernovae of this type, which have been shown to be standard candles with intrinsically small luminosity dispersions, are fainter and more distant for $z > 0.5$ than would be predicted from standard cosmological models with $\Omega = 1$.

C. Global Constraints on the Presence and Nature of Dark Matter

1. Nucleosynthesis

There are few direct lines of evidence for the total mass density of the universe. One of the most direct comes from the analysis of the abundance of the light elements. Specifically, in the early moments of the Big Bang, within the first few minutes, the isotopes of hydrogen and helium froze out. While ^{4}He can be generated in stars by nuclear reactions, deuterium cannot. In the Big Bang, however, it can be produced by several different reactions,

$$\mathrm{p} + \mathrm{n} \rightarrow {}^2\mathrm{D} + \gamma, \qquad \mathrm{p} + \mathrm{p} + \mathrm{e} \rightarrow {}^2\mathrm{D} + \gamma,$$
$$^3\mathrm{He} + \gamma \rightarrow {}^2\mathrm{D} + \mathrm{n},$$

and the deuterium can be destroyed very efficiently by

$$^2\mathrm{D} + \mathrm{n} \rightarrow {}^3\mathrm{H}, \qquad {}^2\mathrm{D} + \mathrm{p} \rightarrow {}^3\mathrm{He} + \gamma.$$

The critical feature of all of these reactions is that they depend on the rate of expansion of the universe during the nucleosynthetic epoch. The drop in both the density and temperature serves to throttle the reactions, producing several useful indicators of the density.

The argument continues: if the rate of expansion is very rapid, that is, if the density is much lower than closure, then the deuterium can be rapidly produced but not effectively destroyed because of the drop in the reaction rate for the subsequent parts of the nuclear network. However, if the density is high, the rate of expansion will be slow enough for the chain of reactions to go to equilibrium abundances

of ^{4}He because of the consumption of the ^{2}D. Thus, the D/H ratio, the mass fraction of ^{4}He, and the ^{3}He/^{4}He ratio can be used to constrain the value of $\Omega = \rho/\rho_c$, where ρ_c is the critical density required for a flat universe ($2 \times 10^{-29} h^2$ g cm^{-3}). The larger is Ω, the smaller is the D/H ratio. Current observations of the primordial abundances of helium isotopes and of deuterium provide $\Omega_{\text{matter}} \approx 0.1-0.3$.

This appears to provide a number significantly lower than that obtained from the virial masses for clusters and from clustering arguments. Since the baryons in the early universe determined the abundances of the elements that emerged from the Big Bang, the abundances of the light isotopes provide a strong constraint on the fraction of DM that can be in baryonic form, luminous or not.

2. Isotropy of the CBR

Recent balloon observations, especially the BOOMERANG experiment, have probed the intermediate angular scales of the CBR fluctuations. These data show a strong peaking on a scale of about 1°, which corresponds to an angular wave number of $l \approx 200$. This is presumably revealing the acoustic (pressure) fluctuations that were generated during the inflationary epoch and survived by stretching through the expansion to $z \approx 1000$ and indicate the size of the horizon. For angular scales smaller than about 7 arcmin $\times\, \Omega^{1/2}$, the perturbations are damped because they are lower than the Silk mass, see below, Eq. (23). The spectrum is best reproduced by an open ΛCDM model (CDM with a cosmological constant). Further improvements in the angular correlation tests will be achieved after MAP and PLANCK are launched in this decade.

3. The Formation of Galaxies and Clusters of Galaxies

Recent work on the distribution of galaxies has centered on the idea that the visible matter does not represent the distribution of mass overall. The picture that is developing along these lines, called "biased galaxy formation," takes as its starting point the assumption that galaxies are unusual phenomena in the mass spectrum. One usually assumes that the galaxies are the result of perturbations in the expanding universe at some epoch during the radiation-dominated period and therefore are representative of the matter distribution. However, biasing argues that the galaxies are the product of unusually large density fluctuations and that the subsequent development of these perturbations is dominated by the smoother background distribution of dark matter—that which never coalesces into galactic mass objects. Several mechanisms have been suggested, all of which may be reasonable in some part of the evolution of the early universe. One picture uses explosions, or bubbles, formed from the first generation of stars and protogalaxies to redistribute mass, alter the structure of the microwave background, and erase the initial perturbation spectrum.

Gravitational clustering alone is insufficient to explain the dramatic density contrasts observed between clusters and voids, unless one invokes large perturbations at the epoch of decoupling of matter and radiation. The current density contrast is of order 2–10, which implies that at the recombination era the perturbations in the density must be of order 10^{-3}. Either the fluctuations in the density at this period are isothermal, in which case the density variation is not reflected in the temperature fluctuations in the cosmic background radiation, or there must be some other mechanism for the formation of the galaxies. In an adiabatic variation, $\delta T/T = (1/3) \times \delta\rho/\rho$, which is clearly ruled out by observations on scales from arcseconds to degrees to a level of 10^{-5}. If the density fluctuations are smaller than this value, there is not enough time for them to grow to the sizes implied by the large-scale structure by the present epoch. Further, quasars also appear to show clustering and to be members of clusters, at $z = 1-4$, so this pushes the growth of nonlinear perturbations to even earlier periods in the history of the universe. If the density of matter is really the cosmological critical value, that is, if $\Omega = 1$, then there cannot be a simple explanation of the distribution of luminous matter simply by the effects of primordial fluctuations.

Biasing mechanisms can be combined with the DM in assuming that the massive particles are the ones which show the large-scale perturbations. Here, the difference between hot and cold DM scenarios shows up most clearly. If the matter is formed hot and stays hot, it will damp out all of the small-scale perturbations. If formed cold, these will be the ones that will grow most rapidly, with gravitational effects later sorting the smallest fragments of the Big Bang into the larger scale hierarchical structures of clusters and superclusters. The topologies of the resultant universes differ between these two extremes, with the cold matter showing the larger contrast between voids and clusters and an apperance that is best characterized by filaments rather than lumps. It should be added that biasing can act on both scenarios and that the final appearance of the simulations can depend on the statistical method chosen to determine the biasing.

4. Cold versus Hot Dark Matter: New Results

The choice between cold versus hot dark matter scenarios depends on the smoothness of the mass distribution but particularly on the concentration of galaxies within the large-scale filaments that link clusters. Too hot a DM

component and all the substructure washes out, too cold and it grows too quickly. As with Goldilocks, there is always a third option—warm DM—but the choice of particle is less clear. The distinction between the different scenarios rests on whether the particles are relativistic and/or massless, or thermal and/or baryonic, in which case the former support the largest physical perturbations but prevent the coalescence of substructure, while the latter exacerbate its growth.

D. The Particles from the Beginning of Time

The initial scale at which all of the particle theories start is with the Planck mass, the scale at which quantum perturbations are on the same scale for gravity as the event horizon. This gives

$$m_{\rm Planck} = \left(\frac{hc}{2\pi G}\right)^{1/2} \approx 1.7 \times 10^{19}\ {\rm GeV}. \qquad (19)$$

This is the scale at which the ultimate unification of forces is achieved, independent of the particle theories. It depends only on the gravitational coupling constant. It is then a question of where the next *grand unification* scale occurs (the GUT transition). This is usually placed at $10^{12}-10^{15}$ GeV, considerably later and cooler for the background. The particles of which the background is assumed to be composed are usually assumed to arise subsequent to this stage in the expansion, and will therefore have masses well above those of the baryons, in general, but considerably below that of the Planck scale.

1. Inflationary Universe and Soon After

The most recent interest by theorists in the need for DM comes from the *inflationary universe* model. In this picture, originated by Guth and Linde, the universal expansion begins in a vacuum state that undergoes a rapid expansion leading to an initially flat space–time, which then suffers a phase transition at the time when the temperature falls below the Planck mass. At this point, the universe reinflates, now containing matter and photons, and it is this state which characterizes the current universe. One of the compelling questions that this scenario addresses is why the universe is so nearly flat. That is, the universe, were it significantly different from a critical solution at the initial instant, would depart by large factors from this state by the present epoch. The same is true for the isotropy. The universe at the time of appearance of matter must have been completely isotropized, something which cannot be understood within the framework of noninflationary pictures of the Big Bang.

The fact that, in these models, $\Omega = 1$ is a requirement, while all other determinations of the density parameter yield far smaller values (open universes), fuels the search for some form of nonbaryonic DM. The ratio $\Omega/\Omega_{\rm observed}$ is approximately 10–100, of the same order as that required from clusters of galaxies.

During the expansion, particles can be created and annihilated through interactions; thus,

$$\frac{dn}{dt} = -D(T)n^2 - 3\frac{\dot{R}}{R}n + C(t), \qquad (20)$$

where the first term is the annihilation, which is temperature dependent; the second is the dilution of the number density because of the expansion; and the third is the creation term, which depends on time and (implicitly) temperature. Let us examine what happens if the rate of creation depends on a particle whose rate of creation is in equilibrium with its destruction. Then, if a two-body interaction is responsible for the creation of new particles and they have a mass m, the Saha equation provides the number in equilibrium,

$$n_{\rm eq} \sim (mT)^{3/2}\exp(-m/T), \qquad (21)$$

and the rate of particle production becomes a function of the particle mass. Thus, the more massive particles can be destroyed effectively in the early universe but will survive subsequently because of the rapid expansion of the background and dilution of the number density. Any DM candidate will therefore have a sensitive dependence on its mass for the epoch at which the separation from radiation occurs as well as the strength of the particle's interaction with matter and radiation.

Following their freeze-out phase, the particles will be coupled to the expanding radiation and matter until their path length becomes comparable with the horizon scale. This phase, called *decoupling*, also depends on the strength of the interaction with the radiation and matter. After this epoch, the particles can freely move in the universe, which is now "optically thin" to them. It is this free-streaming that determines the scale over which the particles can allow for perturbations to grow. For instance, in the case of the massive cold particles, those created with very low velocity dispersions, the scale over which they can freely move without interacting gravitationally with the baryonic matter is determined by the mass of the baryon clumps. For cold DM, this means that the slow particles, those with energies of less than 10 eV, can be trapped within galactic potentials. Those with energies of 100 eV would still effectively be trapped by clusters. If the particles are hot, they cannot be trapped by these clumps and can, in fact, erase the mass perturbations that would lead to the formation of these smaller structures.

As mentioned in discussing galaxy formation in general, simulations of the DM show that for the hot particles, the mass scale that survives the primordial

perturbations and damping is the size of superclusters, about $10^{15}-10^{16}M_{\odot}$. The superclusters thus form first, then these fragment and separate from the background and form clusters and their constituent galaxies from the "top down." If the particles are cold, they cannot erase the smallest scales of the perturbations, on the order of galaxy size ($10^{12}-10^{13}M_{\odot}$), and clusters and superclusters form gravitationally from the "bottom up." The problem is that there is little observational evidence to distinguish which of these is the most likely explanation. In effect, it is the simulations of the galactic interactions, placed in an expanding cosmology and with very restrictive assumptions often being made about the interactions between galaxies within the clusters thus formed, which are used as argumentation for one or the other scenario. In fact, it is for this reason that the models must be called *scenarios* since there cannot be detailed analytic solution of the complex problem and the solutions are only sufficient within the limitations of the precise assumptions that have been applied in the calculations.

The critical mass for gravitational instability and galaxy formation is the *Jeans mass*, the length (and therefore mass) scale at which material perturbations become self-gravitating and separate out in the expanding background. For species in equilibrium with radiation, it depends on both the number density of the dominant particle and its mass as well as on the temperature of the background radiation. In the case of the expanding universe, since during the radiation-dominated era the mass density and temperature are intimately and inseparably linked, the Jeans mass depends only on the mass of the particle that is assumed to be the dominant matter species:

$$M_{\rm J} \sim \rho_i \left(\frac{a_{\rm s}^2}{G\rho_i}\right)^{3/2} = \left(\frac{T^3}{Gm_i^2 n_i}\right)^{1/2}$$
$$\approx 3\times 10^{-9} m_{i,\rm GeV}^{-1} M_{\odot}, \qquad (22)$$

where a_s is the sound speed and n_i is the number density of the dominant species. The smaller the mass of the particle, the larger the initially unstable mass. This collapse must compete against the fact that viscosity from the interaction with background photons tends to damp out the fluctuations in the expanding medium. The critical mass for stability against dissipation within the radiation background is given by the *Silk mass*, below which the photons, by scattering and the effective viscosity that is represents, damp perturbations. This gives a mass of

$$M_{\rm s} = 3\times 10^{13}(\Omega h^2)^{-5/2} M_{\odot}. \qquad (23)$$

Below this mass, which is of the same order as a galactic mass, perturbations are damped out during the early stages of the expansion. This provides a minimum scale on which galaxies can be conceived to be formed. Other particles, however, can also serve to damp out the perturbations at an even earlier epoch if they are hot enough.

2. The Particle Zoo

a. Baryons. Baryons, the protons and neutrons, are the basic building blocks of ordinary matter. They interact via the strong and electromagnetic forces and constitute the material out of which all luminous material is composed.

The basic form of baryonic matter, hot or cold gas, can be ruled out on several counts. One has been discussed in the section on nucleosynthesis in the Big Bang; the baryonic density is constrained by isotopic ratios to be small. Another constraint is provided by the absorption lines formed from cold neutral gas through the intergalactic medium in lines of sight to distant quasars. There are not sufficiently high column densities observed along any of the lines of sight to explain the missing mass. The absorption from Lyman α, the ground-state transition of neutral hydrogen, is sufficiently strong and well enough observed that were the gas in neutral form, it would certainly show this transition viewed against distant light sources. While such narrow absorption lines are seen in QSO spectra, they are not sufficiently strong to account alone for the dark matter.

In addition, hot gas is observed in clusters of galaxies in the form of diffuse X-ray halos surrounding the galaxies and cospatial with the extent of their distribution. Here again, the densities required to explain the observations are not sufficient to bind the clusters or to account for the larger scale missing mass since the X-ray emission seems to be confined only to the clusters and not to pervade the universe at large scale.

It is possible that the matter could be in the form of black holes formed from ordinary material in a collapsed state, but the required number, along with the difficulties associated with their formation, makes this alternative tantalizing but not very useful at present. Of course, football- to planet-sized objects could be present in very large numbers, forming the lowest mass end of the mass spectrum that characterizes stars. Aside from the usual problem of not being able theoretically to form these objects in the required numbers, there are also the constraints of the isotropy of the microwave background and of the upper limit to the contribution of such masses to the infrared background. Since these objects would behave like ordinary matter, they should come into thermal equilibrium with the background radiation and hence be observable as fluctuations in the background radiation on the scale of clusters of galaxies and smaller. Such a variation in the temperature of the CBR is not observed. In addition, it would have been necessary for the baryons to have

participated in the nucleosynthetic events of the first few minutes, and this also constrains the rate of expansion to rule out these as the primary components of the universe. If inflation did occur and the universe is really an $\Omega = 1$ world, then the critical missing mass cannot be made up by baryons without doing significant damage to the understanding of the nuclear processing of the baryonic matter.

b. Neutrinos. The electron neutrino (ν_e) is the lightest lepton, and in fact the lightest fermion, making it an interesting particle for explaining DM. It cannot decay into lower mass species if it is only available in one helicity, and thus would survive from the epoch at which it decoupled from matter and radiation during the radiation-dominated era of the expansion.

It is well known that neutrino processes, being moderated by the weak force, permit the particles to escape detection by many of the classical tests. That these are weak particles means that they decouple from the expansion sooner than the photons and can freely stream on scales larger than those of the photons, also remaining hotter than the photon gas and therefore more extended in their distribution if they accrete onto galaxies or in clusters of galaxies. Should the neutrino have a large enough mass, about 30 eV, the predicted abundance of the three known species is sufficient to close the universe and possibly account for the gross features of the missing mass. Limits place this maximum mass for the electron neutrino species as $\leq$10 eV. Other problems with this explanation include the rates of nucleosynthesis in the early universe, particularly the neutron to proton ratio, which is fixed by the abundance of light leptons and the number of neutrino flavors.

The argument that the ν_e mass must be nonvanishing is important. If the neutrino is massive, it may be nonrelativistic. This implies that it decoupled from the microwave background at some time before the baryons, but later than the massless particles. Since the particle decoupling depends on the temperature of the radiation, this determines the mean free-streaming velocity. The capture efficiency of baryons to this background of particles is dependent on their mass and velocity. The escape velocity from a galaxy is several hundred km sec^{-1}, while for galaxy clusters it is much higher. Thus, for galactic halo DM to be explained by ν_e, it is necessary that the particles be cold, that is, they must have free-streaming velocities less than 100 km sec^{-1}. The restriction of the particle mass to less than 10 eV by the observations of the width of the burst of neutrinos from Supernova 1987 A in the Large Magellanic Cloud likely adds fuel to the argument that the neutrino is an unlikely candidate for DM.

If ν_e is the dominant component of DM, an analog of the Silk mass is possible, below which fluctuations are damped because of the weak interaction of matter with the relic neutrinos. The density parameter is now given by $\Omega = 0.01h^{-2}m_\nu$ (eV) and

$$M_\nu = 5 \times 10^{14}\left(\Omega_\nu h^2\right)^{-2} M_\odot, \tag{24}$$

which is closer to the scale of clusters than galaxies. With current limits, it looks as if ν_e can neither provide $\Omega_\nu = 1$ nor seriously alter the mass scale for galaxy formation, but may nonetheless be important in nucleosynthesis.

c. Alternative particles. After exhausting classical explanations of DM, one must appeal to far more exotic candidates. It is this feature of cosmology that has become so important for particle theorists. Denied laboratory access to the states of matter that were present in the early stages of the Big Bang, they have attempted to use the universe as a probe of the conditions of energy and length that dominate the smallest scales of particle interactions at a fundamental level. Above the scale of energy represented by proton decay experiments, the only probable site for the study of these extreme states of matter is in the cosmos.

Many grand unified theories, or GUTs, contain broken symmetries that must be mediated by the addition of new fields. In order to carry these interactions, particles have to be added to the known spectrum from electroweak and quantum chromodynamics (QCD) theory. One such particle, the axion, is a moderator of the charge-parity (CP) violation observed for hadrons. It is a light particle, but not massless, which is weakly interacting with matter and should therefore decouple early in its history from the background radiation. Being of light mass and stable, it will survive to the epoch of galaxy formation and may provide the particles needed for the formation of the gravitational potentials that collect matter to form the galaxy-scale perturbations.

It is perhaps easiest to state that the axion has the attractive feature that its properties can largely be inferred from models of DM behavior rather than the other way around, and that the formation of these particles, or something much like them, is a requirement in most GUTs. Simulations of galaxy clustering make use of these and other particles with very generalized properties in order to limit the classes of possible explanations for the DM of galaxies and clusters of galaxies.

Supersymmetry, often called SUSY, is the ultimate exploitation of particle symmetries—the unification of fermions and bosons. Since there appears to be a characteristic scale of mass at which this can be effected and since supersymmetry assigns to each particle of one type a partner of the opposite type but of different mass, it is possible that the different particles decouple from the primordial radiation at different times (different temperature scales

at which they are no longer in thermal equilibrium) and subsequently freely move through the universe at large. This provides a natural explanation for the mass scales observed in the expanding hierarchy of the universe.

The photon is the lightest boson, massless and with integer spin. It is a stable particle. Its supersymmetric (SUSY) partner is the *photino*, the lowest mass of the SUSY zoo. The photino freezes out of the background at very high temperatures, of order $T_{\text{photino}} \approx 4(m_{\text{SUSY,f}}/100\ (\text{GeV})^{4/3}$ MeV, where $m_{\text{SUSY,f}}$ is the mass of the SUSY particle (a fermion) by which the photino interacts with the background fermions. If the mass is very small, less than 4 MeV, the decoupling occurs early in the expansion, leading to relativistic particles and free streaming on a large scale; above this, the particle can come into equilibrium with the background through interactions that will reduce its streaming velocity, and annihilation processes, like those of monopoles, will dominate its equilibrium abundance. The final abundance of the photino is dominated by the strength of the SUSY interactions at temperatures higher than T_{photino}.

If the gravitino, the partner of the graviton, exists, then it is possible for the photino to decay into this SUSY fermion and a photon. Limits on the half-life can be as short as seconds or longer than days, suggesting that massive photinos may be responsible for mediating some of the nucleosynthesis in the late epoch of the expansion. If the photinos decay into photons, this raises the temperature of the background at precisely the time at which it may alter the formation of the light isotopes, photodisintegrating ^{4}He and changing the final abundances from nucleosynthesis. Recent attempts to place limits on the processes have shown that the photinos, in order to be viable candidates for DM, must be stable, or have a large mass (greater than a few GeV) if they are nonrelativistic, or have a very low mass ($<$100 eV). There are problems with large numbers of low-mass particles for the same reason there are problems with ν_e, since these would tend to wash out much of the smallscale structure while not being well bound to galaxy halos.

Strings are massive objects that are the product of GUTs. They are linear, one-dimensional objects that behave like particles. Arguments about their structure and behavior have shown that, as gravitating objects, they can serve as sites for promoting galaxy formation, and may be responsible for the biasing of the DM to form potential wells that accrete the baryons out of which galaxies are composed. There is no experimental support for their existence, but there is considerable interest in them from the point of view of particle theories. Strings have the property that they can either come in infinite linear or closed forms. The closed forms serve as the best candidates for inclusion in the DM scenarios.

IX. FUTURE PROSPECTS

The current status of DM is very confusing, especially in light of the wealth of possible models for its explanation. None of the particles presently known can explain the behavior of matter on scales above the galactic, while there are a number of hypothetical particles that can do so for larger scales. The masses of galaxies are such that, in order to explain their halos, the particles of which they are composed must be nonrelativistic, that is, they must have velocities less than a few hundred km sec^{-1}. Thus, cold DM appears the best candidate for the constituents of galactic halos. On the scale of clusters and superclusters, however, it is still not clear whether this is a firm limit. Biased galaxy formation seems a viable explanation of the distribution of luminous matter, but here again there is a wealth of explanation and not much data with which to test it. A summary of the evidence for DM is given in Table I, and a list of candidates is given in Table II.

The best candidate for testing some of the models is space observation. The Hubble Space Telescope is a high-resolution optical and ultraviolet spectroscopic and imaging instrument capable of reaching 30th magnitude and of performing high-resolution observations of galaxies to at least $z = 4$. Surveys of redshift distributions in clusters of galaxies with very high velocity resolution should aid greatly in the details of virial mass calculations, while the imaging should delineate the extent to which interactions between cluster galaxies have played a role in the formation of the observed galactic distributions. The Cosmic Background Explorer (COBE), launched in November 1989, was designed to look at the isotropy of the microwave background, near the peak of the CBR spectrum. COBE has placed strict limits on the scales on which the background shows temperature variations, thereby delimiting the scale of adiabatic perturbations in the expanding universe at the recombination epoch.

TABLE I Summary of Evidence for Dark Matter in Different Environments

Evidence	$\langle M/L \rangle$	Ω
Galactic stars and clusters	1	0.001
Vertical galactic gravitational acceleration	2–3	0.002
Disk galaxy dynamics	10	0.01
Irregular and dwarf galaxies	10–100	0.101–0.1
Binary galaxies and small groups	10–100	0.01–0.1
Rich clusters and superclusters	100–300	0.2 ± 0.1
Large-scale structure	1000	0.5–1.0
Baryosynthesis	10–100	$\leq$0.1
Inflationary universe	1000h	1.0

TABLE II Nonbaryonic Dark-Matter Candidates and Their Properties[a]

Candidate/particle	Approximate mass	Predicted by	Astrophysical effects
G(R)	—	Non-Newtonian gravitation	Mimics DM on large scales
Λ (cosmological constant)	—	General relativity	Provides $\Omega = 1$ without DM
Axion, majoron, Goldstone boson	10^{-5} eV	QCD; PQ symmetry breaking	Cold DM
Ordinary neutrino	10–100 eV	GUTs	Hot DM
Light higgsino, photino, gravitino, axino, sneutrino[b]	10–100 eV	SUSY/SUGR	Hot DM
Para-Photon	20–400 eV	Modified QED	Hot/warm DM
Right-handed neutrino	500 eV	Superweak interaction	Warm DM
Gravitino, etc.[b]	500 eV	SUSY/SUGR	Warm DM
Photino, gravitino, axino, mirror-particle, simpson neutrino[b]	keV	SUSY/SUGR	Warm /cold DM
Photino, sneutrino, higgsino, gluino, heavy neutrino[b]	MeV	SUSY/SUGR	Cold DM
Shadow matter	MeV	SUSY/SUGR	Hot/cold (like baryons)
Preon	20–200 TeV	Composite models	Cold DM
Monopoles	10^{16} GeV	GUTs	Cold DM
Pyrgon, maximon, Perry Pole, newtorites, Schwarzschild	10^{19} GeV	Higher dimension theories	Cold DM
Supersymmetric strings	10^{19} GeV	SUSY/SUGR	Cold DM
Quark nuggets, nuclearites	10^{15} g	QCD, GUTs	cold DM
Primordial (mini) black holes	10^{15-30} g	General relativity	Cold DM
Cosmic strings, domain walls	10^{8-10} M	GUTs	Promote galaxy formation, though small contributor to Ω

[a] QCD, quantum chromodynamics; PQ, Peccei–Quinn; GUTs, grand unified theories; SUSY, supersymmetry; SUGR, supergravity; QFD, quantum electrodynamics.

[b] Of these various supersymmertric partners predicted by assorted versions of SUSY/SUGR, only one, the lightest, can be stable and contribute to Ω, but the theories do not at present tells us which one it will be or the mass to be expected.

Laboratory tests of particle theories should also contribute to the DM problem. The plans for the Superconducting Supercollider show that it should be able to detect the Higgs boson, responsible for the mass of particles in grand unified theories, and this should also feed the modeling of supersymmetric interactions. As the knowledge of the behavior of the electroweak bosons increases, we should also have a clearer picture of the role played by SUSY in the early universe and in the generation of the particles which are responsible for (at least) some of the DM.

On a final note, it may be said, possibly, we have this all wrong. The study of matter that cannot be seen but only felt on very large scales is obviously one driven by models and calculations. These have assumptions built into them that may or may not be justified in the context of the particular studies. Therefore, it may be the case that DM is more an expression of our ignorance of the details of the universe at large distances and on the cosmological scale than we currently believe. However, the need for invoking DM is very widespread in astrophysics: it is required in many explanations, models for it come in many varieties, its presence is indicated by different methods of mass determination, and it is a phenomenon that involves only classical dynamics. Most simple alternatives proposed so far have been very specifically tailored to the individual problems and are often too *ad hoc* to serve as fundamental explanations of all of the phenomena that require the presence of some form of DM. This is a field rich in speculation, but it is also a field rich in quantitative results. As more data are accumulated on the dynamics of galaxies and clusters of galaxies, we will clearly be able to distinguish between the "everything we know is wrong" school and that which detects the fingerprint of the early universe and its processes in the current world.

APPENDIX: THE DISCRETE VIRIAL THEOREM

Consider the motion of a particle about the center of mass of a cluster. The gravitational potential is the result of the interaction with all particles j, not the same as i, for the ith particle:

$$\Phi \equiv -G \sum_{i<j} \frac{m_j}{R_{ij}}, \tag{25}$$

where the masses are allowed to be different. The equations of motion for this particle is

$$m_i \mathbf{v}_i = -m_i \nabla \Phi. \tag{26}$$

Taking the product of this equation with the position of the ith particle (the scalar product) gives

$$\sum_i \mathbf{r_i} m_i \ddot{\mathbf{r}} = -2T + \frac{1}{2}\ddot{\mathrm{I}}, \tag{27}$$

where I is the moment of inertia of the cluster and T is the kinetic energy of the particles. For discrete particles, this is easily calculated. Now assume that R_{ij} is the scalar distance between the ith and jth particles, Then the equation of motion yields the discrete vitrial theorem:

$$\ddot{\mathrm{I}} = \sum_i m_i v_i^2 - G \sum_i \sum_{i<j} \frac{m_i m_j}{R_{ij}} = 2T + W. \tag{28}$$

Here W is the gravitational energy of the cluster, summed over all masses. This is the usual form of the virial theorem, with the additional term for the secular variation of the moment of inertia included. It is usually assumed that the system has been started in equilibrium, so that the geometry of the configuration is constant. This implies that one can ignore the variation in I. Now, assume that all particles have the same mass, m. This gives

$$-G \sum_{i<j} m_i^2 \langle R_{ij} \rangle^{-1} + \sum_i m_i \langle v_i^2 \rangle = 0. \tag{29}$$

We obtain only a two-dimensional spatial picture and a one-dimensional velocity picture of a three-dimensional distribution of galaxies in a cluster. The virial theorem applies to the fullphase space of the constituent masses and is three dimensional, of course, in spite of the fact that we can only see the line-of-sight velocities. Therefore, in order to obtain mass estimates from the virial constraint, we must make some assumption about the isotropy of the velocity distribution. The simplest and conventional assumption is that the radial velocity is related to the mean square velocity by $\langle v_i^2 \rangle = 3\langle V_{\mathrm{rad},i}^2 \rangle$ when averaged over the various orientations of the cluster. The *virial mass estimator* is then given by

$$M_{\mathrm{VT}} = \frac{3\pi N}{2G} \left(\frac{\sum_i V_{\mathrm{rad},i}^2}{\sum_{i<j} \rho_{ij}^{-1}} \right), \tag{30}$$

where now N is the number of (identical) particles and ρ_{ij} is the projected separation of the objects i and j in the sky. The coefficient results from assuming that these are randomly oriented and that $\langle R_{ij}^{-1} \rangle = (2/\pi)\rho_{ij}^{-1}$.

Several assumptions have been built into this derivation, among which are the assumptions of isotropy of the orbits of stars about the center of the cluster and the stability of the shape of the cluster. Notice that there will be an additional term in the equation of motion if there is an oscillation of the cluster with time and that this will reduce the estimated virial mass (although it will in general be a small term since it is inversely proportional to the square of the Hubble time). An additional potential is contributed by the DM, but it may not enter into the virial argument in the same way as this discrete particle picture. In the preceding discussion, the constituent moving galaxies were assumed to be the only cause of the gravitational field of the clusters. Now if there is a large, rigid mass distribution that forms the potential in which the galaxies move, the mass estimates from the virial theorem are incorrect. There will always be an additional term, W_{DM}, which adds to the gravitational energy of the visible galaxies but does not contribute to the mass, and this can alter the determination of the mass responsible for the binding of the visible tracers of the dynamics. One should exercise caution before accepting uncritically the statements of the virial arguments by carefully examining whether they are based on valid, perhaps case-by-case, assumptions.

SEE ALSO THE FOLLOWING ARTICLES

COSMIC INFLATION • COSMOLOGY • DENSE MATTER PHYSICS • GALACTIC STRUCTURE AND EVOLUTION • NEUTRINO ASTRONOMY • STELLAR STRUCTURE AND EVOLUTION

BIBLIOGRAPHY

Aaronson, M., Bothun, G., Mould, J., Huchra, J., Schommer, R. A., and Cornell, M. E. (1986). "A distance scale from the infrared magnitude/HI velocity–width relation. V. Distance moduli to 10 galaxy clusters, and positive detection of bulk supercluster motion toward the microwave anisotropy," *Astrophys. J.* **302,** 536.

Athanassoula, E., Bosma, A., and Papaioannou, S. (1987). "Halo parameters of spiral galaxies," *Astron. Astrophys.* **179,** 23.

Bahcall, □., and Soniera, □. (1981).

Bahcall, N. A., Ostriker, J. P., Perlmutter, S., and Steinhardt, P. J. (1999). "The cosmic triangle: Revealing the structure of the universe," *Science* **284,** 1481.

Bergström, L. (1999) "Non-baryonic dark matter," *Nucl. Phys. B (Proc. Suppl.)* **70,** 31.

Blumenthal, G. R., Faber, S. M., Primack, J. R., and Rees, M. J. (1984). "Formation of galaxies and large-scale structure with cold dark matter," *Nature* **311,** 517.

Boesgaard, A. M., and Steigman, G. (1985). "Big bang nucleosynthesis: Theories and observations," *Annu. Rev. Astron. Astrophys.* **23,** 319.

Dekel, A., and Rees, M. J. (1987). "Physical mechanisms for biased galaxy formation," *Nature* **326,** 455.

Dressler, A. (1984). "The evolution of galaxies in clusters," *Annu. Rev. Astron. Astrophys.* **22,** 185.

Faber, S. M., and Gallagher, J. S. (1979). "Masses and mass-to-light ratios of galaxies," *Annu. Rev. Astron. Astrophys.* **17,** 135.

Geller, M., and Huchra, J. (1990). *Science* **246,** 897.

Holt, S. S., Bennett, C. L., and Trimple, V. (eds.). (1991). "After the First Three Minutes," American Institute of Physics Press, New York.

Kashlinsky, A., and Jones, B. J. T. (1991). "Large-scale structure of the universe," *Nature* **349,** 753.

Lynden-Bell, D., *et al*. (1988). *Ap. J.* **326,** 19.

Lin, and Shu. (1965).

Ostriker, and Peebles. (1973).

Partridge, B. (2000), "The universe as a laboratory for gravity," *Class. Quantum Grav.* **17,** 2411.

Peebles, J. (1980). "Large-Scale Structure of the Universe," Princeton University Press, Princeton, NJ.

Peebles, P. J. E. (1993). "Principles of Physical Cosmology," Princeton University Press, Princeton, NJ.

Rubin, V. (1983). "Dark matter in spiral galaxies," *Sci. Am.* **248** (6), 96.

Saunders, W., *et al*. (1991). "The density field of the local universe," *Nature* **349,** 32.

Silk, J. (2000). "Cosmology and structure formation," *Nucl. Phys. B (Proc. Suppl.)* **81,** 2.

Sato, K., (ed.). (1999). "Cosmological Parameters and the Evolution of the Universe," Kluwer, Dordrecht, The Netherlands.

Spiro, M., Aubourg, E., and Palanque-Delabroille, N. (1999), "Baryonic dark matter," *Nucl. Phys. B (Proc. Suppl.)* **70,** 14.

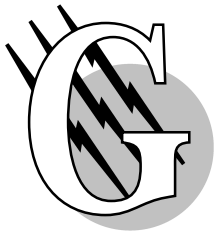
Trimble, V. (1988). "Dark matter in the universe: Where, what, why?" *Contemp. Phys.* **29,** 373.

van Moorsel, G. A. (1987). "Dark matter associated with binary galaxies," *Astron. Astrophys*. **176,** 13.

Vilenkin, A. (1985). "Cosmic strings and domain walls," *Phys. Rep.* **121,** 264.

White, S. D. M., Frenk, C. S., Davis, M., and Efstathiou, G. (1987). "Clusters, filaments, and voids in a universe dominated by cold dark matter," *Astrophys. J.* **313,** 505.

Zeld'ovich, Ya. B. (1984). "Structure of the universe," *Sov. Sci. Rev. E Astrophys. Space Phys.* **3,** 1.



Galactic Structure and Evolution

John P. Huchra
Harvard-Smithsonian Center for Astrophysics

GLOSSARY

Globular cluster A dense, symmetrical cluster of 10^5 to 10^6 stars generally found in the halo of the galaxy. Globular clusters are thought to represent remnants of the formation of galaxies.

Halo The extended, generally low surface brightness, spherical outer region of a galaxy, usually populated by globular clusters and population II stars. Sometimes also containing very hot gas.

H II region A region of ionized hydrogen surrounding hot, usually young, stars. These regions are distinguished spectroscopically by their very strong emission lines.

Hubble constant The constant in the linear relation between velocity and distance in simple cosmolical models. $D = V/H_0$, where H_0 is given in km sec^{-1} Mpc^{-1}. Distance and velocity are measured in Megaparsecs and in kilometers per second, respectively. Current values for H_0 range between 50 and 100 in those units.

Magnitude Logarithmic unit of relative brightness or luminosity. The magnitude scale is defined as $-2.5^* \log(\textit{flux}) + \textit{Constant}$ so that brighter objects have smaller magnitudes. Apparent magnitude is defined relative to the brightness of the A0 star Vega (given $\textit{magnitude} = 0$) and absolute magnitude is defined as the magnitude of an object if it were placed at a distance of 10 parsecs.

Metallicity The average abundance of elements heavier than H and He in astronomical objects. This is usually measured relative to the metal abundance of the Sun and quoted as "Fe/H," since iron is a major source of line in optical spectra.

Missing mass The excess mass found from dynamical studies of galaxies and systems of galaxies over and above the mass calculated for these systems from their stellar content.

Parsec, kiloparsec, Megaparsec 1 parsec is the distance at which 1 Astronomical Unit (the Earth orbit radius) subtends 1 arcsec. 1 pc $= 3.086 \times 10^{18}$ cm $= 3.26$ light years.

Population I, II Stars in the galaxy are classified into two general categories. Population II stars were formed at the time of the galaxy's formation, are of low metallicity, and are usually found in the halo or in globular clusters. Population I stars, like the sun, are younger, more metal rich and form the disk of the galaxy.

Solar mass, luminosity

$$L_\odot = 3.826 \times 10^{33} \text{ ergs sec}^{-1};$$
$$M_\odot = 1.989 \times 10^{33} \text{ gm}.$$

Surface brightness Luminosity per unit area on the sky, usually given in magnitudes per square arcsecond. In Euclidean space, surface brightness is distance independent since both the apparent luminosity and area decrease as the square of the distance.

THE OBJECT of the study of galaxies as individual objects is twofold—(1) understand the present morphological appearences of galaxies and their internal dynamics, and (2) to understand the integrated luminosity and energy distributions of galaxies—in both the framework and the timescale of a cosmological model and evolution. The morphology of a galaxy is determined by its formation and dynamical evolution, mass, angular momentum content, and the pattern of star formation that has occurred in it. The Brightness and spectrum of a galaxy are determined by its present stellar population which is in turn a function of the galaxy's star formation history and the evolution of those stars which is also related to the dynamical history of the galaxy and its environment. Significant advances have been made in detailed quantitative modeling of galaxy morphology and internal dynamics. Galaxy form and luminosity evolution are much less well understood despite their great importance for the use of galaxies as cosmological probes.

The study of galactic structure and evolution began only in the early 20th century with the advent of large reflectors and new photographic and spectroscopic techniques that allowed astronomers to determine crude distances to external galaxies and place them in the model universes being developed by Einstein, de Sitter, Friedmann, Lemaître and others. The key discoveries were Harlow Shapley's of a relation between the period and luminosity of Cepheid variable stars, and Edwin Hubble's use of that relation to determine distances to the nearest bright galaxies. Hubble then was able to calibrate other distance indicators (such as the brightest stars in galaxies, which are 10–100 times brighter than Cepheids) to estimate distances to further galaxies. This led to his discovery in 1929 of the redshift-distance relation which not only established the expansion of the universe as predicted by Einstein but also established that the age or timescale of the universe was much greater than 100 million years. Current observational estimates for the age of the universe range between 14 and 17 billion years in the basic Hot Big Bang model, the Friedmann–Lemaître cosmological model which has been favored for the last few decades.

In the past few years, the parameters of the cosmological model have been narrowed down by astronomers and physicists; the current view is that the Einstein–DeSitter globally flat model is most likley correct, that the Universe will probably expand forever, and that matter makes up about one third of the "content" of the Universe with one sixth of that (or about 4–5% of the total mass–energy density of the Universe) in ordinary baryonic matter—the stuff we are made of. The work of astronomers who concentrate on the study of galaxies is to understand the formation and evolution of galaxies and larger structures such as galaxy clusters from the time ∼15 billion years ago when, as seen in the Cosmic Microwave Background, the Universe was uniform and homogeneous to one part in 100,000.

I. GALAXY MORPHOLOGY—THE HUBBLE SEQUENCE

The initial steps in the study of galactic structure were the development of classification schemes that could be tied to physical properties such as angular momentum content, gas content, mass, and age. Another of Hubble's major contributions to extragalactic astronomy was the introduction of such a scheme based only on the galaxy's visual (or, more correctly, blue light) appearance. This morphological classification scheme, known as the Hubble Sequence, has been modified and expanded by Allan Sandage, Gerard de Vaucouleurs, and Sidney van den Bergh. The Hubble Sequence now forms the basis for the study of galactic structure. There are other classification schemes for galaxies based on such properties as the appearance of their spectra or the existence of star like nuclei or diffuse halos; these are of more specialized use.

Hubble's basic scheme can be described as a "tuning fork" with elliptical galaxies along the handle, the two families of spirals, normal and barred, along the tines, and irregular galaxies at the end (see Fig. 1). Elliptical galaxies

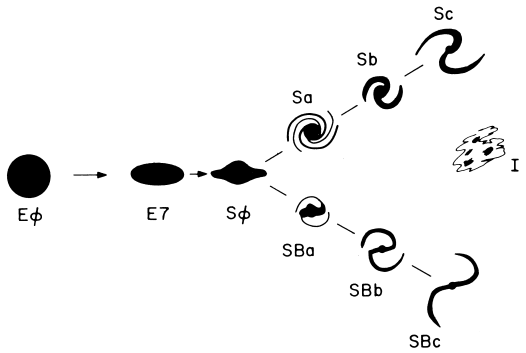


FIGURE 1 Hubble's "tuning fork" diagram (adapted from Hubble, E. 1936.) The Realm of the Nebulae.

FIGURE 2 The EO galaxy NGC 3379 and its comparison SBO galaxy, NGC 3384. This image was obtained with a Charge Coupled Device (CCD) camera at the F. L. Whipple Observatory by S. Kent. Slight defects in the image are due to cosmetic defects in the CCD.

are very regular and smooth in appearance and contain little or no dust or young stars. They are subclassified by ellipticity which is a measure of their apparent flattening. Ellipticity is computed as $e = 10(a-b)/a$, where a and b are the major and minor axis diameters of the galaxy. A typical elliptical galaxy is shown in Fig. 2. A galaxy with a circular appearance has $e = 0$ and is thus classified E0. These are at the tip of the tuning fork's handle. The flattest elliptical galaxies that have been found have $e \approx 7$, and are designated E7. Both normal and barred spiral galaxies range in form from "early" type, designated Sa, through type Sb, to "late" type Sc. Spiral classification is based on three criteria: (1) the ratio of luminosity in the central bulge to that in the disk, (2) the winding of the spiral arms, and (3) the contrast or degree of resolution of the arms into stars and H II regions. Early type galaxies have tightly wound arms of low contrast and large bulge-to-disk ratios. A typical spiral galaxy is shown in Fig. 3. The important transition region between elliptical and spiral galaxies is occupied by lenticular galaxies that are designated S0. It was originally hypothesized that spiral galaxies evolved into S0 galaxies and then ellipticals by winding up their arms and using up their available supply of gas in star formation. This is now known to be false.

Irregular galaxies are split into two types. Type I or Magellanic irregulars are characterized by an almost complete lack of symmetry, no nucleus, and are usually resolved into stars and H II regions. The Magellanic Clouds, the nearest galaxies to our own, are examples of this type. Type II irregular galaxies are objects that are not easily classified—galaxies that have undergone violent dynamical interactions, star formation events, "explosions," or have features uncharacteristic of their underlying class such as a strong dust absorption lane in an elliptical galaxy.

The modifications and extensions to Hubble's scheme added by Sandage and de Vaucouleurs include the addition of later classes for spirals, Sd and Sm, between

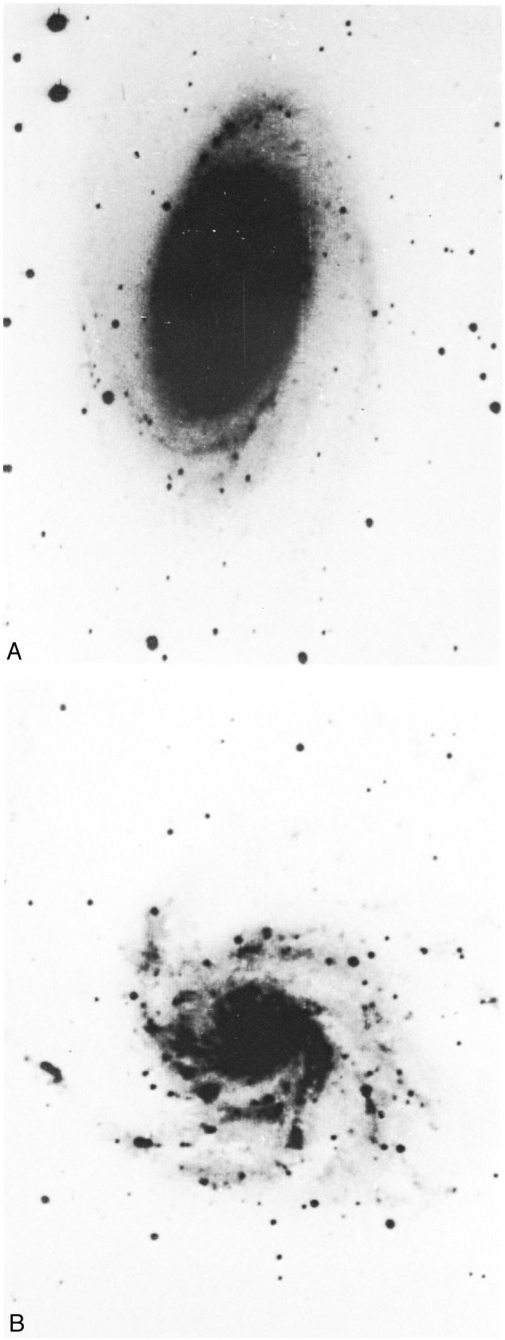


FIGURE 3 (a) The Sb spiral galaxy Messier 81 (=NGC 3031), and (b) the Sc spiral M101 (=NGC 5457 a.k.a. the Pinwheel). The CCD images courtesy of S. Kent.

TABLE I De Vaucouleurs' *T* Classification

T Type	Description
−6	Compact elliptical
−5	Elliptical, dwarf elliptical E, dE
−4	Elliptical E
−3	Lenticular L−, SO−
−2	Lenticular L, SO
−1	Lenticular L+, SO+
0	SO/a, SO-a
1	Sa
2	Sab
3	Sb
4	Sbc
5	Sc
6	Scd
7	Sd
8	Sdm
9	Sm, Magellanic Spiral
10	Im, Irr I, Magellanic Irregular, Dwarf Irregular
11	Compact Irregular, Extragalactic HII Region

classes Sc and Irr I, the further subclassification of spirals into intermediate types such as S0/a, Sab, and Scd, and the inclusion of information about inner and outer ring structures in spirals. S0 galaxies have also been subdivided into classes based either on the evidence for dust absorption in their disks, or, for SB0's, the intensity of their bar components. Van den Bergh discovered that the contrast and developement of spiral arms were correlated with the galaxy's luminosity. Spirals and irregulars can be broken into nine luminosity classes (I, I–II, II, . . . V), but there is considerable scatter (≈0.6–1.0 magnitude) and thus considerable overlap in the luminosities associated with each class. De Vaucouleurs also devised a numerical scaling for morphological type, called the T type and illustrated in Table I, for his *Second Reference Catalog of Bright Galaxies*.

The most recent "modification" to the Hubble sequence is van den Bergh's recognition of an additional sequence between lenticular galaxies and normal spirals. This sequence of spirals, dubbed *Anemic*, has objects designated Aa, ABa, Ab, etc. These Anemic spirals have diffuse spiral features, are usually of low surface brightness, and are gas poor relative to normal spirals of the same form.

II. GALACTIC STRUCTURE

The quantitative understanding of galactic structure is based on only two observables. These are the surface brightness distribution (including the structure of the arms in spiral galaxies) and the line-of-sight (radial) velocity field. The surface brightness at any point in a galaxy's image is the integral along the line-of-sight through the galaxy of the light produced by stars and hot gas. Measurements of the velocity field are made spectroscopically (either optically or in the radio at the 21-cm emission line of neutral H), and represent the integral along the line-of-sight of the velocities of individual objects (stars, gas clouds) times their luminosity. Dust along the line-of-sight in a galaxy obscures the light from objects behind it; in dusty, edge-on spiral galaxies only the near side of the disk can be observed in visible light. Extinction by dust is a scattering process and thus is a function of wavelength; all galaxies are optically thin at radio wavelengths. Elliptical galxies are considered optically thin at all wavelengths.

A. Elliptical Galaxies

Most galaxies can be readily decomposed into two main structural components, disk and bulge. Elliptical galaxies are all bulge. The radial brightness profile of the bulge component is usually parameterized by one of three laws. The earliest and simplest is an empirical relation called the Hubble law,

$$\mu(r) = \mu_o(1 + r/r_o)^{-2},$$

and is parametrized in terms of a central surface brightness, μ_o, and a scale length, r_o, at which the brightness falls to half its central value. At large radius, the profiles are falling as $1/r^2$. At radii less than the scale length, the profile flattens to μ_o. Giant elliptical galaxies typically have $\mu_o \approx 16$ mag/arcsec^2. A better relation is the empirical $r^{1/4}$ law proposed by de Vaucouleurs

$$\mu(r) = \mu_e \exp\left[-7.67\left((r/r_e)^{1/4} - 1\right)\right],$$

where r_e is the effective radius and corresponds to the radius that encloses 1/2 of the total integrated luminosity of the galaxy, and μ_e is the surface brightness at that radius, approximately 1/2000 of the central surface brightness. The third relation is semiempirical and was derived from dynamical models calculated by King to fit the brightness profiles of globular star clusters. These models can be parametrized as

$$\mu(r) = \mu_K\left[\left(1 + r^2/r_c^2\right)^{-1/2} - \left(1 + r_t^2/r_c^2\right)^{-1/2}\right]^2,$$

where r_c again represents the core radius where the surface brightness falls to ≈0.5, r_t is the truncation or tidal radius beyond which the surface brightness rapidly decreases, and μ_K is approximately the central surface brightness. Isolated elliptical galaxies are best fit by models with $r_t/r_c \approx 100$–200. Small elliptical galaxies residing in the gravitational potential wells of more luminous galaxies (like the dwarf neighbors of our own galaxy) are tidally

stripped and have $r_t/r_c \approx 10$. Figure 4 shows examples of the Hubble and de Vaucouleurs laws and the King models.

Dwarf elliptical galaxies, designated dE, are low luminosity, very low surface brightness objects. Because of their low surface brightness, they are difficult to identify against the brightness of the night sky (airglow). The nearest dwarf elliptical galaxies, Ursa Minor, Draco, Sculptor, and Fornax, are satellites of our own galaxy, and are resolved into individual stars with large telescopes. Although dE galaxies do not contribute significantly to the total luminosity of our own Local Group of galaxies, they dominate its numbers (Table II). As mentioned earlier, dwarf galaxies are often tidally truncated by the gravitational field of neighboring massive galaxies. If M is the mass of the large galaxy, m is the mass of the dwarf, and R is their separation, then the tidal radius, r_t, is given by

$$r_t = R\left(\frac{m}{3M}\right)^{1/3}.$$

Frequently the central brightest galaxy in a galaxy cluster exhibits a visibly extended halo to large radii. These objects were first noted by W. W. Morgan and are designated "cD" galaxies in his classification scheme. Unlike ordinary giant E galaxies, whose brightness profiles exhibit the truncation characteristic of de Vaucouleurs or King profiles at radii of 50 to 100 kpc, cD galaxies have profiles which fall as $1/r^2$ or shallower to radii in excess of 100 kpc. "D" galaxies are slightly less luminous and have weaker halos; D galaxies are found at maxima in the density distribution of galaxies. The extended halos of these objects are considered to be the result of dynamical processes that occur during either the formation or during the subsequent evolution of galaxies in dense regions.

TABLE II Presently Known Local Group Members

Name	Type	B_T[a]	Distance (kpc)	Luminosity ($10^9 L_\odot$)
Andromeda	Sb	4.38	730	14.7
Milky Way	Sbc	—	—	(10.0)[b]
M33	Scd	6.26	900	3.9
LMC	SBm	0.63	50	2.2
SMC	Im	2.79	60	0.4
NGC 205	E5	8.83	730	0.24
M32	E2	9.01	730	0.20
IC 1613	Im	10.00	740	0.09
NGC 6822	Im	9.35	520	0.08
NGC 185	dE3	10.13	730	0.07
NGC 147	dE5	10.37	730	0.06
Fornax	dE	9.1	130	0.006
And I	dE	13.5	730	0.003
And II	dE	13.5	730	0.003
And III	dE	13.5	730	0.003
Leo I	dE	11.27	230	0.0026
Sculptor	dE	10.5	85	0.007
Leo II	dE	12.85	230	0.0006
Draco	dE	12.0	80	0.00016
LGS 3	?	(17.5)[b]	730	0.00008
Ursa Minor	dE	13.2	75	0.00005
Carina	dE	—	170	—

[a] B_T is the total, integrated blue apparent magnitude.
[b] Quantities in parentheses are estimated.

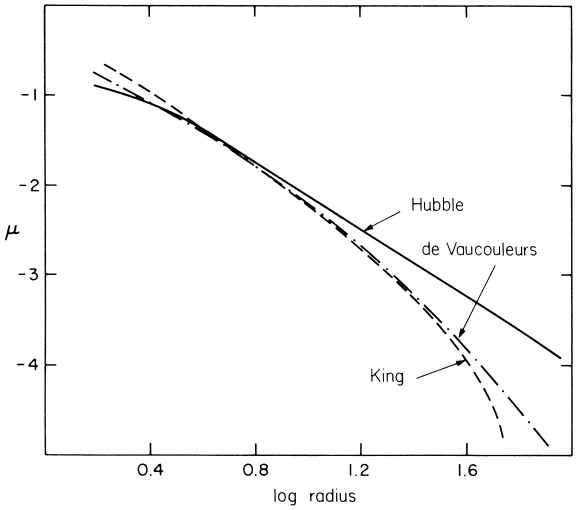


FIGURE 4 The three commonly fit surface brightness profiles for elliptical galaxies and the bulge component of spiral galaxies. The Hubble law (solid line) has core radius $r_0 = 1.0$ unit, the de Vaucouleurs law (dash-dot line) has an effective radius $r_e = 5.0$ units, and the King model (dashed line) has a core radius $r_c = 1.0$, and a tidal radius $r_t = 80.0$ units. All three profiles have been scaled vertically in surface brightness to approximately agree at $r = 5.0$ units ($\log r = 0.7$).

The internal dynamics of spheroidal systems (E galaxies and the bulges of spirals) is understood in terms of a self-gravitating, essentially collisionless gas of stars. These systems appear dynamically relaxed; they are basically in thermodynamic equilibrium as isothermal spheres with Maxwellian velocity distributions. Faber and Jackson noted in 1976 that the luminosities of E galaxies were well correlated with their internal velocity dispersions.

$$L \approx \sigma^4.$$

Here, the velocity dispersion, σ, is a measure of the random velocities of stars along the line-of-sight.

The collisional or two-body relaxation time, t_r, for stars is approximately

$$t_r \sim 2 \times 10^8 \frac{V^3}{M^2 \rho} \text{ years},$$

where V is the mean velocity in km sec^{-1}, ρ is the density of stars per cubic parseconds and M is the mean stellar mass in $M_\odot$. This relaxation time in galaxies is 10^{14} to 10^{18} years—much longer than the age of the universe. Two-body relaxation is generally not important in the internal dynamics of galaxies, but is significant in globular clusters.

The relaxed appearance of galactic spheroids is probably due to the process called violent relaxation. This is a statistical mechanical process described by Lynden-Bell, where individual stars primarily feel the mean gravitational potential of the system. If this potential fluctuates rapidly with time, as in the initial collapse of a galaxy, then the energy of *individual* stars is not conserved. The results of numerical experiments are similar to galaxy spheroids.

For the past decade and a half, the determination and modeling of the true shapes of elliptical galaxies has been a major problem in galaxian dynamics. Most elliptical galaxies are somehat flattened. Early workers assumed that this flattening was rotationally supported as in disk galaxies (see following). Bertola and others observed, however, that the rotational velocites of E galaxies are insufficient to support their shapes (Fig. 5). The velocity dispersion is a measure of the random kinetic energy in the system.

To resolve this problem, Binney, Schwarzchild, and others suggested that E galaxies might be prolate (cigar shaped) or even triaxial systems instead of oblate (disk-like) spheroids. Current work favors the view that most flattened elliptical galaxies are triaxial and have internal stellar velocity distributions which are anisotropic. A small fraction do rotate fast enough to support their shapes.

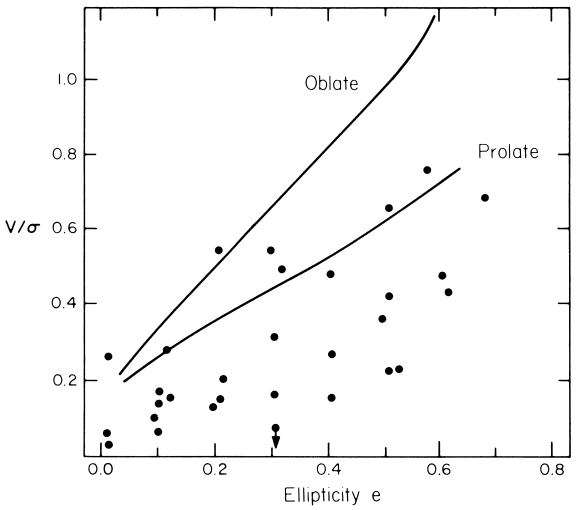


FIGURE 5 The ratio of rotation velocity to central velocity disperson *versus* the ellipticity, *e*, for elliptical galaxies. Solid lines are oblate and prolate spheroid models with isotropic velocity distributions from Binney. Data points are from Bertola, Cappacioli, Illingworth, Kormendy and others.

B. S0 Galaxies

Spiral and S0 galaxies can be decomposed into two surface brightness components, the bulge and disk. Disks have brightness profiles that fall exponentially

$$\mu(r) = \mu_0 \exp(-r/r_s).$$

Bulges have been described earlier. Disks are rotating; their rotational velocity at any radius is presumed to balance the gravitational attraction of the material inside. A typical rotation curve (velocity of rotation as a function of radius) is shown in Fig. 6.

Disks are not infinitely thin. The thickness of the disk depends on the balance between the surface mass density in the disk (gravitational potential) and the kinetic energy in motion perpendicular to the disk. This can be a function of both the initial formation of the system and its later interaction with other galaxies, etc. Tidal interaction with other galaxies will "puff up" a galactic disk of stars. Disks of S0 galaxies are composed of old stars and do not exhibit any indications of recent star formation and associated gas or dust—that is part of the definition of an S0.

Bulges of S0 galaxies and spirals **do** rotate and appear to be simply rotationally flattened oblate spheroids. We will return to this point later when we discuss galaxy formation.

C. Spiral Galaxies

Unlike the smooth and uncomplicated disks of S0 galaxies, the disks of spiral galaxies generally exhibit significant

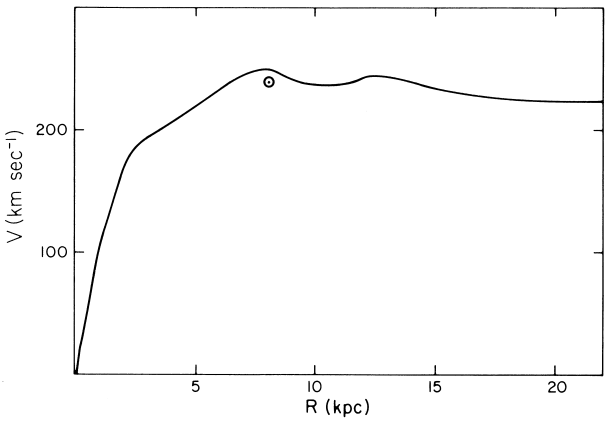


FIGURE 6 A typical rotation curve for a moderately bright spiral galaxy like our own (adapted from the work of V. Rubin *et al.* and M. Roberts). For reference, the rotational velocity of our own galaxy at the approximate radial distance of the sun, 220 km sec^{-1} at 8 pc, is marked with an $\odot$.

amounts of interstellar gas and dust and distinct spiral patterns. The disk rotates differentially (see Fig. 6) and the spiral pattern traces the distribution of recent star formation in the galaxy. Regions of recent star formation have a higher surface brightness than the background disk. Although the spiral is a result of the rotation of the material in the disk, the pattern need not (and generally does not) rotate with the same speed as the material.

The rotation of a spiral galaxy can be described in terms of a velocity rotation curve, $v(r)$, or the angular rotation rate, $\Omega(r)$, where $v = r\Omega$. In 1927, Oort first measured the local differential rotation of our galaxy by studying the motions of nearby stars. He described that rotation in terms of two constants, now known as Oort's Constants, which are measures of the local shear (A) and vorticity (B):

$$A = -\frac{r}{2}\frac{d\Omega}{dr}, \qquad B = -\frac{1}{2r}\frac{d}{dr}(r^2\Omega).$$

The values adopted by the IAU (International Astronomical Union) in 1964 are $A = 15$ km sec^{-1} kpc^{-1} and $B = -10$ km sec^{-1} kpc^{-1}, with the Sun at a distance of $r_o = 10$ kpc from the center of the galaxy, and the rotation rate at the sun $V_\odot = 250$ km sec^{-1}. Current estimates support a smaller Sun-Galactic-center distance ($\sim$8 kpc) and a smaller rotation velocity ($\sim$230 km sec^{-1}) as well as slightly different values for A and B.

Studies of rotation in spiral galaxies are undertaken by long-slit spectroscopy at differenct position angles in the optical or by either integrated (total intensity, big beam) measurements or interferometric mapping in the 21-cm line of neutral hydrogen. The neutral hydrogen (HI) in spiral galaxies is primarily found outside their central regions, reaching a maximum in surface density several kiloparseconds from the center. The gas at the center is mostly in the form of molecular hydrogen, H_2, as deduced from carbon monoxide (CO) maps. From such detailed studies by Rubin, Roberts, and others we know that the rotation curves for spirals generally rise very steeply within a few kiloparseconds of their centers then flatten and stay at an almost constant velocity as far as they can be measured. This result is rather startling because the luminosity in galaxies is falling rapidly at large radii. If the light and mass were distributed similarly, then the rotational velocity should fall off as $1/R$ at large radii as predicted by Kepler's laws. Only a small number of galaxies show the expected Keplerian falloff, leading to the conclusion that the mass in spiral galaxies is **not** distributed as the light.

As in elliptical galaxies, Fisher and Tully noted in 1976 that the luminosities of spiral galaxies are correlated with their internal motions, in particular rotational velocity. The best form of this relation is seen in the near infrared

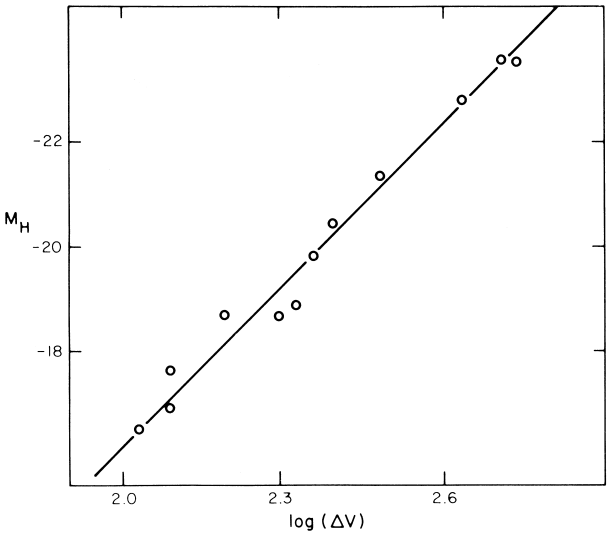


FIGURE 7 The relation between the maximum rotation velocity of spiral galaxies and their absolute infrared (H band = 1.6 μ) magnitude [Adapted from Aaronson, M., Mould, J., and Huchra, J. P. (1980) *Astrophys. J.* **237,** 655.]

where the effects of internal extinction by dust on the luminosities are minimized (Fig. 7). There the relation is approximately

$$L = (\Delta V)^4,$$

where ΔV is the full width of the H I profile measured at either 20 or 50% of the peak. As the H I distribution peaks outside the region where the rotational velocity has flattened, the H I profile is sharp sided and is double peaked for galaxies inclined to the line of sight.

The regularity of the spiral structure seen in these galaxies is exceptional. If the spiral pattern was merely tied to the matter distribution, differential rotation would "erase" it in a few rotation periods. A typical rotation time is a few hundred million years, or $\approx$1/100th the age of the universe. To explain the persistence of spiral structure, Lin and Shu introduced the Density Wave theory in 1964. In this model, the spiral pattern is the star formation produced in a shock wave induced by a density wave propagating in the galaxy disk. The spiral pattern is in solid body rotation with a pattern speed Ω_p. The main features of such a density wave are the corotation radius, where the pattern and rotation angular speeds are the same, and the inner and outer Lindblad resonances, where

$$\Omega_p = \Omega \pm \kappa/m;$$

m is the mode of oscillation and κ is the epicyclic frequency in the disk,

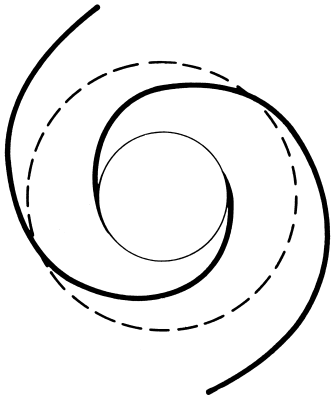


FIGURE 8 Typical two-armed spiral density wave pattern. Solid spiral represents the shock front in the gas. This follows closely behind the gravitational potential perturbation. The outer dashed circle is the corotation radius, where the matter in the disk is rotating at the same speed as the spiral pattern. The inner, thin circle represents the inner Lindblad resonance.

$$\kappa^2 = r^{-3}\frac{d(r^4\Omega^2)}{dr},$$

or

$$\kappa = \frac{2\Omega}{(1 - A/B)^{1/2}}.$$

Figure 8 depicts the features of a density wave. The spiral pattern only exists between the Lindblad resonances. Galaxies in which a single mode dominates are called "Grand Design" spirals.

Two alternative models have been proposed to account for spiral patterns, the Stochastic Star Formation (SSF) model of Seiden and Gerola, and the tidal encounter model. In the first, regions of star formation induce star formation in neighboring regions. With proper adjustment of the rotation and propagation timescales, reasonable spiral patterns result. In the second, spiral structure is the result of galaxy interactions. A possible example of this process is shown in Fig. 9. It is likely that all three processes operate in nature, with density waves producing the most regular spirals, SSF producing "flocculent" spirals, and tidal interactions producing systems like the Whirlpool. A significant fraction (∼10%) of all galaxies show some form of interaction with their neighbors. Not all such systems contain spiral galaxies, however.

D. Irregular Galaxies

Galaxies are classified as Irregular for several different reasons. In the morphological progression of the Hubble sequence the true Irregulars are the Magellanic Irregulars, the Irr I's, which are galaxies with no developed spiral structure that are usually dominated by large numbers of star forming regions. Irregular II's and other peculiar objects have been extensively cataloged by Arp and his coworkers, by Vorontsov–Velyaminov and by Zwicky. These objects are usually given labels to describe their peculiar properties such as "compact," which usually indicates abnormally high surface brightness or a very steep brightness profile, "posteruptive," which usually indicates the existence of jets or filaments of material near the galaxy, "interacting," or "patchy."

There are also galaxies in the form of rings which are thought to be produced by a slow, head-on collision of two galaxies, one of which must be a gas-rich spiral. The collision removes the nucleus of the spiral leaving a nearly round ripple of star formation similar to the ripples produced when a rock is dropped into a lake. Such ring galaxies almost always have compact companion galaxies which are the likely culprits.

The Magellanic irregulars are almost always dwarf galaxies (low luminosity), are very rich in neutral hydrogen, and have relatively young stellar populations. Their internal kinematics may show evidence for regular structure or may be chaotic in nature. These galaxies are generally of low mass; the largest such systems have internal velocity dispersions (usually measured by the width of their 21-cm hydrogen line) less than 100 km sec^{-1}. Their detailed internal dynamics have only been poorly studied until now. There are some indications that star formation proceeds in these galaxies as in the SSF theory mentioned earlier; however evidence also exists for H II regions aligned with possible shock fronts. Although they do not contribute significantly to the total luminosity density of the universe, these galaxies and the dwarf ellipticals dominate the total number of galaxies.

III. INTEGRATED PROPERTIES OF GALAXIES

A. Luminosity Function

The luminosity function or space density of galaxies, $\phi(L)$ is the number of galaxies in a given luminosity range per unit volume. This function is usually calculated from magnituded limited samples of galaxies with distance information. Distances to all but the nearest galaxies are determined from their radial velocities and the Hubble constant. (Note that in the Local Supercluster where the velocity field is disturbed by the gravity of large mass concentrations it is necessary to apply additional corrections to distances measured this way). Figure 10 is the differential luminosity function for field galaxies derived from a recent large survey of galaxy redshifts. The luminosity function is nearly flat at faint magnitudes and falls exponentially

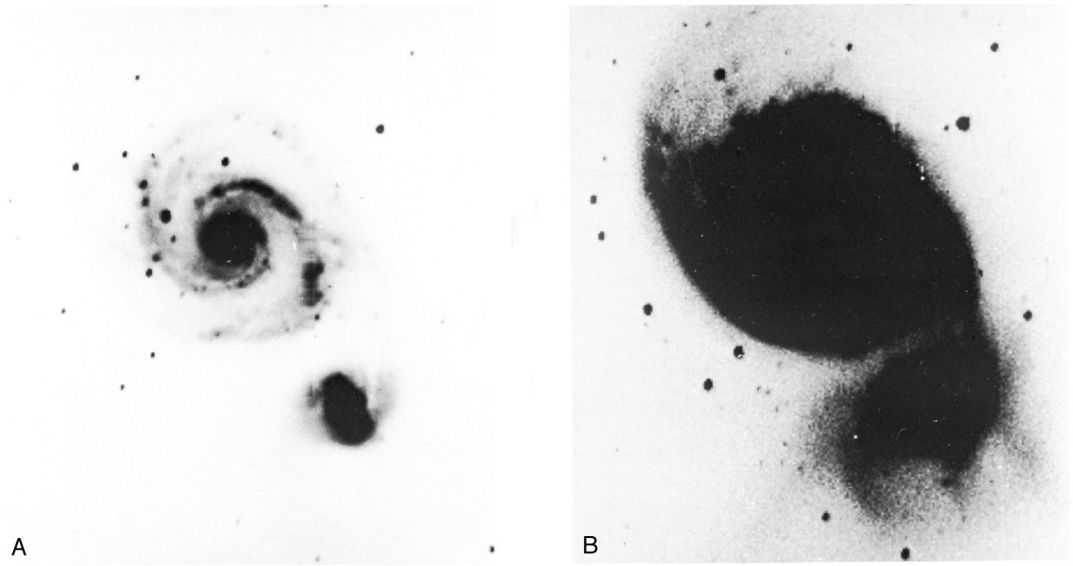


FIGURE 9 The interacting galaxy pair NGC 5194 + 5195, also known as M51 or the Whirlpool Nebula. This galaxy pair is also known as Arp 85, VV 1, and Ho 526 from the catalogs of peculiar, interacting and binary galaxies of Arp, Vorontsov–Velyaminov and Holmberg. The bright spiral, NGC 5194, is classified Sbc; its companion, NGC 5195, is classified SBO P. (a) is a low contrast display of the image to show the interior structure of the galaxies; (b) is a high contrast display of the same image to show the effects of interaction on the outer parts of the galaxies. (CCD photo courtesy of S. Kent.)

at the bright end. A useful parametrization of the $\phi(L)$, derived by Schechter, is

$$\phi(L) = \phi_0 L^{-1}(L/L^*)^{\alpha} \exp(-L/L^*).$$

L^* is the characteristic luminosity near the knee, ϕ_0 is the normalization and α is the slope at the faint end. For a Hubble constant of 100 km sec^{-1} Mpc^{-1} and with blue (B) magnitudes from the Zwicky Catalog of Galaxies and of Clusters of Galaxies

$$L_B^* \approx 8.6 \times 10^9 L_{\odot}, \quad M_B^* \approx -19.37,$$
$$\phi_0 \approx 0.015 \text{ gal Mpc}^{-3},$$

and

$$\alpha \approx -1.25.$$

M_B^* is the characteristic blue absolute magnitude. The Schechter function has the interesting property that the integral luminosity density, useful in cosmology, is just given by

$$L_{int} = \phi_0 L^* \Gamma(\alpha + 2)$$

where Γ is the incomplete gamma function.

Figure 11 shows the luminosity function of galaxies as a function of morphological type.

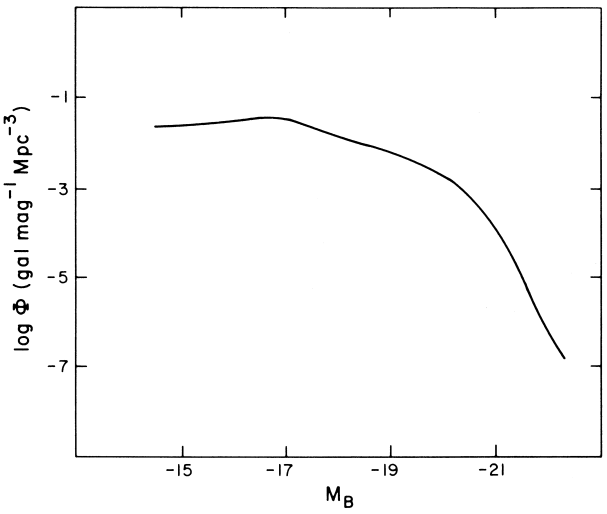


FIGURE 10 The differential galaxy luminosity function, $\phi(L)$, in units of galaxies per magnitude interval per cubic megaparsec, derived from the Center for Astrophysics Redshift Survey.

B. Spectral Energy Distributions

The observed integrated spectra of normal galaxies are functions of stellar population, star formation rate, mean metal content, gas content, and dust content. These properties correlate with, and in some cases are causually related to, galaxy morphology. In our own galaxy there are two relatively distinct populations of stars as discovered by Baade in 1944. Population II stars are the old metal poor stars that form the halo of the galaxy. Globular clusters are population II objects. Population I stars

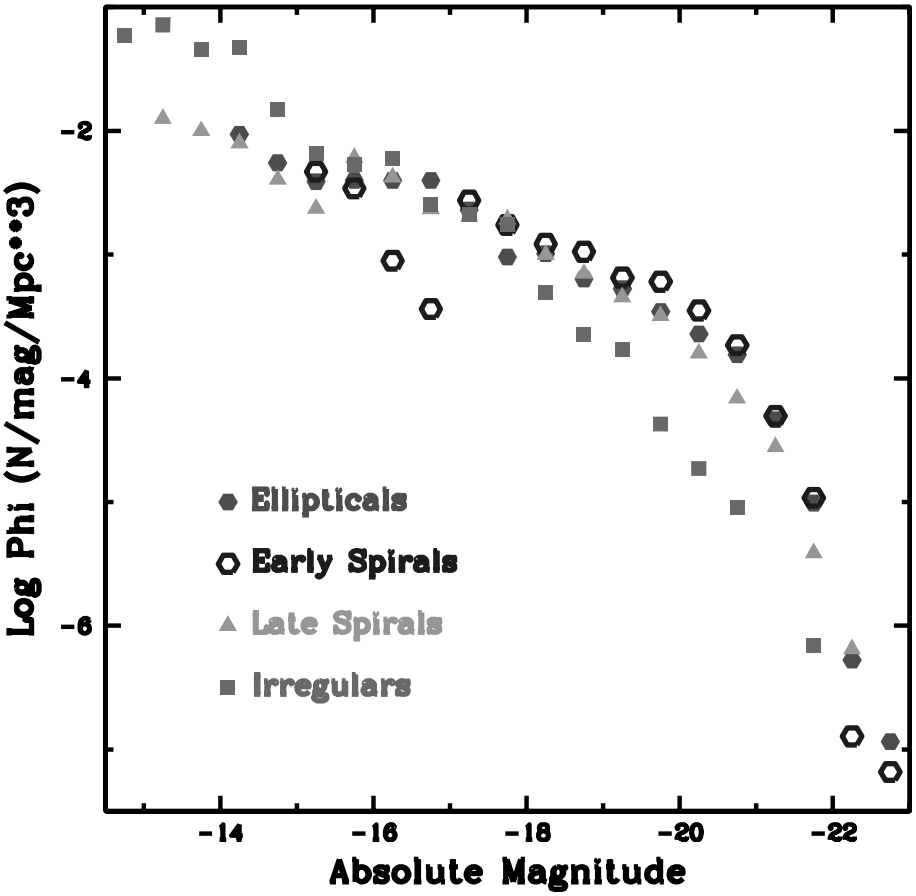


FIGURE 11

are younger and more metal rich. They form the disk of the galaxy. The sun is an intermediate age population I star.

Figure 12a is an example of the optical spectral energy distribution of an old, gasless stellar population. This type of population is typical of the bulge and old disk populations in galaxies. It is dominated by the light of G and K giant stars. The strongest spectral features are absorption lines of (originating in the atmospheres of the giant stars) calcium, iron, magnesium, and sodium, usually in low ionization states, as well as molecular bands of cyanogen, magnesium hydride and, in the red, titanium oxide. The very strong, factor of 2, break in the apparent continuum shortward of 4000 Å is the primary distinguishing characteristic of high redshift normal galaxies. The strengths of the absorption features in old bulge populations is primarily a function of metallicity. In populations where star formation has taken place recently (on timescales of $\leq 10^9$ years), most absorption features in the integrated spectra will appear weaker due to the contribution to the continuum emission from hotter, weaker lined A and F stars. The Balmer lines of hydrogen, which are strength in A and F stars, increase in strength. Table III lists some of the common and strong absorption and emission lines seen in normal galaxy spectra.

Figure 12b is the spectrum of a galaxy who's light is dominated by very young stars and hot gas. The spectrum resembles that of an H II region. This is an irregular galaxy that is undergoing an intense "burst" of star formation. Its optical spectrum is dominated by strong, sharp emission lines of H, He, and the light elements N, O, and S from the photoionized gas. The continuum is primarily from hot O and B stars with a small contribution from the free–free, two-photon, and Paschen and Balmer continuum emission from the gas.

Population synthesis is the attempt to reproduce the observed spectral energy distributions of galaxies by the summation of spectra of well-observed galactic stars, model stars, and models for the emission from the gas photoionized by hot stars in the synthesis. An important parameter in such studies is the Initial Mass Function (IMF), which is the differential distribution of stars as a function of mass in star forming regions. The simplest approximation for the IMF is a powerlaw in the mass, M,

$$N(M)\,dM \;=\; (M/M_{\odot})^{-\alpha}.$$

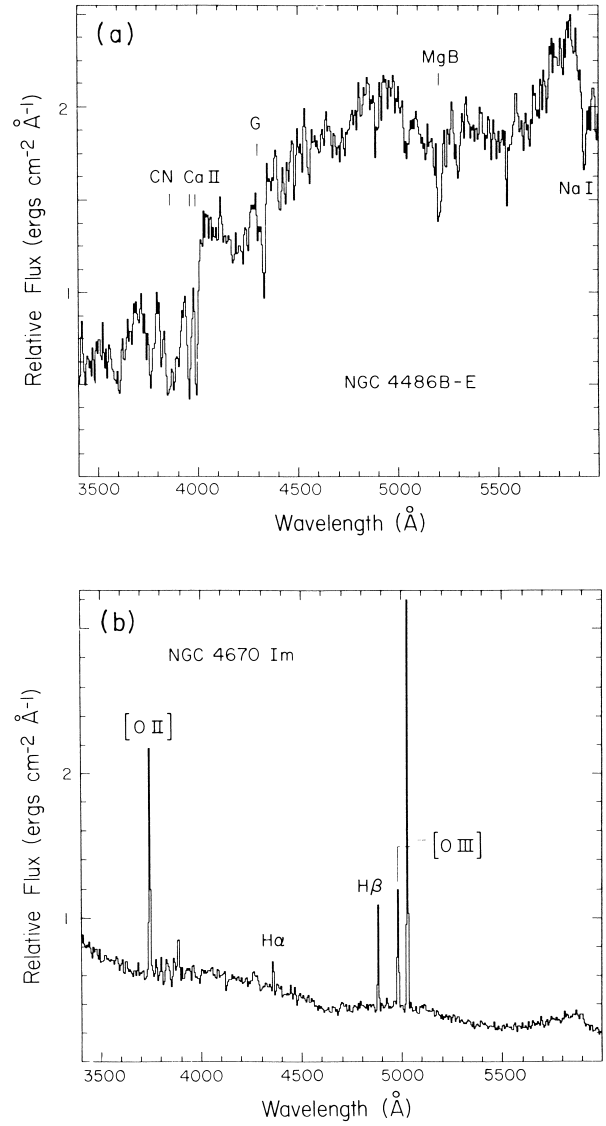


FIGURE 12 Galaxy Spectral energy distributions. (a) NGC 4486B, typical of an old, evolved stellar population. The prominent absorption features of Ca, Na, Mg, the CN molecule and the G band blend of Fe and Cr are marked. (b) NGC 4670, typical of a strong emission line galaxy or extragalactic H II region. The O and H emission lines are marked.

TABLE III Common Spectral Features in Galaxies

λ_0	Element	Comments[a]
	Absorption lines	
3810+	CN	(molecular band)
3933.68	Ca II	K
3968.49, 3970.08	Ca II + He	H
4101.75	Hδ	
4165+	CN	(molecular band)
4226.73	Ca I	
4305.5	Fe + Cr	G band
4340.48	Hδ	
4383.55	Fe I	
4861.34	Hβ	F
5167.3, 5172.7, 5183.6	Mg I	b
5208.0	MgH	(molecular band)
5270.28	Fe I	
5889.98, 5895.94	Na I	D
8542.0	Ca II	
	Emission lines[b]	
3726.0	[O II]	
3729.0	[O II]	
3868.74	[Ne III]	
3889.05	Hζ	
3970.07	Hε	
4101.74	Hδ	
4340.47	Hγ	
4363.21	[O III]	
4861.34	Hβ	F
4958.91	[O III]	
5006.84	[O III]	
5875.65	He I	
6548.10	[N II]	
6562.82	Hα	C
6583.60	[N II]	
6717.10	[S II]	
6731.30	[S II]	

[a] Letters in comments refer to Fraunhoffer's original designation for lines in the solar spectrum.

[b] Brackets around species indicate that the transition is forbidden.

The value of α found by Salpeter for the solar neighborhood is $\approx$2.35. The real IMF is slightly steeper at the high mass end ($M > 10\,M_\odot$), and slightly flatter at the low mass end ($M > 1\,M_\odot$). Results from population synthesis indicate that the mix of stars in most luminous galaxies is very similar to our own in age and metallicity.

C. Galaxy Masses

The masses of galaxies are measured by a variety of techniques based on measurements of system sizes and relative motion, and the assumption that the system under study is gravitationally bound. Masses are sometimes inferred from studies of stellar populations, but these are extremely uncertain as the lower mass limit of the IMF is essentially unknown. It is customary to quote the mass-to-light ratios for galaxies rather than masses. This is both because galaxy masses span several orders of magnitude so it is easier to compare M/L since mass and luminosity are reasonably well correlated for individual morphological classes, and because the average value of M/L is key in the determination of the mean matter density in the

universe. In specifying masses or mass-to-light ratios for galaxies, it is necessary to specify the Hubble constant, as scale lengths vary as the distance while luminosities vary inversely as the square of the distance. It is also useful to specify the magnitude system used for determining luminosity both because different systems measure light to different radii and because galaxies have a wide range of colors and are rarely the same color as the Sun.

Masses of individual spiral galaxies are determined from their rotation curves. Spiral galaxies are circularly symmetric, so the apparent radial velocities relative to their centers can be corrected for inclination. Then the velocity of the outermost point of the rotation curve is a measure of the enclosed mass. If the mass were distributed spherically, the problem would reduce to the classical circular orbit of radius, R, around point mass, M,

$$\frac{1}{2}mV^2 = \frac{GmM}{R}$$

where m is the test particle mass, and V is its orbital velocity, implying

$$M = \frac{1}{2}\frac{V^2R}{G}.$$

Because the actual distribution is flattened, a small correction must be applied. If the mass distributions of spiral galaxies fall with radius as their optical light, then rotation curves would turnover to a Keplerian falloff, $V \propto R^{-1/2}$. Observations of neutral hydrogen rotation curves made at 21 cm now extend beyond the optical diameters of many galaxies and are still flat. This implies that the masses of spirals are increasing linearly with radius, which in turn, implies that the local mass-to-light ratio in the outer parts of spirals is increasing exponentially.

Masses for individual E galaxies are determined from their velocity dispersions and measures of their core radii or effective radii. For a system in equilibrium, the Virial Theorem states that, averaged over time,

$$\Omega + 2T = 0,$$

where T is the system kinetic energy and Ω is the system gravitational potential energy. For a simple spherical system and assuming that the line-of-sight velocity dispersion, σ, is a measure of the mass weighted velocities of stars relative to the center of mass

$$M\sigma^2 = G\int_0^R \frac{M(r)\,dM}{r}.$$

If the galaxy is well approximated by a de Vaucouleurs Law, then $\Omega = -0.33GM^2r_e$. More accurate mass-to-light ratios can be determined for the cores of elliptical galaxies alone by comparison with models.

Masses for binary galaxies are also derived from simple orbital calculations, but, unlike rotation curve masses, are subject to two significant uncertainties. The first is the lack of knowledge of the orbital eccentricity (orbital angular momentum). The second is the lack of information about the projection angle on the sky. The combination of these two problems make the determination of the mass for an individual binary impossible. The formula $M = (V^2R)/(2G)$ produces a minimum mass, the "true" mass is

$$M_T = \frac{V^2R}{2G}\,\frac{1}{\cos^3 i\cos^2\phi},$$

where i is the angle between the galaxies and the plane of the sky and ϕ is the angle between the true orbital velocity and the plane defined by the two galaxies and the observer. The correct determination of the masses of galaxies in binary systems requires a large statistical sample to average over projections as well as a model for the selection effects that define the sample (e.g., very wide binaries are missed; the effect of missing wide pairs depends on the distribution of orbital eccentricities).

Mass-to-light ratios for galaxies in groups or clusters are also determined from the dimensions of the system and its velocity dispersion via the Virial Theorem or a variant called the Projected Mass method. The Virial Theorem mass for a cluster of N galaxies with measured velocities is

$$M_{VT} = \frac{3\pi N}{2G}\,\frac{\sum_i^N V_i^2}{\sum_{i<j} 1/r_{ij}},$$

where r_{ij} is the separation between the ith and jth galaxy, and V_i is the velocity difference between the ith galaxy and the mean cluster velocity. The projected mass is

$$M_P = \frac{f_p}{GN}\left(\sum_i^N V_i^2 r_i\right),$$

where r_i is the separation of the ith galaxy from the centroid. The quantity f_p depends on the distribution of orbital eccentrities for the galaxies and is equal to $64/\pi$ for radial orbits and $32/\pi$ for isotropic orbits.

Table IV summarizes the current state of mass-to-light determinations for individual galaxies and galaxies in systems via the techniques discussed earlier. These are scaled to a Hubble Constant of 100 km sec^{-1}, and the Zwicky

TABLE IV Mass-to-Light Ratios for Galaxies[a]

Type	Method	M/L
Spiral	Rotation Curves	12
Elliptical	Dispersions	20
All	Binaries	100
All	Galaxy Groups	350
All	Galaxy Clusters	400

[a] In solar units, $M_\odot/L_\odot$.

catalog magnitude system used earlier for the Luminosity Function.

In almost all galaxies, the expected mass-to-light ratio from population synthesis is very small, on the order of 1 ore lower in solar units, because the light of the galaxy is dominated by either old giant stars, with luminosities several hundred Suns and masses less than the Sun, or by young, hot main sequence stars with even lower mass-to-light ratios. The large mass-to-light ratios that result from the dynamical studies have given rise to what is called the "missing mass" problem. Astronomers as yet have not been able to determine what constitutes the mass that binds clusters and groups of galaxies and forms the halos of large galaxies. Possibilities include extreme red dwarf (low luminosity) stars, massive stellar remnants (black holes), exotic elementary particles (axions, neutrinos, etc.), and the possibilty that the dynamical state of clusters is much more complex than the existing simple models.

D. The Fundamental Plane

Much of the work on global properties of galaxies over the last 20 years has centered on finding global relationships like the Faber–Jackson relation and Tully–Fisher relation between velocity dispersion or rotation velocity and galaxy luminosity. The drivers for this are twofold: first to search for redshift independent distance estimators, and second to discover if such relations hold clues for the study of galaxy formation. The most useful of such relations discovered so far is the fundamental plane for early type galaxies, particularly elliptical galaxies. Ellipticals form a multiparameter family. The parameters that describe the global properties of elliptical galaxies as discovered by principal component analysis are the galaxy's size, surface brightness (related to stellar density), and velocity dispersion (related to mass). A typical example of such a relation is shown in Fig. 13, from Jorgensen *et al.* (1996). The equation describing this relation is

$$\log r_e = 1.35 \log \sigma - 0.82 \log \langle I \rangle_e,$$

where r_e is the effective (or half-light) radius, σ is the line-of-sight velocity dispersion and $\langle I \rangle_e$ is the mean surface brightness (luminosity inside the effective radius divided by the area). There are other variants of this relation such as the $D_n - \sigma$ relation which has been used by several groups to study the motions of galaxies relative to the uniform expansion of the Universe and thus derive estimates of the mean mass density of the Universe relative to the critical density.

E. Gas Content

Neutral hydrogen (H I) was first detected in galaxies in 1953 by Kerr and coworkers in the 21-cm (1420.40575 MHz) radio emission line. Since then it has been found that almost every spiral and irregular galaxy contains considerable neutral gas. In spiral galaxies, the H I emission distribution usually shows a central minimum and peaks in a ring which covers the prominent spiral arms. The fractional gas mass ranges from a few percent for early type spirals (Sa galaxies) to more than 50% for some Magellanic irregulars. Neutral hydrogen is only rarely detected in elliptical an S0 galaxies. The gas mass fraction in these objects is usually less than 0.1%. In spiral galaxies, H I can usually be detected at 21 cm out to two or three times the radius of the galaxy's optical image on the Palomar Schmidt Sky Survey plates.

Ionized gas is found in galaxies in H II regions (H II refers to singly ionized hydrogen), in nuclear emission regions, and in the diffuse interstellar medium. H II regions are regions of photoionized gas around hot, usually young stars. The gas temperature is of the order of 10^4 K, and depends on the surface temperatures of the ionizing stars and cooling processes in the gas which depend strongly on its element abundances. Low metallicity H II regions are hotter than those with high metallicity. Ionized gas is often seen in the nuclei of galaxies and is either the result of star formation (as in H II regions) or the result of photoionization by a central nonthermal energy source. In our galaxy, there is a diffuse interstellar medium composed of gas and dust. Much of the volume (although not much mass) of the galaxy is in this state with the gas ionized by the diffuse stellar radiatation field. Because its density is so low, ionized gas in this state takes a long time to recombine. (To a first approximation, the recombination time for diffuse, ionized hydrogen is

$$\tau \sim \frac{10^5}{n_e},$$

where n_e is the electron density in cm^{-3}, and τ is in years.) Ionized gas is also found in supernova remnants.

Carbon monoxide (CO), which is found in cool molecular clouds in the galaxy, has been detected in several other

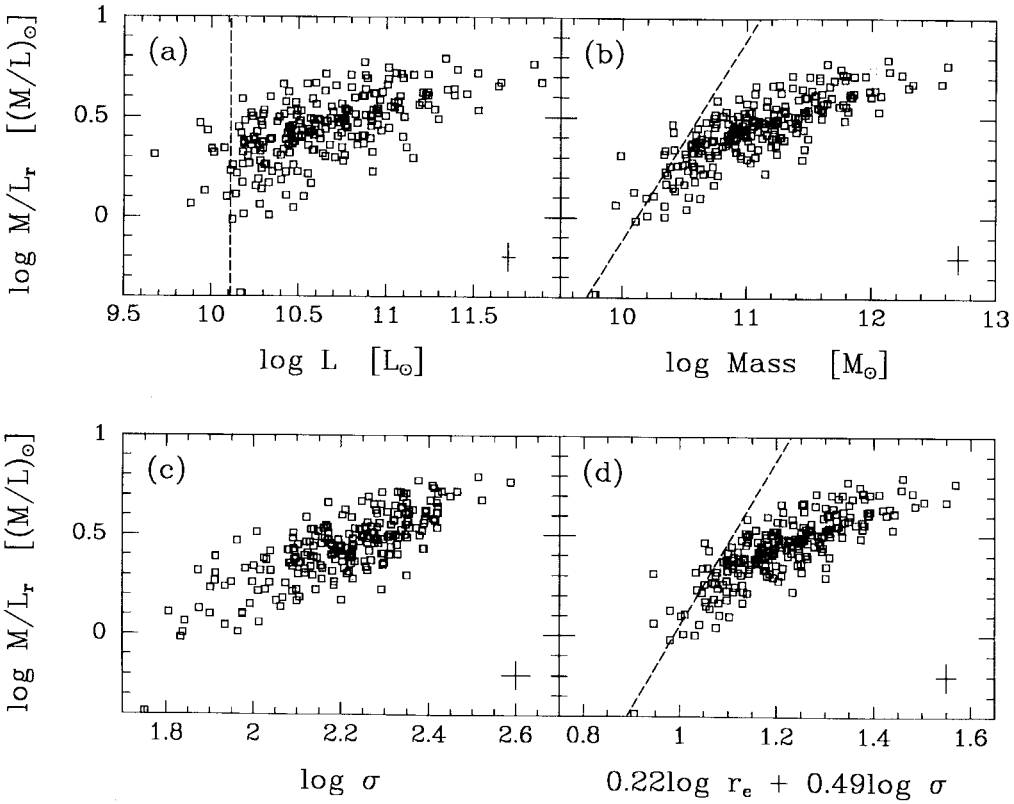


FIGURE 13 The M/L ratio as a function of the luminosity, the mass, the velocity dispersion and the combination $0.22\log r_c + 0.49\log\sigma$. $\log M/L = 2\log\sigma - \log\langle I\rangle_e - \log r_c +$ constant. r_c is in kpc; M/L, mass and luminosity in solar units. For $H_0 = 50$ km sec^{-1} Mpc^{-1} and Mass $= 5\sigma^2 r_c/G$ the constant is -0.73. The dashed lines in (a), (b) and (d) show the selection effect due to a limiting magnitude of -20.45 mag in Gunn r. This is the limit for the Coma cluster sample. (From the Royal Astronomical Society, **280,** 173, 1996.)

nearby spiral and irregular galaxies and even in higher redshift starburst galaxies. CO clouds are associated with regions of both massive and low mass star formation. In sprial galaxies, the CO emission often fills in the central H I minimum. There is an excellent correlation between the gas content of a galaxy and its current star formation rate as measured by integrated colors or the strength of emission from H II regions.

F. Radio Emission

All galaxies that are actively forming stars emit radio radiation. This emission is from a combination of free–free emission from hot, ionized gas in H II regions, and from supernova remnants produced when massive young stars reach the endpoint of their evolution. In addition, certain galaxies, usually ellipticals, have strong sources of nonthermal (i.e., not from stars, hot gas, or dust) emission in their nuclei. These galaxies, called radio galaxies, output the greatest part of their total emission at radio and far-infrared wavelengths. The radio emission is usually synchrotron radiation and is characterized by large polarizations and self-absorption at long wavelengths. A more detailed treatment of central energy sources can be found in another chapter.

G. X-Ray Emission

As noted earlier all galaxies emit X-ray radiation from their stellar components—X-ray binaries, stellar chromospheres, young supernova remnants, neutron stars, etc. More massive objects, particulary elliptical galaxies, have recently been found by Forman and Jones with the Einstein X-ray Observatory to have X-ray halos, probably of hot gas. A small class of the most massive elliptical galaxies which usually reside at the centers of rich clusters of galaxies also appear to be accreting gas from the surrounding galaxy cluster. This has been seen as cooler X-ray emission centered on the brightest cluster galaxy which sits in the middle of the hot cluster gas. This phenomenon is called a "cooling flow," and results when the hot cluster gas collapses on a central massive object and becomes dense

enough to cool efficiently. This process is evidenced by strong optical emission lines as well as radio emission. Cooling flows may be sites of low mass star formation at the centers of galaxy clusters.

Active galactic nuclei—Seyfert 1 and 2 galaxies (discoverd by C. Seyfert in 1943), and quasars are also usually strong X-ray emitters, although the majority are **not** strong radio sources. The X-ray emission in these galaxies is also nonthermal and is probably either direct synchrotron emission or synchrotron-self-Compton emission.

IV. GALAXY FORMATION AND EVOLUTION

A. Galaxy Formation

The problem of galaxy formation is one that remains as yet unsolved. The fundamental observation is that galaxies exist and take many forms. The interiors of most galaxies are many orders of magnitude more dense than the surrounding intergalactic medium.

In the hot Big Bang model, the simplest description of galaxy formation is the gravitational collapse of density fluctuations (perturbations) that are large enough to be bound after the matter in the universe recombines. At early times, the atoms in the universe—mostly hydrogen and helium—are still ionized so electron scattering is important and radiation pressure inhibits the growth or formation of perturbations. Before recombination, the universe is said to be in the *Radiation Era*, because of the dominance of radiation pressure. The period of recombination, that is, hydrogen and helium atoms are formed from protons, α particles, and electrons, is called the *Decoupling Era*, because the universe becomes essentially transparent to radiation. After that the universe is *Matter Dominated* as gravitational forces dominate the formation of structure. The cosmic microwave background radiation, postulated by Gamow and colleagues in the 1950s and detected by Penzias and Wilson in 1965, is the relic radiation field from the primeval fireball and represents a snapshot of the universe at decoupling.

In this simple picture, matter is distributed homogeneously and uniformly before decoupling because the radiation field will tend to smooth out perturbations in the matter. After decoupling statistical fluctuations will form in either the matter density or the velocity field (turbulence). The amplitude of fluctuations which are large enough will grow, and the fluctuations can fragment and, if gravitationally bound, collapse to form globular clusters, galaxies, or larger structures. A simple criterion for the growth of fluctuations in a gaseous medium was derived by Jeans in 1928. The Jeans wavelength, λ_J, is given by

$$\lambda_J = c_s \left(\frac{\pi}{G\rho} \right)^{1/2},$$

where c_s is the sound speed in the medium, and ρ is its density. The sound speed is

$$c_s \sim c/\sqrt{3}$$

in the radiation dominated era, and

$$c_s = \left(\frac{5kT}{3m_P} \right)^{1/2}$$

after recomination. If a fluctuation is larger than that in the Jeans length, gravitational forces can overcome internal pressure. The mass enclosed in such a perturbation, the "Jeans mass," is just

$$M_J \approx \rho \lambda_J^3 \approx \frac{c_s^3}{G^{3/2}\rho^{1/2}}.$$

Before recombination, this mass is a few times $10^{15}\ M_\odot$, which is comparable to the mass of a cluster of galaxies. After recombination, the Jeans mass plunges to $\approx 10^6\ M_\odot$, or about the mass of a globular star cluster. The amplitude $(\delta\rho/\rho)$ required for fluctuations to become gravitationally bound and collapse out of the expanding universe is dependent on the mean mass density of the universe. The ratio of the actual mass density to the density required to close the universe is usually denoted Ω.

$$\Omega = \frac{8\pi G\rho}{3H_o^2},$$

where H_o is the Hubble Constant. If Ω is large, i.e., near unity, small perturbations can collapse.

In the 1970s work on a variety of problems dealing with the existence and form of large-scale structures in the universe made it clear that the formation of galaxies and larger structures had to be considered together. Galaxies cluster on very large scales. Peebles and collaborators introduced the description of clustering in terms of low-order correlation functions. The two-point correlation function, $\xi(r)$, is defined in terms of

$$\delta P = N\,[1 + \xi(r)]\,\delta V,$$

where δP is the excess probability of finding a galaxy in volume δV, at radius r from a galaxy. N is the mean number density of galaxies. Current best measurements of the galaxy 2-point correlation function indicate that it can be approximated as a power law of form

$$\xi(r) = (r/r_o)^{-\gamma},$$

with and index and correlation length (amplitude) of

$$\gamma = 1.8, \quad \text{and} \quad r_o = 5h^{-1}\,\text{Mpc},$$

(where h is the Hubble Constant in units of 100 km sec^{-1} Mpc^{-1}). Power has been found in galaxy clustering on the largest scales observed to date ($\sim$100 Mpc), and the amplitude of clustering of clusters of galaxies is 5 to 10 times that of individual galaxies. In addition, work in the last decade has shown that there are large, almost empty regions of space, called *Voids*, and that most galaxies are usually found in large extended structures that appear filamentary or shell like, with the remainder in denser clusters.

There are currently two major theories for the formation of galaxies and large-scale structures in the universe. The oldest is the gravitational instability plus heirarchical clustering picture primarily championed by Peebles and coworkers. This is a "bottom-up" model where the smallest structures, galaxies, form first by gravitational collapse and then are clustered by gravity. This model has difficulty in explaining the largest structures we see, essentially because gravity does not have time in the age of the universe to significantly affect structure on very large scales. A more recent theory, often called the *pancake* theory, is based on the assumption that the initial perturbations grow adiabatically, as opposed to isothermally, so that smaller, galaxy-sized perturbations are intially damped. In this model the larger structures form first with galaxies fragmenting out later. If dissipation-less material is present, such as significant amounts of cold or hot dark matter (e.g., massive neutrinos), collapse will usually be in one direction first, which produces flattened or pancake-like systems. Models of this kind are somewhat better matches to the spatial observations of structure, but the hot dark matter models fail to produce the observed relative velocities of galaxies. The Hubble flow, or general expansion of the universe, is fairly cool, with galaxies outside of rich, collapsed clusters, moving only slowly ($\sigma < 350$ km sec^{-1}) with respect to the flow.

There are alternative models of galaxy formation, for example, the explosive hypothesis of Ostriker and Cowie. In this model, a generation of extremely massive, pregalactic stars is formed and goes supernova, producing spherical shocks that sweep and compress material soon after recombination. These shells then fragment and the fragments collapse to form galaxies. There is some recent evidence that favors this model.

The starting point for all of the aforementioned models are the observations of small-scale fluctuations in the microwave background radiation. Perturbations produced either before or after recombination appear as perturbations in the microwave background on scales of a few arc minutes. One of the major cosmological results of the 1990s was the discovery of these fluctuations at amplitides of one part in a hundred-thousand, $\Delta T/T < 10^{-5}$, with the COBE (Cosmic Background Explorer) satellite. More recently, higher spatial resolution observations with ground-based telescopes at the South Pole and with balloon-borne telescopes have measured the power spectrum of these fluctuations and appear to strongly confirm that the Universe is geometrically flat.

This is exactly as predicted by an amplification of the Big Bang model called *Inflation*, which was introduced in the early 1980s by A. Guth and later P. Steinhardt and A. Linde. In inflationary models, the dynamics of the early universe is dominated by processes described in th Grand Unified Theories (GUTS). In these models, the universe undergoes a period of tremendous inflation (expansion) at a time near 10^{-35} sec after the Big Bang. If inflation is correct, galaxy formation becomes easier because fluctuations in the very early universe are inflated to scales which are not damped out and can exist *before* decoupling. The problem of galaxy formation and the formation of large-scale structure is then the problem of following the growth of the perturbations we see in the early Universe until they become the objects we see today.

B. Population Evolution

The appearance of a galaxy changes with time (evolves) because its stellar population changes. This occurs because the characteristics of individual stars change with time. and because new stars are being formed in most galaxies. A galaxy's appearance thus reflects its integrated star formation history and the evolution of its gaseous content. The population evolution of galaxies is relatively well understood although detailed models exist for our galaxy and only a few others.

After the initial "formation" of the galaxy, the higher mass stars in the first generation evolve more rapidly than the lower mass stars. For example, the evolutionary timescale for a 100 $M_\odot$ star is only a few million years, while that for a 1 $M_\odot$ star is nearly 10 billion years. Elements heavier than hydrogen and helium are produced in the cores of these stars and are then ejected into the interstellar medium, either by stellar mass loss or supernova explosions, thus enriching the metal content of the gas. Stars formed later and later have increasing metallicities—the oldest and most metal poor stars in our own galaxy have heavy element abundances only 10^{-3} to 10^{-4} solar. The youngest stars have abundances a few times solar. It is possible for the average metallicity of a galaxy's gas to decrease with time if it accretes primordial low metallicity gas from the intergalactic medium faster than its high mass stars eject material.

Elliptical galaxies are objects in which most of the gas was turned into stars in the first few percent of the age of the universe. Only a few exceptional ellipticals—galaxies with cooling flows, for example—show any sign of current star formation. These are objects in which the initial star

formation episode was extremely efficient, with almost all of the galaxy's gas being turned into stars. At present, elliptical galaxies, probably get very slightly fainter and redder as a function of time. Their light is dominated by red giant stars that have just evolved off the main sequence, and the number of these stars is a slowly decreasing function of time for an initial mass function with the Salpeter slope. These stars are between 0.5 and 1 $M_\odot$. A competing process, main sequence brighteneing, can occur in systems where stars still on the main sequence contribute significantly to the light—that is, for systems with steep initial mass functions. Stars on the main sequence evolve up it slightly, becoming brighter and hotter, just before evolving into red giants.

Spiral and irregular galaxies have integrated star formation rates that are more nearly constant as a function of time. In these objects, the gas is used up slowly or replenished by infall. It is probable that in many of these galaxies, star formation is episodic, occurring in bursts caused by either passage of spiral density waves through dense regions, significant infall of additional gas, or interaction with another galaxy. The photometric properties of these galaxies are usually determined by the ratio of the amount of current star formation to the integrated star formation history. The optical light in objects with as little as 1% of their mass involved in recent, $<10^8$-year-old, star formation will be dominated by newly formed stars. Galaxies with constant star formation rates get brighter as a function of time for any reasonable assumption about the initial mass function.

It is common to model the evolutionary history of galaxies by assuming an unchanging initial mass function and parameterizing the star formation rate in terms of the available gas mass or density, or in terms of a simple functional form, usually and exponential. Examples of the possible color and luminosity evolution of an elliptical and a spiral galaxy are shown in Fig. 14. These models assume an exponentially decreasing star formation rate,

$$\Psi(t) \sim Ae^{-\beta t},$$

where β is the inverse of the decay time ($\beta = 0$ is a constant star formation rate). Properties of galaxies along the Hubble sequence are well approximated by a continuous distribution of exponentials, from constant star formation rates (Sd and Im galaxies) to initial bursts with little subsequent star formation (E and S0 galaxies).

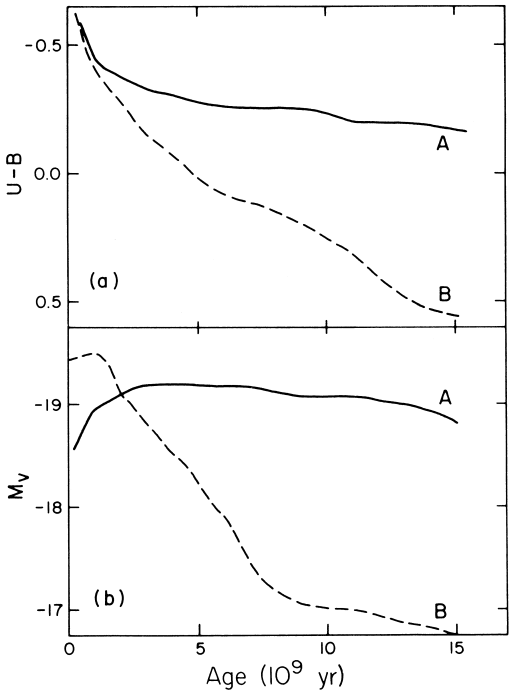


FIGURE 14 Luminosity and color evolution for two model galaxies. (a) The color U–B represents the logarithmic difference (in magnitudes) between two broad bandpasses approximately 800 Å wide centered on 3600 Å(U) and 4400 Å(B). More negative U–B's are bluer. (b) The variable M_V is the absolute visual magnitude. Model A has a star formation rate that is nearly constant as a function of time. This might be typical of a spiral galaxy. Model B has an exponentially decreasing star formation rate—15 *e*-folds in 15 billion years. This would be typical of an E or SO galaxy with a small amount of present day star formation.

C. Dynamical Evolution

There are several processes responsible for the dynamical evolution of galaxies, tidal encounters and collisions, mergers, and dynamical friction. If galaxies were uniformly distributed in space, the probability, P_i, that a galaxy would have undergone a close interaction or merger in time t is

$$P_i = \pi R^2 \langle v_{rel} \rangle N t,$$

where R is the size of a galaxy, $\langle v_{rel} \rangle$ is the mean relative velocity, and N is the number density of galaxies. For the average bright galaxy R is about 10 kpc (for a Hubble Constant of 100 km sec^{-1} Mpc^{-1}, $\langle v_{rel} \rangle$ is about 300 km sec^{-1} and P_i is thus less than 10^{-4} in a Hubble time. Galaxies are clustered, however, and this significantly increases the probability that any individual object has undergone an interaction. As stated earlier, on the order of 10%, all galaxies show some evidence of dynamical interaction.

Tidal encounters and collisions that produce observable results are still relatively rare events in the field. Some examples of such events were discussed earlier, e.g., the Whirlpool and ring galaxies. Detailed models of individual events have been made by Toomre and others and the models compare well with observed structures and velocity fields. Encounters at large relative velocity usually produce small effects because the interaction time is short

and the stellar components of galaxies can easily pass through one another. Encounters have large effects when the relative velocities are comparable or smaller than the internal velocity fields of the galaxies involved. The effects of encounters between spiral galaxies are also enhanced if the spin and orbital angular momenta of the galaxies are aligned. Fast collisions, however, may be possibly responsible for sweeping gas from early type galaxies in galaxy clusters, although other mechanisms (ablation by the hot intracluster medium, for example) probably dominate. Tidal encounters between protogalaxies were originally thought to produce most of the observed angular momentum in the universe; however, numerical simulations fail to produce the observed amount. The origin of angular momentum remains a mystery.

It was similary proposed that mergers of spiral galaxies might produce elliptical galaxies, a process which could explain the larger fraction of early type galaxies seen in rich clusters, where interactions are more likely to take place. Unfortunately, the photometric properties of elliptical galaxies as well as their relatively large globular cluster populations strongly argue against this hypothesis. As in tidal encounters, the "efficiency" of mergers depends on the relative encounter velocity; encounters at low relative velocities are much more likely to produce mergers than fast encounters. In particular, the line-of-sight velocity dispersion of galaxies in rich clusters is on the order of 1000 km sec^{-1}, which is much higher than the stellar dispersions internal to the individual galaxies.

Dynamical friction is a specialized case of "encounter" whereby a satellite object slowly loses its orbital energy when orbiting inside the halo of a more massive galaxy. An object moving in the halo of a galaxy produces a gravitational wake which itself can exert drag on the moving object. This effect was first described by Chandrasekhar in 1960. An object of mass M moving through a uniform halo of stars of volume density n with velocity v will suffer a drag of

$$\frac{dv}{dt} = -4\pi G^2 Mnv^{-2}\left[\phi(x) - x\phi'(x)\right] \ln \Lambda,$$

where ϕ is the error function, $x = 2^{-1/2}v/\sigma$, and Λ is the ratio of the maximum and minimum impact parameters considered. σ is the velocity dispersion of the stars in the halo.

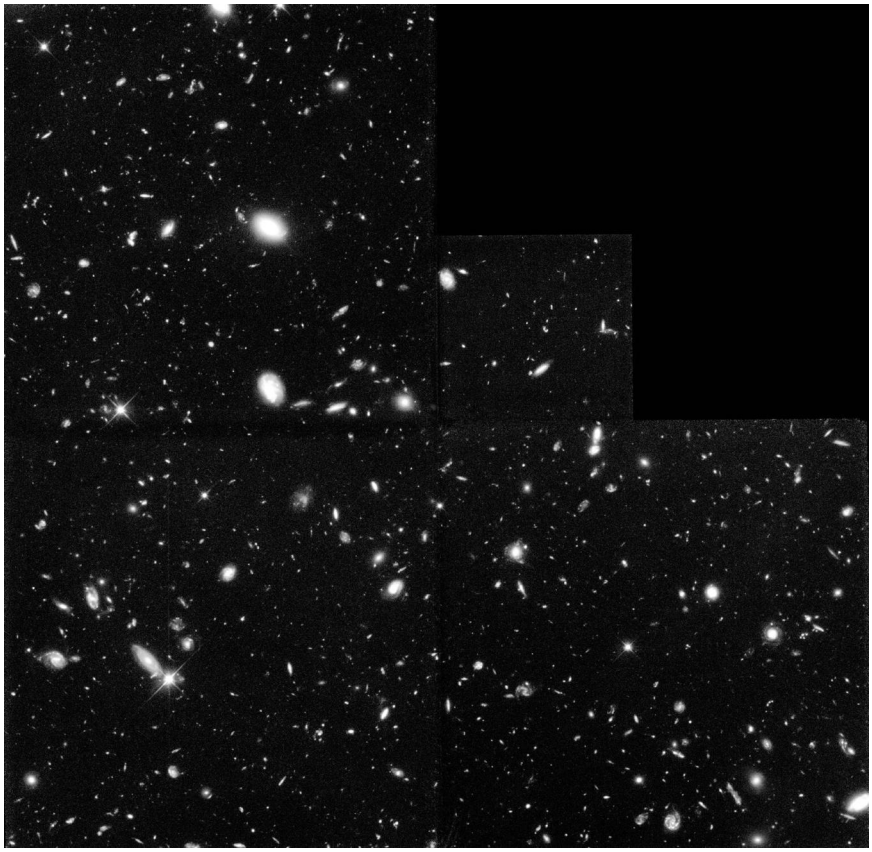


FIGURE 15

Mergers via dynamical friction as well as direct mergers of more massive galaxies almost certainly occur at the very centers of rich clusters of galaxies where the central cD galaxy is often accompanied by a host of satellite galaxies. Such processes probably account for the cD galaxy's extended halo, depressed central surface brightness, and excess luminosity relative to unperturbed bright ellipticals.

D. The High Redshift Universe

The launch of the Hubble Space Telescope in 1990 and the advent of a large number of new and powerful 8-m class ground-based optical telescopes has significantly improved our ability to see the Universe as it was a billion years or so after the Big Bang. Perhaps the best example of our view is the Hubble Deep field (Fig. 15). In this image, we see many galaxies at refshifts of 3 to 4, or as they would appear when the Universe is only a little older than a billion years. These objects are generally morphologically peculiar and show signs of both strong star formation and intense dynamical interaction. This and other observations indicate that the rate of formation of stars in the Universe probably peaked when the Universe was about 1/5th its present age and has declined significantly in the last 5–10 billion years. They also indicate that merging of galaxies and parts of galaxies is an important aspect of the evolutionalry scenario at redshifts greater than 1.

V. SUMMARY

As of this writing, no individual objects have been observed at redshifts greater than 6, so there is a large unexplored region of time and space between the time of recombination (the formation of the Cosmic microwave background at a redshift of ∼1000 or an age of a few 100,000 years) and the first observable objects. These "dark ages" will be explored with a new set of spaceborne telescopes such as SIRTF (the Space InfraRed Telescope Facility) and the NGST (Next generation Space Telescope).

SEE ALSO THE FOLLOWING ARTICLES

COSMIC INFLATION • COSMOLOGY • INTERSTELLAR MATTER • QUASARS • SOLAR PHYSICS • STAR CLUSTERS • STELLAR SPECTROSCOPY • STELLAR STRUCTURE AND EVOLUTION • SUPERNOVAE

BIBLIOGRAPHY

Bertin, G. (2000). "Dynamics of Galaxies," Cambridge University Press, Cambridge, UK.
Binggeli, B., and Buser, R. (eds.) (1995). "The Deep Universe," Springer, Berlin, Germany.
Binney, J., and Merrifield, M. (1998). "Galactic Astronomy," Princeton, Princeton, NJ.
Binney, J., and Tremaine, S. (1987). "Galactic Dynamics," Princeton, Princeton, NJ.
Bok, B. J., and Bok, P. (1974). "The Milky Way," Harvard University Press, Cambridge.
Bothun, G. (1998). "Modern Cosmological Observations and Problems," Taylor and Francis, London, UK.
Corwin, H., and Bottinelli, L. (eds.) (1989). "The World of Galaxies," Springer-Verlag, New York.
Elmegreen, D. M. (1998). "Galaxies and Galactic Structure," Prentice Hall, Upper Saddle River, NJ.
Ferris, T. (1982). "Galaxies," Stewart, Tabori and Chang, New York.
Hodge, P. (1986). "Galaxies and Cosmology," Harvard University Press, Cambridge.
Hodge, P. (ed.), (1984). "The Universe of Galaxies," Freeman and Co., San Francisco, CA.
Peebeles, P. J. E. (1971). "Physical Cosmology," Princeton University Press, Princeton NJ.
Sandage, A. (1961). "The Hubble Atlas of Galaxies," Carnegie Institution of Washington, Washington, DC.
Sandage, A., Sandage, M., and Kristian, J. (eds.) (1975). "Galaxies and the Universe," University of Chicago Press, Chicago.
Shu, F. (1982). "The Physical Universe," University Science Books, Mill Valley, CA.
Silk, J. (1989). "The Big Bang," Freeman and Co., San Francisco, CA.
Sparke, L. S., and Gallagher, J. S. (2000). "Galaxies in the Universe," Cambridge University Press, Cambridge, UK.

Interstellar Matter

Donald G. York
University of Chicago

GLOSSARY

Absorption Removal of energy by atoms, molecules, or solids from a beam of radiation, with re-emission at wavelengths other than the absorbing wavelength.
Cosmic rays Atomic nuclei or electrons accelerated to energies of more than 1 MeV by unknown processes.
Dissociation Breakup of a molecule into smaller molecules or atoms.
Emission Process whereby excited energy states in atoms and molecules produce radiation as electrons within the particles relax to lower energy states.
Galaxy Aggregate of stars (numbered 10^6–10^{12}) gravitationally bound in orbits typically 10 kpc in size.
Intercloud medium Space and material between the interstellar clouds, generally thought of as the largest volume in the interstellar medium.
Interstellar clouds Condensations in the interstellar medium of various densities and temperatures, typically several parsecs to several tens of parsecs in size.
Interstellar medium Space between the stars and the dust, gas, and fields that fill it.
Ionization Removal of electrons from atoms or molecules by any of several processes.
Nucleosynthesis Formation of heavy elements from lighter elements through fusion.
Parsec Measure of distance, equal to 3×10^{18} cm or about 3 light-years.
Polarization Preferential rather than random alignment of the electric vector of incoming radiation.
Recombination Capture of an electron by an atomic or molecular ion.
Scattering Redirection of light without changing its wavelength.
Shock front Pressure discontinuity within which low-energy particles are accelerated and ionized or dissociated.

AS FAR AS WE KNOW, galaxies can be regarded as the main building blocks of the universe. They are the central feature in modern astronomical research, all other fields being aimed at understanding where they came from and what goes on within them. The galaxies we can see

are made up of stars, but the largest part of the volume of the galaxies is filled with dust and gas—the most pure vacuum in nature, save only the space between the galaxies themselves. By mass, the interstellar material accounts for about 10% of our galaxy. The space between the stars is referred to as the interstellar medium. According to modern theories, the existence of a visible galaxy as an active, star-forming entity depends upon the passage of atoms made in stars out into this medium. The processes that occur therein are thought to lead to formation of new generations of stars.

The detailed process of removal of atoms from stars and their aggregation into condensing clouds that form new stars in interstellar space is complex. The elaboration of the process is a major thrust in modern astronomical research. We attempt here to describe, from an empirical point of view, what is known about the interstellar medium and the central role played by the gas and dust found there in the unfolding of the larger story of the life, birth, and death of galaxies. Following an overview, the article proceeds to describe the environment of interstellar space, the physical processes thought to be important, and the diagnostic techniques used in the field. The central ideas on the nature of the interstellar medium are then summarized. Near the end of the article, the particular details of the region near the sun are generalized to elaborate on areas of research on interstellar material in other galaxies and to discuss the cosmological importance of interstellar medium research.

I. THE INTERSTELLAR MEDIUM—AN OVERVIEW

A. Clouds

The interstellar medium in our galaxy contains four readily detectable components, three of which are described in terms of clouds of atoms and molecules. The brightest components, are the so-called HII regions, regions of ionized hydrogen that emit radiation when the protons (H^+) and electrons recombine. The regions are constantly reionized by radiation from nearby stars. They are readily visible near stars with temperatures above 30,000 K (O stars). The Great Nebula in Orion, around the multiple star θ Orionis, is the best-known example of the HII region because it is visible to the naked eye.

The second component consists of dark clouds that are very cold and emit no visible radiation, but do emit far-infrared (100-μm) and millimeter molecular line radiation. Absorption of background star light by dust in these clouds is responsible for the dark bands along the Milky Way in Scorpio and in Cygnus (the Cygnus rift) and for the famous dark spot in the southern sky known as the Coal Sack.

A third component, detectable only with sophisticated instrumentation, consists of diffuse clouds. These are lower in mass than either HII regions or dark clouds and are warmer than dark clouds but are not warm enough to emit significant radiation. They contain too little dust to cause discernible extinction to the human eye.

B. The Intercloud Medium

A fourth component is a very hot phase, with $T \sim 10^6$ K, thought to fill much of the void between the stars. This phase has been detected only indirectly, and its direct detection represents the most important frontier in the field. Gas with $T \sim 3 \times 10^5$ K (detected in the form of OVI or O^{+5}) is ubiquitous and is thought to imply the presence of the hotter 10^6 K material. Soft X rays detected from gas near the sun may arise in the 10^6 K gas. X rays from supernova remnants at 10^7 K are easily detectable. As these remnants expand and cool, they should produce large cavities at 10^6 K.

Figure 1 illustrates the location of interstellar gas of various types in a schematic spiral galaxy, seen from above.

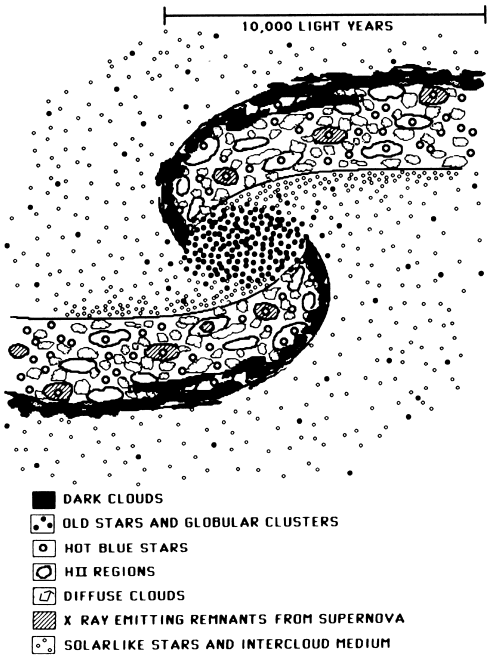


FIGURE 1 Relative location, within the spiral arms of a galaxy, of various types of interstellar material. The space between interstellar clouds is known as the intercloud medium. Solar-like stars permeate the spiral arms, but are omitted here for clarity.

II. PHYSICAL ENVIRONMENT IN INTERSTELLAR SPACE

A. Galactic Radiation Field

Interstellar gas is ionized by radiation from many sources. The dominant source of high-energy photons capable of ionizing hydrogen ($E > 13.6$ eV) is the O and B stars ($T > 10^4$ K). These stars primarily reside in spiral arms, but they produce a diffuse source of radiation with the spectral shape of a 25,000 K black body and an energy density of about 1 eV/cm^3 (equivalent to the radiation detected from a single B1 star at a distance of about 30 light-years). Since the hottest, most massive stars form in groups, there are local maxima in this radiation field, discernible by the copious emission of HII regions.

The only place unaffected by this pervasive field is the interior of dark clouds. Hydrogen atoms at the exterior of clouds shield interior atoms from the high-energy photons ($E > 13.6$ eV, the ionization potential of HI), but lower energy photons may penetrate. However, in the dark clouds, dust at the exterior provides a continuous opacity source, and so even visible light cannot penetrate.

Other sources of radiation include white dwarfs, hot cores of dying stars, normal stars like the sun, and distant galaxies and QSOs (the diffuse extragalactic background). The first two sources produce high-energy photons, but not in copious amounts. Normal stars are an important source of photons only at $\lambda > 4000$ Å.

While X-ray sources permeate space, the X rays themselves have little measurable effect on interstellar clouds. X rays probably serve to warm and ionize cloud edges and to produce traces of high-ionization material.

Supernovae from time to time provide radiation bursts of great intensity, but, except perhaps in the distant halo, these have little global effect. (The mechanical impact of the blast wave from the explosion, on the other hand, is very important, as discussed later.)

B. Magnetic Field

A weak magnetic field exists in space, though its scale length and degree of order are uncertain. The strength is roughly 10^{-6} G. The field was originally discerned when it was discovered that light from stars is linearly polarized ($\sim$1%). From star to star the polarization changes only slowly in strength and direction, suggesting that the polarizing source is not localized to the vicinity of the stars, but rather is interstellar in origin. It is thought that the field aligns small solid particles, which in turn lead to polarization of starlight. Since many elements are kept mostly ionized by the radiation field in space, the magnetic field constrains the motion of the gas (ions) and may be responsible for some leakage of gas from the galaxy.

C. Cosmic Rays

Very high energy particles (10^6–10^{21} eV) penetrate most of interstellar space. In the dark clouds, where ionizing photons do not penetrate, collisions between atoms and cosmic rays provide the only source of ions. Cosmic rays probably provide the basis of much of the chemistry in dark clouds, since, except for H_2, the molecules seen are the result of molecule–ion exchange reactions. Collisions between cosmic rays and atoms are thought to produce most of the observed lithium, beryllium, and boron through spallation. The energy density of cosmic rays is about 1 eV/cm^3, similar to that of stellar photons. Cosmic rays interact with atoms, leading to pion production. These decay to γ rays. The diffuse glow of γ rays throughout the galactic disk is accounted for by this process.

D. Cosmic Background Radiation

Light from other galaxies, from distant quasars, and even from the primordial radiation has a measurable impact on the interstellar medium. There are about 400 photons/cm^3 (or about 1 eV/cm^3, since $\lambda \sim 1$ mm) in our galaxy from the radiation bath of the Big Bang, now cooled to 2.71 K. At the remote parts of the galaxy, the integrated light of the extragalactic nebulae at 5000 Å, and of the distant quasars at $E > 13.6$ eV ($\lambda > 912$ Å) is comparable to galactic sources of such radiation.

E. Mechanical Energy Input

Stars, in the course of their evolution, propel particles into space, often at thousands of kilometers per second. These may come from small stellar flares; from massive dumps of the envelope of one star onto a binary companion, leading to an explosion (novae or supernovae); from winds driven by, among other things, radiation pressure, especially in the hottest stars; or from detonation of stellar interiors, leading to supernovae. These particle flows generate pressure on the interstellar clouds, and, in regions of massive stars and supernovae, substantially heat the gas and reshape the clouds. The random motion of interstellar clouds is largely the result of the constant stirring caused by supernovae.

III. PHYSICAL PROCESSES

A. Absorption and Emission of Radiation

Up to a point the physics of interstellar clouds is easy to ascertain. Atoms and simple molecules are easy to detect spectroscopically. Various effects lead to excitation or ionization of atoms and molecules. The extremely low

densities (10^{-3}–10^5 particles/cm^3) and the resulting great distances between atoms (millimeters to meters) means they seldom interact. When they do, the restoring radiative processes are usually fast enough that the atoms return spontaneously to their original energy states. The net result is quantized emission that is detectable at the earth. The strength and wavelength of the emission can be used to discern the density, temperature, abundance, velocity, and physical environment of the emitting region.

The same physics implies that most atoms and molecules are in their ground electronic state, with very few of the higher energy states populated by electrons. Hence, when light from a given star is intercepted by an atom, only a very restricted band of photons corresponds to an energy that the atom can absorb. In these few cases, the absorption is followed immediately by return of the electron, excited by the absorbed photon, to its original state, and subsequent re-emission of an identical photon. There is only a tiny chance that the re-emitted photon will have exactly the same path of travel as the one originally absorbed. To an observer using proper equipment, it appears that a photon has been removed from the beam. Since there are 10^{17}–10^{21} atoms in a 1-cm^2 column toward typical stars, the many repetitions of the absorption/re-emission process leave a notch in the spectrum called an "absorption line" because of the appearance of the features as seen in a spectrograph—dark lines across the bright continuous spectrum of a star. The absorption line strengths can be used to measure the physical conditions of the absorbing region.

B. Formation and Destruction of Molecules

In the densest regions (10–10^5 atoms/cm^3), molecules are known to exist in their lowest electronic states. The most abundant molecule, H_2, is thought to be formed on the surfaces of grains, small-sized (<1-μm) particles discussed later. Neutral hydrogen and molecular hydrogen are ionized by cosmic rays (see preceding discussion) as are other species such as neutral oxygen. A complicated chain of reactions (ion–molecular reactions) then ensues, since molecules and ions exchange electronic charge at rapid rates under the conditions present in interstellar clouds. Neutral molecules are then formed by recombination of ionized molecules and ambient electrons.

Some molecules are too abundant to be formed in the reaction scheme noted (such as ammonia, NH_3). Perhaps formation on grains is important in such cases. While the production of H_2 on grains can be generalized, production of other molecules cannot be predicted without specific knowledge of the makeup of the grains, which we now lack (see later discussion).

Some molecules may be formed in diffuse clouds by shock-triggered processes. Molecules are destroyed by chemical reactions (to make other molecules) or by radiation. In general, the higher the photodestruction cross section of a molecule, the more deeply it will be buried (in an equilibrium between destruction and formation processes) in the denser regions of clouds.

The most ubiquitous molecules are H_2 and CO (10^{-5}–0.5 molecules per H atom). Two- and three-atom molecules are common (e.g., CH, CN, HCN, and HCO^+, with 10^{-13} to 10^{-8} molecules per H atom). The most massive molecule now known is $HC_{13}N$. However, there is speculation that molecules as large as C_{60} may be present in large numbers. Because the nature of grains is poorly understood, it is possible that some observations that require small grains could be accounted for by large molecules (see later discussion).

C. Ionization and Recombination

Atoms and molecules in clouds are infrequently hit by photons or particles (atoms, molecules, or grains). When sufficient energy exists in the collision, an electron may be removed (ionization), producing a positively charged particle. Coulomb attraction between these particles and free electrons leads to recombination, reinstating the original state of charge. Recombination may be to an excited state of the neutral species. Subsequent decay of the electron to the ground state, through the quantized energy levels of the neutral atom or molecule, leads to emission of photons. The best-known examples are the Balmer lines of hydrogen seen in recombination from HII regions, the strongest of which is the Hα line at 6563 Å, which leads to the red appearance of some emission nebulae in color photographs. Sometimes the recombination process leads to population of atoms in "forbidden" states not formally permitted to be populated by absorption from the ground state because of quantum mechanical selection rules. These states are normally suppressed by collisional de-excitation in laboratory gases, but at the densities encountered in space ($<10^4$ ions/cm^3), radiative decay occurs before collisional de-excitation can occur. The best-known cases are the oxygen forbidden lines, particularly λ5007 of O^{+2}, which gives a greenish appearance to some nebulae.

D. Shock Fronts

Discontinuities in pressure occur in the interstellar medium, caused by explosions of stars, for instance. Propagation of such a pressure shock through the medium, into clouds in particular, leads to observable consequences. Shocks propagating at velocities above 300 km/sec randomly thermalize the gas impinged upon to temperatures above 10^6 K, where cooling by radiation of a gas with cosmic or solar abundance is inefficient. Such shocks are referred to as adiabatic and lead to growth of large cavities

of hot, ionized gas, transparent to radiation of most wavelengths. Eventually, particularly when a dense cloud is hit by a shock, the shock speed is reduced; hence it thermalizes gas to a lower temperature, and cooling is more important. The radiation that cools the gas is just the recombination radiation discussed earlier. Thus, if a supernova explosion occurs in a low-density region containing higher density (hence cooler) clouds, the propagating shock wave will lead to observable radiation from cooling just where the clouds are. Supernova remnants, such as Vela or the Cygnus Loop, are thus largely filamentary in visible light due to recombination radiation from cloud edges (e.g., Hα or λ5007; see previous subsection).

E. Dust Scattering

Several observation of stars indicate the presence of dust particles in space. Historically, the first and chief indication of their presence is that stars of the same temperature (based on stellar spectra) have different colors. In general, fainter, more distant stars are redder. Some clouds have so much dust they extinguish all background stars at $\lambda < 2\ \mu$m.

Models that try to account for the detailed features of the reddening suggest the existence of silicate grains and perhaps some graphite grains. Sizes range from 100 to 1000 Å (0.01–0.1 μm) with perhaps a core of silicates and a mantle or outer shell of amorphous ices. Such grains could produce extinction by direct scattering into directions away from the line of sight of the observer, by pure absorption in the bulk of the grains (with reradiation into longer wavelengths), and by interference of refracted and scattered radiation on the observer's side of the grains. Such grains, if irregular in shape, can be aligned by the interstellar magnetic field, leading to polarization of starlight.

F. Grain Formation and Destruction

Since the exact makeup of the grains is unknown, it is not clear how they form. Specific models for grain makeup allow one to pose questions as to how certain types of grains might form. For instance, cool stars are known to expand and contract periodically (over periods of months). At the maximum expansion, the outer envelopes reach temperatures below 1300 K. At the relevant densities, solid particles can condense out of the gas. The content of the grains depends on the composition of the stellar atmospheres, but the atmospheres seem varied enough to produce, in separate stars, silicate- and carbon-based cores. Such grains may be separated from the infalling, warming gas during the contraction of the atmosphere by radiation pressure, depending upon residual electric charge on the grains.

Subsequently, grains may grow directly in interstellar clouds. In this case, small-grain cores (or perhaps large molecules) must serve as seeds for growth of the remainder of the grain by adhesion of atoms, ions, and molecules with which the seed grains collide. The type of grains formed would then depend on the charge on the grains and the charge on various ions and on the solid-state properties of the grain surface and of monolayers that build up on it.

G. Heating and Cooling of the Gas

Cooling processes in interstellar gas can be directly observed because cooling is mainly through line radiation. Diffuse clouds are cooled mainly by radiation from upper fine structure levels of species such as C^+, which are excited by collisions of atomic or ionic species with electrons, neutral hydrogen, hydrogen ions, or H_2. In dense clouds, this process is important in atoms such as carbon. In hot, ionized regions (HII regions, $T \sim 10^4$ K) recombination radiation of H^0 and O^+ carries away energy from the gas, whereas at temperatures of 10^5 K (shocks in supernova remnants) recombination and subsequent radiation from forbidden levels of O^{+2} and other multiply ionized species dominates.

Heating of the interstellar gas is poorly understood. The heating sources observable (X rays, cosmic rays, exothermic molecular reactions) are inadequate to explain the directly observed cooling rates, given the temperatures observed in diffuse clouds ($T \sim 100$ K). The prime candidate for the primary heating mechanism is photoelectron emission from grains caused by normal starlight at $\lambda \lesssim 3000$ Å striking very small grains.

Heating of dense molecular clouds ($T \sim 10$ K) is also uncertain. Obvious gas-phase processes appear to be inadequate. An additional possibility in such clouds is that star formation activity (directly by infrared radiation or indirectly by winds or shocks caused by forming stars) heats the clouds. Localized shocked regions with $T \sim 25$ K have now been directly observed in dense molecular clouds through emission from rotationally and vibrationally excited H_2.

IV. DIAGNOSTIC TECHNIQUES

A. Neutral Hydrogen Emission

The spin of the nucleus (proton) of hydrogen and the electron can be parallel or antiparallel. This fact splits the ground state of hydrogen into two levels so close together that the populations are normally in equilibrium. In practice this means that the higher energy state is more frequently populated and that the resulting emission, in spite of its low probability, yields a detectable emission

at $\lambda = 21$ cm. In many cases, when the column density in each cloud (or velocity component) is not too high, the power detected at 21 cm is directly proportional to the number of hydrogen atoms on the line of sight of the main lobe of the radiation pattern of the radio telescope. The total number of atoms is generally expressed as a column density, N_{HI}, in units of atoms per square centimeter. The detection is not dependent on the *volume* density of the gas (denoted as n_{HI} atoms/cm^3) and yields no direct information on the distance of the emitting atoms: all atoms in the velocity range to which the receiver channel is sensitive are counted. Generally, the local expansion of the universe removes atoms not in our galaxy from the receiver frequency by Doppler shifting their 21-cm emission to longer wavelengths.

B. Molecular Emission

Molecules, although generally in the ground electronic state, are excited by collisional or radiative processes to higher rotational and vibration levels. Of the some 40 molecular species known in interstellar clouds, most are detected in the millimeter-wavelength region through rotational excitation. Hydrogen has been detected in rotational and vibrational emission and may be detectable in electronic emission (fluorescence).

If the mechanism populating the higher level of a transition is known, the emission strength gives some information about the mechanism. For instance, if emission from several different upper levels of a molecule is detectable, the relative population of the states can be determined. Given molecular constants and the temperature, the density of hydrogen can be determined if excitation is by collisional processes.

There are several unidentified emission features in the near infrared that are apparently related to interstellar material, some of which have been attributed to molecules in the solid phase or to polycyclic aromatic hydrocarbons. Identification of these features is being actively pursued in several laboratories.

C. Atomic Emission

Recombination lines and forbidden emission lines of atoms allow determination of abundances, densities, and temperatures in HII regions. Optical telescopes provide the main data for these lines. Eventually, fine structure emission in the infrared will give us detailed abundance data inside molecular clouds. Such data are currently available only for selected regions.

D. Atomic and Molecular Absorption

Absorption lines, already explained, are used to derive column densities of many species. Because of collisional de-excitation mechanisms, H_2CO is seen in absorption against the microwave cosmic background of only 3 K. Twenty-one-centimeter radiation is absorbed by H atoms in the lowest hyperfine state against background radio continuum sources. Resonance absorption lines of molecules such as C_2, H_2, CN, and CH^+ are seen in the optical and ultraviolet spectral regions. Resonance (ground-state) transitions of most of the first 30 elements in atomic or ionic form are seen as optical or ultraviolet absorption in the spectra of distant stars. In special circumstances, the degree of absorption is related linearly to the number of atoms leading to derived column densities (particles per square centimeter) to be compared with corresponding values of N for hydrogen. The ratio $N(\mathrm{X})/N(\mathrm{HI})$ gives the abundance of the species X, though in practice several ionization states of the species X must be accounted for. Conversely, some ions of heavy elements may be detected when hydrogen is ionized (HI unobservable), and the number of ionized hydrogen atoms must be accounted for (using, for instance, knowledge of the electron density and the intensity of the Balmer recombination radiation).

When ratios $N(\mathrm{X})/N(\mathrm{HI})$ are available for clouds, they are frequently referred to solar abundance ratios. If there are $1/10$ as many atoms of a certain kind per H atom as found in the sun, the element is said to be depleted by a factor of 10 in interstellar space.

E. X-Ray Emission and Absorption

X-ray absorption lines have not yet been detected from the hot gas mentioned earlier. While this is one of the most important measurements to be done in the study of interstellar material, it awaits the arrival of a new generation of very large X-ray telescopes. X-ray absorption edges have been detected, but these are contaminated by circumstellar absorption in the X-ray source itself, with little possibility of the velocity distinction possible in resonance absorption lines.

X-ray emission arises from radiative recombination and collisional excitation followed by radiative decay. Broadband X-ray emission from interstellar gas at 0.1- to 1-keV energies has been detected. High-quality spectra of the resolved emission lines from the hot gas are just becoming available. Such measurements are extremely important because the large hydrogen column densities at distances >100–200 pc absorb such soft X-ray photons. Any detection will thus refer only to very nearby gas, which can therefore be studied without confusion from more distant emission. Broadband X-ray emission at higher energies from the diffuse hot gas cannot yet be separated from a possible diffuse continuum from distant QSOs and Seyfert Galaxies.

Higher energy (>1-keV) X-ray lines and continuum have been detected from supernova remnants and from very hot ($T \sim 10^7$ K) gas falling into distant clusters of galaxies.

F. γ-Ray Emission

As noted earlier, cosmic rays interact with interstellar clouds to produce γ rays. A knowledge of the distribution of interstellar clouds and of the observed distribution of diffuse γ radiation may lead to a detailed knowledge of the distribution of cosmic rays in the galaxy. Our current knowledge is based largely on the cosmic rays detected directly at one point in the galaxy (earth).

V. PROPERTIES OF THE INTERSTELLAR CLOUDS

Given the many possibilities for detection of radiation emitted or modified by interstellar gas, astronomers have pieced together a picture of the interstellar medium. In many cases, the details of the physical processes are vague. In most cases, detailed three-dimensional models cannot be constructed or are very model dependent. On the other hand, in some cases, sufficient knowledge exists to learn about other areas of astrophysics from direct observations. The example of the cosmic-ray distribution has already been given. Others are mentioned subsequently.

Interstellar clouds are complex aggregates of gas at certain velocities, typically moving at ± 6 to 20 km/sec with respect to galactic rotation, itself $\sim$250 km/sec over most of the galaxy. Each cloud is a complex mixture of a volume of gas in a near pressure equilibrium and of isolated regions affected by transient pressure shocks or radiation pulses from star formation or from supernova explosions. Clouds are visible in optical, UV, or X-ray emission (or continuum scattering) when they happen to be close to hot stars and are otherwise detectable in absorption (molecules or atoms) or emission from low-lying excited levels (0.01 eV) or from thermal emission of the grains in the clouds.

A. Temperatures

Molecular clouds are as cold as 10 K. Diffuse clouds are typically 100 K. HII regions have $T \sim 8000$ K, depending on abundances of heavy elements that provide the cooling radiation. Low-column-density regions with $10{,}000 < T < 400{,}000$ K are seen directly, presumably the result of heating at the cloud edges from shocks, X rays, and thermal conduction. Isolated regions with $T > 10^6$ K are seen near sites of supernova explosions.

B. Densities

Densities, as determined from direct observation of excited states of atoms and molecules, are generally inversely proportional to temperature, implying the existence of a quasi-equilibrium state between the various phases of the medium. The effects of sources of disequilibrium in almost all cases last $\lesssim 10^7$ years, or less than $1/10$ of a galactic rotation time, itself $1/10$ of the age of the sun. The product $nT(\mathrm{cm}^{-3}\mathrm{K})$ is $\sim$3000 to within a factor of three where good measurements exist. Thus, the molecular (dark) clouds have $n > 10^2$ cm^{-3}, while in diffuse clouds, $n < 10^2$ cm^{-3}. Higher densities (up to 10^5 cm^{-3}) occur in disequilibrium situations such as star-forming regions inside dense clouds and in HII regions.

C. Abundances: Gas and Solid Phases

By measuring column densities of various elements with respect to hydrogen, making ionization corrections as necessary, abundances of elements in interstellar diffuse clouds can be determined. Normally, the abundances are compared with those determined in the sun.

Different degrees of depletion are found for different elements. Oxygen, nitrogen, carbon, magnesium, sulfur, argon, and zinc show less than a factor of two depletion. Silicon, aluminum, calcium, iron, nickel, manganese, and titanium show depletions of factors of 5–1000. Correlations of depletion with first ionization potential or with the condensation temperature (the temperature of a gas in thermal equilibrium at which gas-phase atoms condense into solid minerals), have been suggested, but none of these scenarios actually fits the data in detail.

The pattern of depletion suggests no connection with nucleosynthetic processes. Those elements that are depleted are presumed to be locked into solid material, called grains. Such particles are required by many other observations attributed to interstellar gas, as discussed earlier. In principle, the unknown makeup of the grains can be determined in detail by noting exactly what is missing in the gas phase. However, since there must be varying sizes and probably types of grains and since the most heavily depleted elements do not constitute enough mass to explain the total extinction per H atom, most of the grains by mass must be in carbon and/or oxygen. Establishing the exact mass of the grains amounts to measuring the depletions of C and O accurately. Although these measurements can now be made with the Hubble Space Telescope, problems of interpretation of deletions remain.

The grain structure (amorphous or crystalline) is not known. There are unidentified broad absorption features, called diffuse interstellar bands, that have been attributed to impurities in crystalline grains. However, these features

may be caused by large molecules. It has been argued that even if grains are formed as crystalline structures, bombardment by cosmic rays would lead to amorphous structures over the life of the galaxy.

Theories of grain formation are uncertain. A general scenario is that they are produced in expanding atmospheres of cool supergiants, perhaps in very small "seed" form. They may then acquire a surface layer, called a mantle, probably in the form of water ice and solid CH_4, NH_3, etc. This growth must occur in cold, dense clouds. The detailed process and the distribution of atoms between minerals and molecules in solid phase are unknown.

D. Evolution

Interstellar clouds can be large, up to 10^6 solar masses, and are often said to be the most massive entities in the galaxy. In this form, they may have a lifetime of more than 10^8 years. They are presumably dissipated as a result of pressure from stars formed within the clouds. Over the lifetime of the galaxy, interstellar clouds eventually turn into stars, the diffuse clouds being the residue from the star formation process. Growth of new molecular clouds from diffuse material is poorly understood. Various processes to compress the clouds have been suggested, including a spiral density wave and supernova blast waves. No one mechanism seems to dominate and several may be applicable. However, the existence of galaxies with up to 50% of their mass in gas and dust and of others with less than 1% of their mass in interstellar material leads to the inference that diffuse material and molecular clouds are eventually converted into stars.

VI. PROPERTIES OF THE INTERCLOUD MEDIUM

A. Temperatures

As already suggested, the medium between the clouds is at temperatures greater than 10^4 K. The detection of soft X rays from space indicates that temperatures of 10^6 K are common locally. Detection of ubiquitous OVI absorption suggests there are regions at $\sim 3 \times 10^5$ K. Various observations suggest widespread warm neutral and ionized hydrogen. Attempts to explain this 10^4 K gas as an apparent smooth distribution caused by large numbers of small clouds with halos have not been successful. Thus there is evidence for widespread intercloud material at a variety of temperatures, though large volumes of gas near 80,000 K are excluded. The more tenuous diffuse clouds, however formed, may be constantly converted to intercloud material through evaporation into a hotter medium.

B. Densities

In accordance with previous comments, all indications are that approximate pressure equilibrium applies in interstellar space. The above temperatures then imply intercloud densities of 10^{-1}–10^{-3} H atoms/cm^3. While direct density measurements in such regions are possible at $T < 30{,}000$ K, through studies of collisionally excited C^+ and N^+, direct determinations in hotter gas are spectroscopically difficult. X-ray emission has not yet been resolved into atomic lines and so is of limited diagnostic value. Direct measurements of ions from 10^4 to 10^7 K in absorption over known path lengths must be combined with emission line data to fully derive the filling factor, hence the density, at different temperatures. Since emission lines arise over long path lengths, velocities must be used to guarantee the identity of the absorption and emission lines thus observed. Such data will not be available for several years.

C. Abundances

The deplection patterns already noted in the discussion of clouds appear qualitatively in all measures of gas at 10^4 K, ionized or neutral. In general, the intercloud gas is less depleted. Perhaps shocks impinging on this diffuse medium lead to spallation of grains and return of some atoms to the gas phase. Data on hotter gas come mainly from absorption lines of OVI, CIV, and NV. Since ionization corrections are not directly determinable and since the total H^+ column densities are not known at the corresponding temperatures of 5×10^4 to 5×10^5 K, abundances are not available. However, the ratios C/O and N/O are the same as in the sun. It is not known whether elements such as iron, calcium, and aluminum are depleted in this hot gas. Optical parity-forbidden transitions of highly ionized iron or calcium may some day answer this important question.

D. Evolution

The evolution of the intercloud medium depends on the injection of ionization energy through supernova blast waves, UV photons, and stellar winds. A single supernova may keep a region of 100-pc diameter ionized for 10^6 years because of the small cooling rate of such hot, low-density gas. Ionizing photons from O stars in a region free of dense clouds may ionize a region as large as 30–100 pc in diameter for 10^6 years before all the stellar nuclear fuel is exhausted. Thus in star-forming regions of galaxies with low ambient densities and with supernova rates of 1 per 10^6 years per (100 pc)^3 and/or comparable rates of massive star formation, a nearly continuous string of overlapping regions of $\sim 10^4$–10^6 K can be maintained. When

lower rates of energy input prevail, intercloud regions will cool and coalesce, forming new clouds. In denser regions, comparable energy input may not be enough to ionize the clouds, except perhaps near the edges of the dense region, for periods as long as 10^8 years.

VII. THE INTERSTELLAR MEDIUM IN OTHER GALAXIES

While much is unknown about the actual balance of mechanisms that affect the distribution of gas temperatures and densities in our own galaxy, the facts that are known make interstellar medium observations in other galaxies an important way to determine properties of those galaxies as a whole. A few examples are mentioned here.

A. Supernova Rates

Supernovae are difficult to find because the visible supernovae occur only once per 30–300 years. Regions with higher rates are often shrouded in dust clouds or extinguished by nearby dust clouds. Thus, little is known about supernova rates and their global effects on galaxies and on the orgin of elements. Nucleosynthesis in massive stars and in the supernova explosions, with subsequent distribution to the diffuse interstellar medium, may have occurred at variable rates, perhaps much more frequently in the early stages of galaxy formation than now.

Studies of interstellar media in other galaxies can shed light on these subjects. X-ray emission and atomic line emission can reveal the presence of supernova remnants, which, since they last up to 10^5 years or more, are easier to find than individual supernovas, which last only a few months. Absorption line measurements and emission line measurements can be used to determine abundances in other galaxies. Absorption line measurements reveal the velocity spread of interstellar gas, the stirring effect caused by supernovae integrated over 10^6–10^7 years. (Quasi-stellar objects, clumps of O stars, or the rare supernova can be used for background sources in such absorption line studies.) By making the above-noted interstellar medium studies on samples of galaxies at different redshifts, the history of galaxy formation can be discerned over a time interval of roughly three-fourths the expansion age of the universe.

B. Cosmic-Ray Fluxes

The origin of cosmic rays is unknown. However, they account for a large amount of the total energy of the galaxy. *In situ* measurements are only possible near earth. However, cosmic rays provide the only explanation of the observed abundances of boron, beryllium, and lithium ($\sim 10^{-9}$–10^{-10} atoms/H atom). Thus the abundance of any of these elements is a function of the integrated cosmic-ray flux over the lifetime of the galaxy. Variations in the ratio [B/H] would imply different cosmic-ray fluxes. Comparison of [B/H] with other parameters related to galaxy history (mass, radius, total Hα flux, and interstellar cloud dispersions) may indicate the history of cosmic rays.

C. History of Element and Grain Formation

As is clear from previous sections, studies of various kinds of clouds in our own galaxy have not led to a clear empirical picture of how the interstellar medium changes with time. The trigger or triggers for star formation are poorly understood, and the history of the clouds themselves is unknown in an empirical sense. Numerous problems exist on the theoretical side as well.

Study of interstellar media in other galaxies should be very important in changing this situation. Absorption and emission measurements that provide data on individual clouds and on ensembles of clouds should allow classifications of the gas phase in galaxies that can be compared with other classifications of galaxies by shape and total luminosity. Key questions include: Do elements form gradually over time, or are they created in bursts at the beginning of the life of the galaxy? Do earlier galaxies have the same extinction per hydrogen atom as is found in our own galaxy, or are differences seen perhaps because grains require long growth times to become large enough to produce extinction at optical wavelengths? Do the many unidentified interstellar features (optical absorption, IR emission and absorption) occur at all epochs, or are they more or less present at earlier times? When did the first stars form? How does star formation depend on the abundance of heavy elements?

D. Cosmological Implications

For various reasons the interstellar medium has proved to be fruitful ground for determining cosmological quantities. Without detailed comparison with other techniques, a few examples are given.

The light elements hydrogen and helium are thought mainly to be of nonstellar origin. They are currently thought to be formed in the Big Bang. Expansion of the early cosmic fireball leads to a small interval of time when the gas has the correct temperatures for fusion of hydrogen to deuterium and of deuterium to helium (^{3}He and ^{4}He). The helium reactions are so rapid that deuterium remains as a trace element, the abundance of which is dependent on the density of matter at the time of nucleosynthesis. By knowing the expansion rate of the universe, the derived density can be extrapolated to current densities.

Comparison with the total light emitted by stars in galaxies suggests that there are 10 times fewer protons and neutrons than implied by the abundance of deuterium. The search for the "missing" baryons is now focused on material in hot gas (10^{50}–10^{70} K), already noted in this article as being hard to detect.

The amount of helium present is similarly important in deriving the properties of matter at the time of nucleosynthesis. In principle, the helium abundance today provides a test of fundamental particle physics because it depends on the number of neutrino types (currently thought to be three) and on the validity of general relativity.

Deuterium abundances are currently best determined in UV absorption line experiments in local interstellar matter. Helium abundances are best determined by measurements of optical recombination lines of He^{2+} and He^{+} from HII regions in dwarf galaxies with low metal abundance.

The preceding comments suggest how interstellar medium studies allow a view of the very earliest stages of the formation of the universe through measurements of relic abundances of deuterium and hydrogen. Earlier comments related the importance of such studies to observing and understanding the formation and evolution of galaxies through studies of abundances and star formation rates. A third example is the study of clustering of galaxies at different redshifts. Since galaxies are very dim at cosmological distances and since they may not have central peaks in their light distributions, they are difficult to detect at $z > 2$. Even at $z = 1$, only the most luminous galaxies are detectable. However, interstellar medium observations of absorption lines depend not on the brightness of the galaxy, but on the brightness of an uncorrelated, more distant object, say a QSO. Thus, galaxies at very high redshift can be studied. Many QSOs side by side will pass through a number of galaxies in the foreground, and, with adequate sampling, the clustering of absorption lines reflects the clustering of galaxies on the line of sight.

Depending on the total matter density of the universe, the galaxies at high redshift will be clustered to a comparable or lesser degree than they are today. Thus, changes in clustering between high-z galaxies (interstellar media) and low-z galaxies (direct images and redshifts) reflect the mean density of the universe. The material measured in this way includes any particles with mass, whereas the density measurement discussed earlier, using deuterium, counts only neutrons and protons. The difference between the mass density determined from changes in clustering and the density determined from deuterium give a measure of the amount of "dark" matter in the universe, that is, matter not made of protons and neutrons. Neither estimate of the mean density (one from deuterium, the other from clustering) explains the apparent density needed to account for the geometry of space. Either some form of matter not clustered with galaxies, or equivalently, a nonzero energy of the vacuum of space may explain the discrepancy. Current estimates are that 0.5% of the mass-energy is in visible stars, 5% is in other forms of matter made of protons and neutrons, 25% is in matter not made of protons and neutrons, and 70% is in vacuum energy.

SEE ALSO THE FOLLOWING ARTICLES

COSMIC RADIATION • COSMOLOGY • GALACTIC STRUCTURE AND EVOLUTION • INFRARED ASTRONOMY • STAR CLUSTERS • STELLAR SPECTROSCOPY • STELLAR STRUCTURE AND EVOLUTION • SUPERNOVAE • ULTRAVIOLET SPACE ASTRONOMY

BIBLIOGRAPHY

Bally, J. (1986). "Interstellar molecular clouds," *Science*, **232,** 185.

Boesgaard, A. M., and Steigman, G. (1985). "Big Bang nucleosynthesis: Theories and observations," *Annu. Rev. Astron. Astrophys*. **23,** 319.

Frisch, P. F. (2000). "The galactic environment of the sun," *Am. Sci*. **88,** 52.

Krauss, L. M. (1999). "Cosmological antigravity," *Sci. Am*. **280**(1), 52.

Mathis, J. S. (1990). "Interstellar dust and extinction," *Annu. Rev. Astron. Astrophys*. **28,** 37.

Savage, B. D., and Sembach, K. (1996). "Interstellar abundances from the absorption line observations with the Hubble Space Telescope," *Annu. Rev. Astron. Astrophys*. **34,** 279.

Shields, G. A. (1990). "Extragalactic HII regions," *Annu. Rev. Astron. Astrophys*. **28,** 525.

Spitzer, L. (1990). "Theories of hot interstellar gas," *Annu. Rev. Astron. Astrophys*. **28,** 37.

Neutron Stars

Steven N. Shore
Indiana University, South Bend

GLOSSARY

Glitch Discontinuous change in the rotational frequency of a pulsar.

Millisecond pulsars Weakly magnetized neutron stars, either isolated or in binary systems in which there is no mass transfer between components, which have periods near the rotational stability limit.

Pulsar Radio pulsars are isolated or nonaccreting binary neutron stars with strong ($>10^9$ G) oblique magnetic fields. X-ray pulsars are strongly magnetized neutron stars imbedded in accretion disks, which are fed from mass-losing companions in close binary systems.

Units Solar mass ($M_\odot$), 2×10^{33} g; solar radius ($R_\odot$), 7×10^{10} cm.

URCA process Neutrino energy loss mechanism whereby electron capture on nuclei produces neutrinos, followed by subsequent beta decay of the resultant nucleus with an additional neutrino emitted. The name derives from the Casino de Urca, at which gamblers lost their money little by little but without seeming to be aware of it. The process was first described by Gamow and Schoenberg.

NEUTRON STARS are the product of supernovae (SN), gravitational collapse events in which the core of a massive star reaches nuclear densities and stabilizes against further collapse. They occur in a limited mass range, from about 0.5 to 4 $M_\odot$, being the product of stars between about 5 and 30 $M_\odot$. If they are rapidly rotating and magnetized, their radio emission can be observed as pulsars. Some occur in binary systems, where mass accretion from a companion produces X-ray and gamma-ray emission over a

wide range of accretion rates. If the rate is in the proper range, periodic nuclear reactions can be initiated, which create episodic mass ejection events called bursters. Magnetic neutron stars in binary systems produce pulsed X-ray emission without strong radio emission. Neutron stars are unifications of nuclear and relativistic physics, requiring gravitation theory to explain their structure and serving as useful laboratories for constraining the nuclear equation of state.

I. HISTORICAL AND CONCEPTUAL INTRODUCTION

The first work on dense equations of state, following the advent of quantum theory in the 1920s, showed that it is possible for the pressure of a gas to depend solely on the density. This implies that it is possible to construct stable cold configurations of matter, independent of the previous thermal history of the matter. The first application of this idea to stellar evolution was by Landau who, in a prescient 1931 paper, discussed the possibility of the end state of evolution being composed of a sphere in which the pressure support is entirely provided by electrons in this extreme state of matter. Referred to as degenerate, the condition results from the fact that, in a quantum gas, the electrons are constrained to have only one particle per "box" of phase space. Landau pointed out that it might be possible for the configuration to depend only on the mass of the particle, which determines the momentum at a given energy, and thus to have objects in which the particle is neutral and heavy, but Fermi-like.

Chandrasekhar, in the early 1930s, derived an upper limit to the mass of such a configuration, above which the pressure exerted by the electrons is always insufficient to overcome the pull of gravity. The ultimate fate of a more massive object would be collapse. While he did not follow the evolution of such a state, it was clear from this work that the upper mass for such stars, called white dwarfs, is of order the mass of the sun. He argued that more massive stars cannot end their lives in hydrostatic equilibrium but must gravitationally collapse.

The discovery of the neutron and Yukawa's meson theory clarified the basic properties of nuclear matter. Application of quantum statistics demonstrated that the neutron is a fermion, a spin of $\frac{1}{2}$. It is consequently subject to the Pauli exclusion principle, like the electron, but since its mass is considerably heavier than the electron, by about a factor of 2000, degenerate neutron configurations are considerably denser than white dwarfs (WD). If a neutron sphere (NS) were formed, it would a sufficiently compact at a given mass that the Newtonian calculations previously used for stellar structure would not suffice for its description. Following the work of Tolman, who in 1939 derived the interior structure metric for a homogeneous nonrotating mass, Oppenheimer and Volkoff, in the same year, derived the equation for relativistic hydrostatic equilibrium. They showed that there is a maximum mass above which no stable configuration for a neutron star is possible for a degenerate neutron gas. While Chandrasekhar had presented a limiting mass, which is an asymptotic upper limit for infinite central density, Oppenheimer and Volkoff showed that the masses above about 1 $M_\odot$ must collapse. A subsequent paper by Oppenheimer and Snyder showed that the ultimate fate of more massive objects was to rapidly collapse past their Schwarzschild radii and become completely relativistic, singular objects, later called black holes (BH) by Wheeler. By the end of the 1930s, it was understood that the end state of stellar evolution was either a stable degenerate configuration of electrons (WD), neutrons (NS), or catastrophic gravitational collapse (BH). This picture has not been significantly altered with time.

Following these calculations, Zwicky and Baade nominated supernovae as the sites for neutron star formation. Having first shown that there is a class of explosive stellar events with energies significantly above the nova outburst, by about three orders of magnitude, which they dubbed "supernovae," they showed that the magnitude of energy observed could be provided by the formation of a neutron star. They argued that collapse of the core of the star from about the radius of the sun to the order of 10 km would be sufficient to eject the envelope of the star (by a process at the time unknown) with the right magnitude of energy observed in the SN events. The only object that could thus be formed is a neutron star.

The first detailed calculations for the structure of such objects awaited the advent of computers in the early 1960s and the improvement in the equations of state for nuclear matter. The fact that degenerate equations give an upper limit of about 0.7 $M_\odot$ is in part an artifact of the softness of the equation of state at lower than nuclear densities. However, from the energy level structure in nuclei, it is possible to specify more exact forms for the nucleon–nucleon (N–N) interaction potential, which can be used to calculate the equation of state. The nuclear, or strong, force depends on the exchange of integer spin particles, π mesons. It is thus possible to have an attractive potential at large separations, which becomes progressively more repulsive at short distances. The central densities, which result in such models, are somewhat higher near the maximum mass, but they do not change the fact that there is a maximum in the stable mass of such a star. The inclusion of exotic particles as the outcome of N–N scattering changes the details of the equation of state at the highest densities, but is still incapable of removing this maximum entirely.

Neutron stars remained theoretical entities until November 28, 1967, when, using a radio telescope designed for the study of scintillation of radio sources at a wavelength of 3.7 m, Bell and Hewish discovered the first pulsar CP $1919+21$. They observed a rapidly varying, periodic, pulsed, point radio source with pulse frequency of about 1 Hz. In rapid succession, several more of these objects were discovered. By the end of the year, Gold had suggested that a pulsar is a rapidly rotating magnetized neutron star. The pulses, he argued, result from the magnetic field being inclined to the rotation axis; it is from the magnetic polar regions that the signal is radiated. The short duration of the pulses compared with the interpulse period supports this view. Pacini showed that such stars must spin down via the emission of low-frequency rotationally generated electromagnetic waves based on the power requirements for the synchrotron luminosity of the Crab Nebula, and such a spindown was indeed observed by Radakrishnan in the Vela pulsar in 1969. The polarization and frequency characteristics of the radio pulses left no doubt as to their nonthermal origin, and implied the presence of extremely strong magnetic fields, of order 10^{12} G (the field of the earth is about 1 G). The formation of such strong fields had previously been suggested by Ginzberg in 1963 as the consequence of flux freezing during stellar gravitational collapse.

The association of PSR $0532+27$ with the Crab Nebula, the remnant of the supernova of 1054, and its optical identification with a faint blue star of featureless continuum and pulsed optical output near the center of symmetry of the nebula, showed that the initial guess for the origin of neutron stars is likely correct. However, with the exception of the Vela pulsar and one or two others (the matter is still in doubt), pulsars are generally not found imbedded in supernova remnants.

The discovery of binary X-ray sources came with the determination of the period of Cen X-3, in the early 1970s following the launch of the UHURU satellite. The spectrum of several other sources, and the mass determinations, notably Her X-1 = Hz Her, have been used for the determination of both the magnetic and mass properties of neutron stars. These sources are regular X-ray pulsars, formed by the structuring of the mass accretion region into a column that corotates with the magnetic field of the star and has a period of order a few seconds. The observation of cyclotron lines in the X-ray from these stars shows that the magnetic fields are $>10^{10}$ G as expected for neutron stars. None has a mass above $4\,M_\odot$ regardless of the mass of the companion.

The first binary pulsar, PSR $1913+16$, was discovered by Manchester and Taylor in 1974. With its superb timing stability relative to the errors associated with velocity determinations for mass-accreting X-ray pulsars, it helped to set, for the first time, accurate mass limits for neutron stars. Only upper limits are possible, given uncertainties in the mass of the companion. About a half dozen more of these have since been found. The discovery of PSR $1937+214$, a pulsar with a period of about 1 msec, in 1982 was soon followed by several others, including several members of binary systems. These pulsars have comparatively weak magnetic fields of order 10^9 G, and appear to result from the spin-up of neutron stars by accretion during phases of rapid mass transfer in a close binary system.

II. FORMATION OF NEUTRON STARS

Neutron stars are formed by gravitational collapse of intermediate mass stars. Nucleosynthesis in such stars can proceed exothermally to build elements up to, but not past, ^{56}Fe because of the progressive increase in the binding energy of nuclei until this point. Following Fe synthesis, the primary loss mechanisms for the stellar core become electron captures onto seed nuclei via the URCA process:

$$\mathrm{e} + (Z, A) \rightarrow (Z-1, A) + \nu_\mathrm{e}, \tag{1}$$

where Z is the atomic number and A the atomic mass, followed by the subsequent decay:

$$(Z-1, A) \rightarrow (Z, A) + \mathrm{e} + \bar{\nu}_\mathrm{e}, \tag{2}$$

which represents a bulk rather than surface cooling. Because the neutrino opacity of normal stellar envelopes is small, these particles can freely escape from the core in rapid order, producing rapid cooling that is very density and temperature sensitive. The core is thus forced to contract, increasing the reaction rates and producing accelerating losses. Finally, with increasing *neutronization*, the core reaches a critical density at which the free-fall time becomes short compared with the sound speed, and the subsequent gravitational collapse cannot be avoided.

The free protons, which result from the destruction of nuclei as the density passes $\approx 10^{11}$ g cm^{-3} and electron capture that converts them to neutrons, alter the equation of state. Should the core mass be small enough, less than the maximum mass allowable for a neutron star, the collapse will halt for the core, and the subsequent release of gravitational potential energy, which is essentially the gravitational binding energy of the final configuration (the collapse takes the core from about 10^6 km to about 10 km) amounts to $\approx 10^{53}$ erg available for the lifting of the envelope.

Most of this energy is radiated away in a neutrino burst that accompanies the formation of the stable core. Neutrino emission from the newly formed neutron star reaccelerates the shock, which initially stalls, and powers the

expulsion of the stellar envelope. The visible supernova event that is the product of the collapse contains but a small portion of the energy liberated by the core formation. While the details depend on the evolutionary history of the progenator star, and the visibility of the shock depends as much on the environment of the collapsed star as the energetic processes powering the ejection, it is generally thought that the supernova event that heralds the end of the collapse should be visible. The neutrino burst detected from SN 1987A in the Large Magellanic Cloud, with about the right order of magnitude and spectrum, as well as duration (a few seconds), provides strong support for the general picture as outlined here.

There are two critical physical problems connected with the formation of the stable configuration that will become a neutron star: how does the equation of state halt the collapse of the core; and how does the final configuration look subsequent to the ejection of the envelope? To answer the first, one must examine the equations of state, since it is the products of the nucleon–nucleon interaction. Then, one must consider the conditions for equilibrium, both mechanical and thermal, and what constraints they provide on the interior structure.

III. EQUATIONS OF STATE: NEUTRON STARS AS NUCLEI

Nuclear physics meets the cosmos nowhere as dramatically as in the context of the interior structure of neutron stars. Since these stars are in essence nuclei, albeit several kilometers across, delimiting their density profiles is an excellent check on our understanding of the structure of nuclear matter.

The simplest equation of state, the relation between pressure, density, and temperature, comes from application of the Pauli exclusion principle to neutrons. Since neutrons are spin-$\frac{1}{2}$ particles, their phase space is strictly limited to single occupation. To show how a simple equation of state can be derived, consider a gas of quantum mechanical, chargeless particles that interact only via the Pauli principle. That is, the number of particles in a volume of phase space, h^3, where h is Planck's constant, is limited by

$$p_{\mathrm{F}}^3 n^{-1} = h^3, \tag{3}$$

which defines the *fermi momentum*, p_{F}, for nucleon density n. The distance between the particles is essentially the de Broglie wavelength. These particles exert a mutual pressure, resulting from the quantum condition, so that $P \approx n(p_{\mathrm{F}}^2/2m) \approx h^2 n^{5/3}/m$. Notice that this is independent of temperature, a condition which is called *degenerate*. The pressure comes primarily from the unavailability of phase space "bins" into which particles can be placed. The degeneracy of a gas is measured by

$$\alpha = \frac{p_{\mathrm{F}}^2}{2mk_{\mathrm{B}}T}, \tag{4}$$

where k_{B} is the Boltzmann constant. T is the temperature, and m is the mass of the particle in question. If α is large, the quantum effects dominate the available phase space volume for the gas and the temperature plays no role in determining the pressure. The pressure is really the average of this over the total distribution function for

$$P \approx \int_0^{E_{\mathrm{F}}} \frac{E^{3/2}\,dE}{\exp[(E/kT) - \alpha] + 1}. \tag{5}$$

Therefore, the pressure is strongly dependent on the density, and not temperature, unless the total thermal energy significantly exceeds the Fermi energy.

White dwarfs and neutron stars are similar, in this picture, because both are composed of spin-$\frac{1}{2}$ particles. Their differences arise chiefly from the mass of the supporting particles, electrons being the primary supporters for white dwarfs. There is, however, a more profound difference between these two stars—neutrons, being baryons, interact with each other via the strong force. Thus, the details of nuclear potentials play critical roles in the equilibrium of such extreme states of matter.

In order to remain bound in nuclei, neutrons must interact attractively for some separations; the stability of matter, however, demands that the core turn repulsive (or at worst have vanishing potential) at very small distances. The finite sizes of nuclei, of order a few fermi (about 10^{-13} cm), comes from the large mass of the π meson, the particle principally responsible for the transmission of the strong force. Degeneracy alone yields a *soft* equation of state, that is, one which asymptotically vanishes at infinite density. The maximum mass such a gas can support against collapse must be much smaller than one which is due to a strongly repulsive interaction at small separation (high density limit). The star that results from a soft pressure law will consequently be smaller in radius and have a higher central density than for a strongly repulsive short-range potential, called a *hard* equation of state. However, it is only once the neutrons and protons become free that such considerations must be included in structure calculations. Below the density at which this occurs, the nuclei are formed into an ionic lattice, through which a degenerate electron gas flows. This portion of the neutron star, the crust, is a solid ranging from densities of about 10^8 to 10^{11} g cm^{-3}. The electrons at these very high densities are not only degenerate but form pairs called BCS superconductors.

The neutrons are not bound in nuclei in neutron star cores. That is because the free energy is greater for the

bound than the "unbound" state, and the neutrons simply diffuse, or "drip" out of the nuclei. This occurs well above the neutronization density, at about 4×10^{11} g cm^{-3}. Its significance for neutron star structure is that until this density is reached, the system can be thought of as a normal ionic gas, or indeed solid, but at higher densities no such state is possible. Pairing becomes important in the reactions of the nucleons, again at densities of the order of the neutron drip state, and a superfluid can result. This involves the formation of N–N pairs, with their subsequent behavior as a bosonic system. The analogous situation occurs much earlier in density for the electrons, which form pairs as in a superconductor and also behave as a superfluid within the ionic lattice that can be set up from the nuclei.

This superfluid is one of the most important components in the interior of the neutron star, not because of its effect on pressure so much as its other dynamical properties. For instance, the heat capacity of a superfluid is essentially zero. The nucleons are completely degenerate and therefore do not have large thermal agitation. It has the property of rotating as a quantum liquid, which means certain properties such as the vorticity are quantized, and therefore exercises a strong dynamical effect on the coupling between the outer and core regions of the neutron star. In effect, the superfluid represents a collective dynamical mode for the neutrons, similar to that observed in nuclei when many-particle states are included in the calculation of nuclear structure. On the role it plays in the rotational history of the star, we shall say more later.

The core region dominates the structure of the neutron star. It is here that the stiffness of the nuclear matter equation of state is most strongly felt. Since the density is about 1–3 particles fm^{-3}, about 10^{15} g cm^{-3}, it is essential that the small distance part of the nuclear potential be properly evaluated. The pressure is given by the gradient, with separation, of the N–N potential; in fact, the maximum mass of the star is essentially determined by the gradient in the N–N potential at distances less than 1 fm. The steeper the gradient in the potential, the stiffer the equation of state; at the same pressure, a lower density is required. The star can be more distended and have a higher moment of inertia, a less centrally condensed structure, a lower density throughout the envelope, and a higher maximum mass than for a softer potential.

The interior structure can, in fact, be observationally studied. The rotational properties of a neutron star depend on its equation of state through its moment of inertia. If the magnetic field can be determined independently, the spin-down rate for pulsars gives an estimate of this quantity. The maximum mass is dependent on the N–N potential. Various models provide substantially different results. While the softest equations of state give a lower limit of about 0.7–1 $M_\odot$ for the maximum mass, the stiffest give 4–5 $M_\odot$.

IV. EQUATIONS OF STRUCTURE

In order to go from the equation of state to the structure of a neutron star, it is essential that general relativity play a role. Electrons, being lower mass than neutrons, produce degenerate configurations that are large, with radii about the size of the earth (about 0.01 $R_\odot$). Nucleons, however, are about 2000 times more massive, so that for degeneracy to be reached and for the N–N potential to play a part in their support, the final configurations must be very much more compact. Neutron stars are so close to the stability limit for a compact configuration and their gravitational binding energies are so close to its rest mass, that relativistic rather than Newtonian equations are required to describe their structure.

To begin with, the mass equation must take into account the fact that the rest mass and the energy density, P, are both involved with the total density. The metric, for simplicity, can be assumed to be that of an isolated, nonrotating (or slowly rotating) point mass outside of the star, which has an increasing mass with the distance from the center of the star (the so-called Schwarzschild interior solution). Take the mass to be scaled in terms of *geometrical* units, so that $G = c = 1$. Then the equation for the mass interior to some point is given by integration on spherical shells:

$$\frac{dm}{dr} = 4\pi r^2 \rho. \tag{6}$$

The pressure gradient is given by the reaction of the matter to compression, which depends on the equation of state. Thus,

$$\frac{dP}{dr} = -\frac{(\rho + P)(m + 4\pi P r^3)}{r^2(1 - 2m/r)}. \tag{7}$$

The gravitational potential, actually the term in the metric that provides the redshift, is given by

$$\frac{d\phi}{dr} - \frac{1}{\rho + P}\frac{dP}{dr}. \tag{8}$$

This last term is the closest link to the actual metric, since $ds^2 = -e^{2\phi}dt^2 + 1/[1 - (2m/r)]dr^2$. For a point mass, the vacuum field has a singularity at $r = 2m$, which in physical units is $r_* = 2\ GM/c^2$, the Schwarzschild radius.

For a star with a surface gravity of 10^{14} cm sec^{-2}, for instance, the gravitational redshift is about 0.2. This redshift reduces the intensity of the radiation seen by a distant observer in a given energy band by a factor of $(1 + z)^3$ and affects the temperature determined from the flux. Thus, in all of the calculations of neutron star properties, it should

be borne in mind that the observational quantities are those detected at infinity, that is by a distant observer, but that the interior values are calculated in the *rest frame* of the neutron star.

These equations, originally based on Tolman's interior solution for the metric of a relativistic nonrotating mass, form the basis of the Oppenheimer–Volkov solution. As one moves progressively farther from the stellar core, the increasing interior mass causes an increasing curvature of space time, which in turn affects the dependence of the pressure on the distance. The gravitational mass, the amount of mass that is felt by the overlying matter outside of a radius r, increases more slowly than would be expected in the Newtonian model, and the pressure therefore changes appropriately.

The equation of state is inserted into these equations, for which the baryon density n is the independent "coordinate," with the initial conditions that the density at the center, ρ_c, is a free variable, with $M(0)=0$. The mass of the configuration is determined by the boundary condition that $\rho(R)=0$ and $P(R)=0$, while $M(R)=M$. Most of the mass of the star is taken up in binding energy, which becomes progressively larger as the mass is increased. Such behavior is already understood from white dwarf models, since degenerate equations of state behave similarly. As mass is added to the star, the star shrinks. The limiting mass is when the radius becomes small enough for a black hole to develop, past which no stable configuration is possible. This is why neutron stars lead necessarily to collapse if they are made massive enough. There is no stable state between them and catastrophe should they begin to collapse.

The maximum mass is a product of general relativity. While for a Newtonian star an infinite central density is asymptotically approached leading to a progressively smaller star (which becomes asymptotically small), there is a lower limit to the size of a relativistic object, the Schwarzschild radius. Degenerate matter produces an approximate mass–radius relation $MR^3 = \text{constant}$, so that for $R = r_*$ there is a fixed mass. Above this mass, a black hole is the inevitable result.

The upper mass limit for neutron stars can be set by the observation that, for at least the X-ray pulsars, the mass of the degenerate object must exceed that of a white dwarf and is always lower than 4 or $5\,M_\odot$. For Cyg X-1 and LMC X-3, the masses appear inconsistent with this value and likely represent cases of black holes as the sources of the X-ray emission.

V. PULSARS

The first observations of radio pulsars served to confirm the existence of neutron stars—over 400 are now known (Table I). The detailed modeling of these objects has centered on the mechanisms that are capable of producing the emission. It is clear that they are rapidly rotating collapsed stars, with periods as short as a few milliseconds. In order to discuss how observations of pulsars help to elucidate the interior properties of neutron stars, it is useful to look at some of the observational constraints they can provide.

First, the fact that they shine means that pulsars possess strong magnetic fields; that they pulse is the result of their rotation and the fact that the magnetic field is not axisymmetric. A rotating, magnetized neutron star loses rotational energy by emission of magnetic dipole radiation:

$$\frac{dE_{\rm rot}}{dt} = \frac{2\mu^2}{3}\,\omega^4 \sin^2\beta, \tag{9}$$

where μ is the magnetic moment ($\mu = B_0 R_0^3$, R_0 is the stellar radius, and B_0 is the surface magnetic field) and β is the angle between the magnetic moment and the rotational axis. The change in the rotation period, $P_{\rm not}$, is given by

$$P_{\rm rot}\dot{P}_{\rm rot} \sim \frac{B^2 R_0^6}{I}, \tag{10}$$

where I is the moment of inertia. Secular variations of the magnetic field, such as decay due to the finite conductivity of the interior matter, produce a change in the deceleration rate. For instance, if the field is exponentially decreasing with time,

$$\dot{\omega} \sim B_0^2 e^{-2t/\tau}\omega^3, \tag{11}$$

where τ is the decay time for the field. Should the alignment angle of the field also depend on time, the pulse profile and the rate of spindown will also change. Unfortunately, the time scales implied by the models are very long, and while these provide some probes of the electromagnetic properties of neutron star interiors, they are not well understood presently.

In a few other ways, pulsars provide us with windows into the interiors of neutron stars, most dramatically in the form of glitches. These will be discussed later in more detail. Glitches are discontinuous changes in the rate of rotation of the star, by $\Delta\omega/\omega \approx 10^{-7}$–$10^{-6}$ superimposed on the secular spindown of isolated pulsars. In addition, however, there is *timing noise*, which is a fluctuation in the rotation rate of the star manifested by changes in the arrival time of pulses not connected with the properties of the interstellar medium. This may be caused by small thermal effects, within the stellar superfluid, which produce random walking between pinning sites of vortices.

Pulsar emission may also reveal properties of neutron star crust, although now only within the context of specific models. The Ruderman–Sutherland model uses the fact that as a magnetic star rotates, it generates a latitude-dependent electric field in the surrounding space.

TABLE I Radio Pulsar Mass Summary

Star	Median mass ($M_\odot$)	68% Central limits	95% Central limits	Notes[a]
Double neutron star binaries				
J1518 + 4904:				
Pulsar	1.56	+0.13/−0.44	+0.20/−1.20	GR, RO
Companion	1.05	+0.45/−0.11	+1.21/−0.14	GR, RO
Average	1.31	±0.035	±0.07	GR
B1534 + 12:				
Pulsar	1.339	±0.003	±0.006	GR
Companion	1.339	±0.003	±0.006	GR
B1913 + 16:				
Pulsar	1.4411	±0.00035	±0.0007	GR
Companion	1.3874	±0.00035	±0.0007	GR
B2127 + 11C:				
Pulsar	1.349	±0.040	±0.080	GR
Companion	1.363	±0.040	±0.080	GR
Average	1.3561	±0.0003	±0.0006	GR
B2303 + 46:				
Pulsar	1.30	+0.13/−0.46	+0.18/−1.08	GR, RO
Companion	1.34	+0.47/−0.13	+1.08/−0.15	GR, RO
Average	1.32	±0.025	±0.05	GR
Neutron star/white dwarf binaries				
J0437 − 4715	—	—	<1.51	$\dot{x}$, $P_b m_2$
J1012 + 5307	1.7	±0.5	±1.0	Opt
J1045 − 4509	—	—	<1.48	$P_b m_2$
J1713 + 0747	1.45	±0.31	±0.62	$P_b m_2$, GR
	1.34	±0.20	±0.40	$P_b m_2$, GR, Opt
B1802 − 07	—	<1.39	<1.45	GR
	1.26	+0.08/−0.17	+0.15/−0.67	GR, $P_b m_2$
J1804 − 2718	—	—	<1.73	$P_b m_2$
B1855 + 09	1.41	±0.10	±0.20	GR
J2019 + 2425	—	—	<1.68	$P_b m_2$
Neutron star/main—sequence binaries				
J0045 − 7319	1.58	±0.34	±0.68	Opt, MS

[a] Assumptions made in mass estimate: (GR) general relativistic binary model; (RO) random orbital orientation, or inclination angle uniform in $\cos i$; (P_bm$_2$) core mass orbital period relation; ($\dot{x}$) propermotion-induced change in the projected semimajor axis; (Opt) optical companion observations; (MS) main-sequence stellar model.

Crust electrons experience an acceleration, initially flowing freely away from the pulsar in the magnetic polar cone, but being trapped closer to the magnetic equator. The pulsar environment fills with charge, which alters the electric field at the stellar surface. Electrons are being ripped out of the crust faster than they can be resupplied from charge migration within the star so that a potential drop builds up over the polar cap. When the potential drop at the poles exceeds that of the rest mass energy for pair creation, $2m_e c^2$, positron–electron pairs, which produce cascades at the polar caps, are created. These, in turn, are responsible for the observed emission. The emission is confined to the polar region, and it is therefore the obliquity of the field that is responsible for the visibility of the pulsar.

There is no evidence for any thermal emission from any known pulsar that is not in a mass-accreting binary system. The neutron star cooling is sufficiently fast and uniform over the stellar surface that there is no evidence

for pulsed thermal X-rays. Instead, any X-ray emission can be attributed either to small synchrotron nebulas in the vicinity of the neutron star or the nonthermal high-energy emission from the polar cascade.

The spindown rate of a pulsar, all other things being equal, is a measure of the age of the star. Provided the magnetic field does not change with time, the age can be estimated from II/II. Usually, this gives ages of about 10^3–10^6 years. There is one class of pulsar, however, for which this estimate is completely unreliable, the milli-second pulsars. For these, the rate of spindown is extremely long, over 10^8 years, although their periods are incredibly small. The answer is to be found, in part, from the very low values of their magnetic fields and in part is due to their rotation being the artifact of their histories. These neutron stars have been, or still are, members of close binary systems in which accretion has torqued the neutron star up to rotational frequencies near the limit of stability. We shall return to this point in Section VIII.

VI. GLITCHES: EVIDENCE FOR SUPERFLUIDS IN NEUTRON STAR INTERIORS

At densities in excess of the neutron drip density, the free nucleons feel an attractive potential. This favors the formation of pairs, which behave like bosons and form a superfluid. Pairing is favored at densities between 10^{11} and 10^{14} g cm^{-3}. This fluid feels no viscosity, except by collisions with the normal fluid, which serves to drag on the particles and couple them to the rotation of the star as a whole. A rotating superfluid quantizes vorticity. These vortices have the property of pinning to irregularities in the crustal substructure by threading through nuclei. Should there be sufficient thermal agitation of the vorticies, they will creep from one nuclear site to another.

The surface is continually slowing down by the emission of low-frequency (the rotation frequency of the star) electromagnetic waves causing the angular velocity gradient to become large across the superfluid mantle. This in turn generates a shear which will produce a drift of the pinning sites. Major depinning events result in glitches; the glitching rate depends on the temperature of the normal fluid and serves as a probe of the internal temperature of the neutron star mantle. Postglitch relaxation of the rotational frequency depends critically on scattering properties in the superfluid for the phonons generated by the event. In effect, the star rings for a while as it settles down to a new rotationally stable state. The detailed behavior of this system has yet to be completely understood, although phenomenological models have been successful in predicting that the glitch behavior should be a property of only young, that is, still relatively hot, neutron stars.

The reader might think this contradicts the fact that the equation of state for the system is degenerate. However, that does not mean that there are no random motions for the particles; it simply implies that the equation of state does not have a temperature dependence. In addition, neutrons in the superfluid are no longer constrained by Fermi statistics, and so have a different behavior than the electrons or free nucleons.

VII. COOLING PROCESSES

Neutron star matter is completely optically thick to photons. Therefore, only the surfaces can cool by the emission of such radiation. If this were the only mode for the decay in the temperature of such objects, one would expect them to be easily detected by X-ray observations for quite a long time after their formation, of order 10^6 yr, since their formation surface temperatures are in excess of 10^{10} K. Other processes than photon emission, however, are far more important in the cooling of such objects, specifically those associated with the emission of neutrinos. For very dense matter, the cross sections for the production of electron and muon neutrinos by collision processes among nucleons are high. The URCA processes, in which neutrinos are the sole emitted particles, have been shown to be important. Specifically, processes of the form

$$\mathrm{n} + \mathrm{n} \rightarrow \mathrm{n} + \mathrm{p} + \mathrm{e} + \nu_\mathrm{e} \tag{12}$$

and

$$\mathrm{n} + \mathrm{p} + \mathrm{e} \rightarrow \mathrm{n} + \mathrm{n} + \nu_\mathrm{e} \tag{13}$$

which are modified URCA processes in which the neutron bystander absorbs some of the momentum necessary for the emission, have been shown to be important. Bremsstrahlung processes, in which the exit channels are essentially the same as the incoming particles with the addition of the radiation of the excess momentum acquired during the collision, can be of the form:

$$\mathrm{n} + \mathrm{p} \rightarrow \mathrm{n} + \mathrm{p} + \nu_\mathrm{e} + \bar{\nu}_\mathrm{e} \tag{14}$$

$$\mathrm{n} + \mathrm{n} \rightarrow \mathrm{n} + \mathrm{n} + \nu_\mathrm{e} + \bar{\nu}_\mathrm{e}. \tag{15}$$

The URCA processes may occur in the crust, where entire nuclei are present:

$$\mathrm{e} + (Z, A) \rightarrow \mathrm{e} + (Z, A) + \nu_\mathrm{e} + \bar{\nu}_\mathrm{e}. \tag{16}$$

One of the earliest suggested processes, by Bahcall and Wolf, is

$$\pi^- + \mathrm{n} \rightarrow \mathrm{n} + \mathrm{e} + \bar{\nu}_\mathrm{e} \tag{17}$$

a weak interaction, like the nn or pn interactions, which for high enough interaction energy is replaced by the (μ, ν_μ)

pair. While the rates for π^- and URCA processes have modest temperature dependence of T^6, the nn and np bremsstrahlung processes are more temperature sensitive, varying like T^8.

Superfluids play a role in altering the rates of the bremsstrahlung processes. This is because of pairing of the baryons, which introduces a gap energy for the free particles so that the rates are suppressed by a factor of $\exp(-\Delta_j/kT)$ where Δ_j is the BCS gap for either protons or neutrons ($j = \mathrm{p}$, n, respectively). Since this gap energy depends on the density, it alters the predicted rates of cooling in a way that can probe the interior structure of the star.

It must be kept in mind that, since neutron stars are relativistic objects, the gravitational metric plays a role in the cooling. Neutrinos and photons escaping from the interior suffer a redshift simply because of the gravitational field. This reduces their luminosity at the stellar surface and affects the cooling rate. Their observed surface temperature and luminosity are related to their "actual" values by $T_s e^{-\phi}$ and $L_s e^{2\phi}$, respectively. Because the structure of the star is virtually independent of temperature, however, the redshift can be determined from the interior model and then included in the calculation of the luminosity after the fact.

No young pulsars, which are isolated neutron stars or binary neutron stars that are not accreting matter from companions, show observable thermal X-ray emission, including the Crab pulsar. The cooling time estimates from all current calculations indicate that the stars should be detectable for at least 10^3–10^4 years, yet the Crab pulsar is less than 1000-year old and is still unobservable in soft X-rays; its inferred surface temperature is less than 2×10^6 K. This is at odds with most of the current theories for neutrino formation, and the properties of the cooling function for neutron stars serve as a useful test of ideas in neutrino and intermediate energy nuclear physics. One possible explanation is that rapid cooling takes place via charged pion condensation because of the rapid increase in the neutrino emission. This occurs at densities of order 3×10^{14} g cm^{-3}.

Recent calculations have included the effects of an atmosphere, although the calculations of the opacities at the probable densities and temperatures associated with neutron stars in the cooling phase are difficult to obtain. The gravities are of order 10^{16} cm^2 sec^{-1}, in comparison with those of a normal star (about 10^4) and a white dwarf (10^8). These show, however, that most of the emission is typical of a blackbody, and that flux redistribution because of the opacity edges at ionization limits alters the shape of the flux profile but cannot explain the lack of detection of thermal emission from young pulsars. Magnetic fields can act to alter the cooling rates as well, but this tends to increase the visibility of the star by suppressing many of the cooling and conduction processes. The apparently rapid cooling associated with neutron stars remains a serious problem.

VIII. NEUTRON STARS IN BINARY SYSTEMS

The number of pulsars known to be in binary systems is steadily growing, including several of the most rapidly rotating objects known—the millisecond pulsars. The presence of neutron stars in binary systems was signaled by the discovery of Cen X-3 and Her X-1, both of which display rapid X-ray pulsation. Accreting material from a companion, which is losing mass because of tidal interactions with the neutron star, is funneled into a column at the magnetic poles. The accretion columns are tied to the stellar surface and corotate with the star. The subsequent discoveries of the gamma burst sources, the rapid burster, and the X-ray burst sources have added to the picture of binaries.

If the neutron star does not possess a strong magnetic field, matter accreted from a companion will pile up on the surface until it undergoes nuclear reactions. Since these occur with underlying degenerate matter, the heat conductivity away from the site is very poor. The consequence is a thermonuclear runaway, resulting in the matter blowing off the surface and producing a characteristic burst profile. The maximum luminosity of the burst is at the $_1$Eddington limit, the luminosity at which the radiative acceleration exceeds that of gravity. The burst signature is a very rapid rise to maximum light (of order 10 msec) and a slower (<1 sec) decay time. Many of the burst sources have been observed to flare on numerous occasions. Some show quasiperiodic variations in the X-ray, as seen by EXOSAT and RXTE. These presumably come from accretion onto the magnetosphere of rapidly rotating neutron stars, with the development of a two-stream and Rayleigh–Taylor instability at the boundary layer. This results in the occasional excitation of X-ray variations as the matter rains onto the stellar surface, mediated by the magnetic field.

Since neutron star formation requires a supernova event, one might think it unlikely that the binary could survive such a cataclysm. Yet clearly many have. This leads to the suggestion that the amount of mass lost from the system must be small compared with the masses of the stars remaining, and that the events must systematically be like type I SN events. This implies that the collapsed star had begun as either a white dwarf or helium star that was induced to collapse, perhaps by the excess accretion of matter. No such events have yet been observed, so the model remains untested except by statistics on the binaries. Many are circular orbits (a few notable exceptions exist, such as Cir X-1), and this requires efficient tidal dissipation on

time scales of 10^6 years to circularize the orbit. The binary X-ray sources have been found in globular clusters, which also contain burst sources, so that the formation of neutron stars is clearly not limited only to the young population of the galaxy. The sources observed for globular clusters may be the result of tidal capture of the neutron star.

Finally, X-ray novae and burst sources are analogs of classical novae, except the mass gainer is a neutron star. Matter accumulates from a tidally distorted close companion and ignites when the pressure reaches the flash point for hydrogen burning, reaching temperatures of 10^9 K or higher. Little mass is required, about $10^{-11}\ M_\odot$, for less than the accretion necessary to bigger an explosion on a white dwarf because of the enormous gravity. Nuclear processing during the burst can bypass iron, extending as far as $M_\odot$ or beyond.

IX. MAGNETIC FIELD GENERATION AND DECAY

While there are some strongly magnetic main sequence stars with fields of several kilogauss, most normal stars show upper limits of about 100 G. If the flux of such a star is somehow conserved in the collapse process, the resultant field would scale as R_*^{-2}, so that the expected field of a neutron star should be of order 10^{12} G. While the details of the magnetic field freezing process have yet to be understood, the fact that this is the right order of magnitude for many neutron star fields is very interesting and indicates that some form of flux freezing must play a role in the stars.

Direct detection of the fields is accomplished through cyclotron emission by electrons in accreting systems. The measured fields, $\sim 10^{12}$ G agree with the strengths inferred for isolated radio pulsars.

A central problem for neutron star studies is the generation of the wide range observed for pulsar magnetic fields. The millisecond pulsars are all objects that have intrinsically low fields, of order 10^9 G or so, and have been spun up by accretion during their earlier histories. It is important to note that the subsequent decay of the rotation is dependent on the magnetic field strength and configuration, so that the weaker the field, the longer the time over which the star will be a rapid rotator.

A model that generates strong magnetic fields after the collapse uses the currents generated in the crust due to the thermal gradients present immediately after the supernova event and while the neutron star is cooling. It is assumed that a small magnetic field (by pulsar standards) may be present initially after the collapse. Since in a solid, heat is transported by electrons moving within the crystal lattice, a crustal temperature gradient produces a net electron current, the *Nernst effect*. The field saturates when the ohmic dissipation overwhelms the amplifications at about the value observed for pulsars, about 10^{12} G. The fields reach peak strength within the crust, not at the surface, where they become essentially quantum limited, at about 10^{14} G. The problem is that this is a local, not global, effect, and it is not clear how the field manages to organize the large-scale dipolar structure thus far observed for these systems. Further, there is the problem of the obliquity. No obvious alternative models have yet been proposed. What one knows is that many neutron stars do possess strong fields, the weakest of which almost coincide in strength with those of the strongest white dwarf stars, and that many of these fields are present even in stars that have been around for a while. The decay times for the fields appear to be long, but here again there are not sufficient data to judge the matter.

The fields of many neutron stars are highly inclined to the rotation axis. Some of this is due to the discovery procedure—one looks for periodically variable emission lines, pulsed X-ray continuum, or pulsed radio emission, so one systematically discovers the highly oblique fields. Highly oblique fields are known to occur in main sequence stars and in white dwarfs and have even recently been discovered for planets (e.g., Uranus), so they may not be a special feature of magnetic field formation in neutron stars. It remains, however, an unexplained phenomenon for pulsars.

Several important departures from normal matter result from the strength of the surface magnetic fields. One is that the electrons in the degenerate gas at the surface quantize in their orbits about the magnetic field lines. The result is that each of the particles, depending on its momentum (and recall that since there is complete degeneracy in the electron gas this ensures that there will be only one particle per magnetic level, or Landau orbital), occupies an energy level. Thus, the electron gas has a strongly polarized response and also highly anisotropic properties. There are several important consequences of this. For one, heat conductivity is strongly direction dependent. Even with the gas being degenerate, there will still be far more efficient heat transfer parallel than perpendicular to the magnetic field lines. This in turn affects the cooling of the neutron star and alters the details of the temperature gradient. In addition, there is also a strong perturbation felt by the electrons still tied to the ions in the crust. This perturbation causes the electron clouds to align with the magnetic field, producing one-dimensional crystals as a dominant crustal structure. There is also an effect on the coupling of the

crust and core via the current represented by the electron gas in the mantle of the neutron star.

X. FUTURE PROSPECTS

Several major problems clearly remain, some of which cut to the core of our understanding of the formation and subsequent evolution of neutron stars. No extragalactic supernovas have yet been observed to produce pulsars, and, with the exception of the Crab and Vela pulsars and RCW 103, supernova remnants do not appear to have associated pulsars. Additionally, no point X-ray sources are associated with Cas A (which exploded in the 1670s as determined by the dynamics of the ejecta and the γ-ray detection of ^{44}Ti decay by COMPTEL) or the remnants of SN 1604, SN 1572, or SN 1006.

Supernova models have still to explain the cutoff between stars that will yield stable final cinders and those that continue to collapse to form black holes. The effects of mass loss in the progenitor structure are still unclear.

The origin and evolution of pulsar magnetic fields, while qualitatively explained by several models, have yet to produce a reason for the large obliquities inferred for neutron star magnetic fields. The interaction of these oblique fields with accreting matter in close, mass-exchanging binary systems is still poorly understood. In particular, much work is still required to pull the observations of the different regions of the accretion disk together, a task that involves correlating observations at X-ray, ultraviolet, optical, and radio wavelengths. The discovery of quasi-periodic X-ray sources has begun the probe of neutron star magnetospheric structures, but much is left to do.

The details of the internal properties of neutron stars, especially the equation of state for the core, the solidification of nuclear matter, whether or not pion condensates or other exotic forms of matter occur, and the effects of superstrong magnetic fields on the equation of state, are among the important problems that will need to be understood before a complete model for neutron stars can be determined.

At this juncture, more than 20 years after the discovery of pulsars, one can at least say that neutron stars exist. They are likely to remain the best available physical laboratory for the study of most extreme stable states of matter for some time to come.

SEE ALSO THE FOLLOWING ARTICLES

BINARY STARS • MAGNETIC FIELDS IN ASTROPHYSICS • NEUTRINOS • NUCLEAR PHYSICS • PULSARS • RELATIVITY, GENERAL • STELLAR STRUCTURE AND EVOLUTION • SUPERNOVAE

BIBLIOGRAPHY

Alpar, M. A., Buccheri, R., Ögelman, H., and van Paradijs, J. (eds.) (1998). "Many Faces of Neutron Stars," Kluwer, Dordrecht, The Netherlands.

Alpar, M. A., Kiziloğlu, Ü., and van Paradijs, J. (eds.) (1995). "The Lives of Neutron Stars: Proc. NATO ASI," Kluwer, Dordrecht, The Netherlands.

Arnett, W. D. (1997). "Supernovae and Nucleosynthesis," Princeton Univ. Press, Princeton, NJ.

Bethe, H. A. (1971). *Annu. Rev. Nuclear Sci.* **21,** 93.

Blandford, R. D., Applegate, J. H., and Hernquist, L. (1983). Thermal origin of neutron star magnetic fields, *M.N.R.A.S.* **204,** 1025.

Drechsel, H., Kondo, Y., and Rahe, J. (eds.) (1987). "Cataclysmic Variables: Recent Multi-frequency Observations and Theoretical Developments, Reidel, Dordrecht, Netherlands.

Glendenning, N. K. (1997). "Compact Stars: Nuclear Physics, Particle Physics, and General Relativity, Springer-Verlag, Berlin.

Joss, P. C., and Rappaport, S. A. (1984). Neutron stars in interacting binary systems, *Annu. Rev. Astron. Astrophys*. **22,** 537.

Lyne, A., and Graham-Smith, F. (1998). "Pulsar Astronomy," 2nd Ed., Cambridge Univ. Press, Cambridge.

Manchester, R. N., and Taylor, J. H. (1977). "Pulsars," Freeman, San Francisco, CA.

Mészáros, P. (1992). "High-Energy Radiation from Magnetized Neutron Stars," Univ. of Chicago Press, Chicago.

Michel, F. C. (1991). "Theory of Neutron Star Magnetospheres," Univ. of Chicago Press, Chicago.

Oppenheimer, J. R., and Volkoff, G. M. (1939). On massive neutron cores, *Phys. Rev.* **55,** 374.

Pethick, C. J., and Ravenhall, D. G. (1995). *Ann. Rev. Nucl. Part. Sci.* **45,** 429.

Prakash, M., Bonibaci, I., Prakash, M. *et al.* (1997). *Phys. Rep.* **280,** 1.

Shapiro, S., and Teukolsky, S. (1983). "Black Holes, White Dwarfs and Neutron Stars: The Physics of Compact Objects," Wiley (Interscience), New York.

Thorsett, S. E., and Chakrabarty, D. (1999). "Neutron stars mass measurements. I, Radio pulsars," *Apj*, **512,** 288.

Tsuruta, S. (1998). "Thermal properties and detectability of neutron stars, II. Thermal evolution of rotation-powered neutron stars," *Phys. Rep.* **292,** 1.

Pulsars

F. Curtis Michel
Rice University

GLOSSARY

Dispersion measure Observational measure of the extent to which radio pulses are delayed by propagation through the interstellar medium; used to estimate distances to individual pulsars.
Drifter Pulsar that exhibits drifting subpulses.
Drifting subpulse Subpulse that moves systematically across the integrated pulse profile.
Integrated pulse profile Apparent pulse shape (intensity versus time) of a pulsar as determined by averaging together a large number of otherwise weak and noisy individual pulses.
Neutron star A star more massive than $\sim$1.4 times the mass of the sun cannot become a white dwarf, but instead must collapse (once it has exhausted its supply of hydrogen) to what is in effect a single atomic nucleus $\sim$10 km in diameter.
Nulling A number of consecutive pulses may be missing in any long train of pulses. A few pulsars exhibit this "nulling."
Subpulse Persistent feature within the integrated pulse profile, often noticeable in single pulses.
Supernova Explosion of a star that temporarily produces a "star" as bright as an entire galaxy ($\sim 10^{11}$ ordinary stars). Left behind are an expanding shell of gas and a remnant collapsed object often observed as a pulsar.
White dwarf Star that has exhausted its internal sources of energy and has therefore shrunk to a dense sphere comparable in size to the earth.

ONE OF THE very important discoveries in astronomy of modern times has been that of the radio pulsars: stellar objects that emit high-intensity pulses of coherent radio emission. These astronomical "clocks" can be extraordinarily stable and provide important probes of interstellar space besides being enigmatic objects themselves.

I. INTRODUCTION

In astronomy, *pulsar* denotes an object emitting sharp, rapid pulses of radio emission with clocklike periodicity of about 1 sec. The discovery of pulsars by Antony Hewish and Jocelyn Bell, announced in 1968, earned both of them

a permanent place in the history of astronomy and Hewish won a Nobel prize.

Pulsars are now believed to be neutron stars, stars of mass comparable to that of the sun but collapsed to a sphere of only ∼10 km in radius (i.e., nearly 100,000 times smaller than our own sun). These neutron stars are thought to be intensely magnetized and in rapid rotation. The rotation serves two roles. First, the rotation of the highly conducting neutron through its own magnetic field induces strong electric fields, which pull charged particles from the surface; second, this magnetic field is seen from different aspects by a distant observer on the earth as the neutron star rotates. Theories generally assume that the radio emission is concentrated into a beam, with electrons being accelerated out of the magnetic polar "caps" and emitting radio waves, which sweep the sky like a lighthouse beam as the star rotates.

A second type of rapidly pulsating object was later discovered to emit X-rays in certain binary systems. These objects, called pulsating X-ray sources, are also believed to be rotating magnetized neutron stars, but in this case the neutron star orbits a companion star that is transferring mass from its outer atmosphere onto the neutron star. This mass influx is diverted by the magnetic field and falls primarily on the poles, which are heated to millions of degrees by this bombardment and emit X-rays. Again, rotation of the neutron star modulates the X-rays seen on earth.

As a class, the two are distinct both observationally (X-ray emission versus radio emission) and theoretically (accretion of mass onto a neutron star versus ejection of charged particles). However, a few pulsars emit both radio waves and X-rays, notably, the so-called Crab pulsar located in the center of the Crab nebula.

II. DISTRIBUTION OF PULSARS

Radio pulsars are dispersed among the ordinary stars in the galaxy, and 99% of them are single objects not in binary systems, again unlike the pulsating X-ray sources and also unlike ordinary stars, half of which are binary. Nor are they visible as stars even to the best telescopes. Virtually all the detectable energy output is in radio waves, with a spectrum of emission that declines rapidly with increasing frequency. This decline would in itself make radio pulsars too feeble to be seen at the much higher frequencies of the visible spectrum. Most of the stars in the galaxy are concentrated in the disk of the galaxy (the Milky Way), and pulsars are markedly concentrated in the same way. The distances to pulsars can be estimated from their observed dispersion measure. Because radio waves interact with the very few electrons in outer space (only ∼30 per liter!), the lower frequency waves in the spectrum travel more slowly, which causes a sharp pulse to be smeared out; the same thing happens to the sharp pulse created by a lightning stroke, giving the Whistler phenomenon when the waves are detected at large distances. The dispersion measure corresponds to the amount of correction for delay necessary to reconstruct a sharp pulse, which also determines the integrated electron density along the path traveled. The above estimate for the electron density then gives an idea of the distance, the closest being ∼80 parsecs (pc) away and some are detected as far away as 55,000 pc.

Most pulsars do not have proper names but are named according to their place in the sky. Thus, the Crab pulsar is also known as PSR $0531+21$, meaning a pulsar (PSR) found at a right ascension of 5 hr and 31 min and a declination of 21° north (plus).

Although ∼700 pulsars have been discovered, the fastest ones are among the most unusual, and it is instructive to examine the properties of the latter. Owing to the precession of the equinoxes, fixed objects in the sky drift to slightly new positions steadily over time. Astronomers have to periodically update these coordinates (which can simply be typed into the controls to point a modern telescope at that spot). Although the pulsar "coordinate names" were never meant to be the way of actually locating them, radio astronomers nevertheless changed from the 1950 coordinate system to the 2000 one a few years ago, which renumbered all the "names." The new labels are preceded by "J," while the old ones are assigned "B." All the names here are the original discovery names ("B"), unless explicitly noted.

III. EXCEPTIONAL PULSARS

The fastest known pulsar is called the Millisecond pulsar, or PSR $1937+214$ (here the extra digit refines the declination to 21.4° north), and has a period of only 1.558 msec, which is quite close to the maximum rotation rate that a neutron star could theoretically have without flying apart owing to the centrifugal forces exceeding gravity. The surface velocity for a typical neutron star according to theory would then be 4×10^9 cm/sec, or 13% the velocity of light. Although extremely rapid, the radio emissions from this pulsar are very similar to those of most other pulsars.

The pulse shape of this pulsar is shown in Fig. 1, showing intensity (I) and degree of linear polarization (L). Although most of the energy is in a concentrated spike, there is a distinctive notch in that spike, and also an interpulse is seen halfway between successive main pulses (one full cycle of emission is shown). Indeed, as a class, pulsars

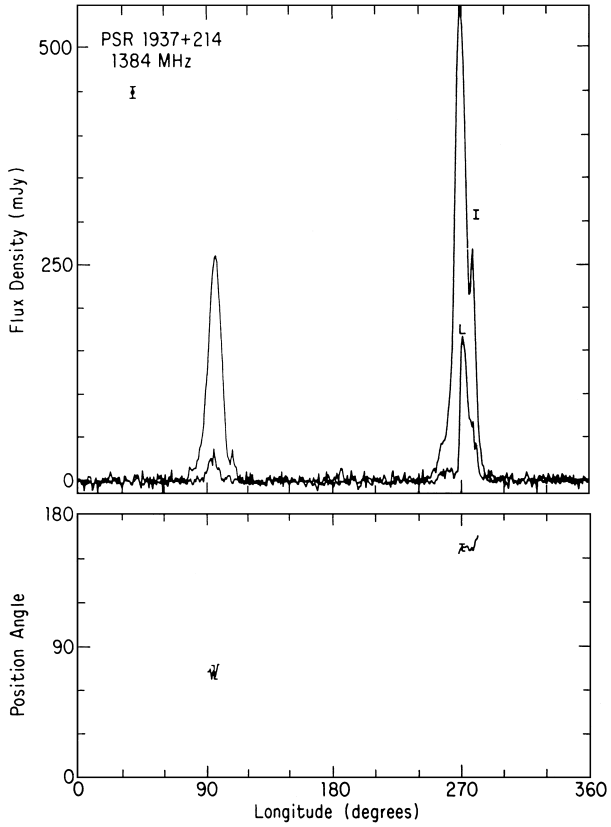


FIGURE 1 Pulse shape of PSR 1937 + 214. Although this pulsar is currently the fastest known, the pulse shape is not particularly remarkable. Total intensity is *I* and linearly polarized intensity is *L*, showing that the main pulse (right) is about 33% polarized near the maximum. The width of the main pulse is about 10°, which is also fairly typical, and interpulses (left), although not exhibited by most pulsars, are not uncommon. The spacing of interpulses very nearly halfway between successive main pulses is often interpreted as the observer seeing both magnetic poles of the pulsar, first one and then the other as the neutron star rotates. The position angle is the axis of polarization projected on the sky (usuaully measured from north), and here we see that the two pulses have just about orthogonal polarization. A "swing" in polarization corresponds to a rapid change in position angle with longitude (not seen here). [Reprinted by permission from Stinebring, D. R., Boriakoff, V., Cordes, J. M., Deich, W., and Wolszczan, A. (1984). "Birth and Evolution of Neutron Stars: Issues Raised by Millisecond Pulsars" (S. P. Reynolds, and D. R. Stinebring, eds.), p. 35, NRAO, Green Bank, WV.]

are similar to fingerprints; each is readily identifiable as such but is nevertheless distinct on close examination. Most, for example, do not have the interpulse, and many would have multiple "components" if we were to view the notch as separating two distinct components, etc. This pulsar is a particularly accurate clock, and the pulse period has been determined to high precision: 1.5578064488737 (±6) msec. Like virtually all other pulsars, it is slowing down, but only very slowly (about 10^{-19} sec/sec).

A second millisecond pulsar, PSR 1957 + 20, has almost the same period (1.607 msec) and has the extraordinary property of being in a binary system and eclipsing its companion. The companion is heated by the pulsar radiation and is seen as a variable star at the orbital period of 9.2 hr. For a likely pulsar mass of around 1.4 solar masses, the mass of the companion turns out to be only 0.022 solar masses, the lightest known star. The eclipses cannot be entirely due to the companion moving in front of the pulsar because they are too large, corresponding to distances at which matter would not be gravitationally bound to the companion. Many think that the companion is essentially a comet evaporating in this system (but the eclipses are quite symmetric, which is a problem). Another idea is that plasma in a magnetosphere about the companion is the occulting agent. We will discuss where the estimate for the pulsar mass comes from in Section IV.

An even "slower" pulsar, PSR 1821 − 24 at 3.054 msec (about 20,000 rpm!), is one of an interesting new class that is found near the centers of globular clusters, M28 in this case (the designation is from a catalog compiled by the comet hunter Messier, who was frustrated by the various other fuzzy objects in the sky). This pulsar is not in a binary system, but the field pulsar PSR 1855 + 09 (period 5.362 msec) is, with a period of 12 days. A "field" object is one that can pop up anywhere, as opposed to objects associated with clusters, etc. The companion is unseen, and its properties are largely unknown, other than that the mass probably exceeds 0.2 solar masses. PSR 1953 + 29, which was discovered shortly after the millisecond pulsar with a period of 6.133 msec, is also a field pulsar but also in a binary system, here with a period of 117 days (about that of Mercury about the sun) and a similarly unseen companion of approximately the same minimum mass. PSR 1620 − 26 is an 11.076 "millisecond" 191-day binary pulsar in the globular cluster M4.

This list of millisecond pulsars is growing continually, especially for pulsars in globular clusters. Once one pulsar is discovered, the dispersion measure to the globular cluster is known. These distant pulsars are too faint to give directly detectable pulses, and the data must be processed with a very large number of trial values for period and dispersion measure until a pulse can be found. Knowing the dispersion measure, therefore, greatly reduces the effort to search for additional pulsars in the globular cluster. For example, the globular cluster M15 is now known to harbor at least five pulsars designated PSR 2127 + 11A, B, C, D, and E in order of discovery, with periods of 110, 56, 30, 4.80, and 4.65 msec, respectively, one of which is thought to be binary and one of which (the slowest) has an increasing period, possibly because it is in a very wide binary system. Such systems may require years of observation to determine the orbital properties. Although the

periods are also in the order of discovery, this seems to be a coincidence.

Until 1982, the fastest known pulsar was the Crab pulsar (PSR $0532+21$, now known as J0534 + 2200), at 33.1 msec, which sits centered on what is probably the remnant of a historical supernova observed by the Chinese in 1054 AD. This remnant, the Crab nebula, is expanding at a measurable rate consistent with such a birthdate. The pulsar is remarkable in that it is also a source of visible light, which is pulsed at the same 33-msec period as the radio, too fast for the eye to follow! As with typical pulsars the radio emission declines rapidly with frequency, and some researchers believe a separate mechanism may cause the pulsar to become visible again at visible frequencies. [Roughly speaking, radio frequencies are of the order 10^9 cycles per second (hertz), whereas visible light is $\sim 10^{15}$ Hz.] Moreover, the high-frequency part of the spectrum extends into X-ray and γ-ray energies (of the order of 10^{21} Hz), again pulsed. The Crab nebula itself has been discovered to be jumping about in response to the pulsar in the optical, and images can be viewed at http://oposite.stsci.edu/pubinfo/pr/96/22.html. New X-ray satellite images show similar structures: http://chandra.harvard.edu/photo/0052/index.html.

A similar pulsar is PSR $0540-693$, with a period of 50 msec and also surrounded by a nebula, resembling quite closely the Crab pulsar except for the distance. This pulsar is in the Large Magellanic Cloud (LMC) some 55 kpc away. Like the Crab pulsar, it emits visible pulses.

A very important pulsar, altrough slightly slower at 59.0 msec, is again one in a binary system and is often called the Hulse–Taylor binary pulsar (PSR $1913+16$) after its discoveres. (About 20% of the known pulsars have right ascensions of 19 hr, which is largely a selection effect owing to the fact that the look direction of the giant fixed radio telescope at Arecibo rotates across the Milky Way at this location.) This pulsar was the first binary pulsar to be discovered and is possibly the most important. The orbital period is only 7 hr and 45 min, and general relativistic effects are important. This binary system is the first single-line binary (i.e., only the pulsar is detected) for which all of the orbital elements have been deduced (by means of the general relativity theory), including observation of the advance of perihelion (the same effect explained for Mercury) at a rate entirely consistent with theory. More importantly, the binary pair are spiraling together at a rate consistent with energy loss by gravitational radiation. Unlike the 6.1-msec binary pulsar, the unseen companion also has a mass of 1.4 solar masses, suggesting that it may also be a neutron star.

Another fast pulsar discovered early on (PSR $0833-45$) is the Vela pulsar associated with the Vela supernova remnant, with a period of 89.2 msec. Like the Crab pulsar, it has high-frequency emissions. Unlike any other pulsar, however, it exhibits extremely large "glitches," wherein its period abruptly decreases by a small but readily detectable amount of about one part in a million. These events repeat at an irregular interval of ~ 3 years and do not have exactly the same behavior each time.

This listing of pulsars illustrates a number of important observational inferences, detailed in Section IV. Most of the next dozen or so pulsars having periods between 100 and 200 msec tend to be isolated pulsars without striking or unusual properties.

IV. WHERE DO PULSARS COME FROM?

The Crab pulsar and surrounding supernova envelope strongly support theoretical arguments that neutron stars form in supernova events. Numerical simulations indicate that a very massive star (10 times that of the sun, say) grows a core as it evolves until it essentially has a dense white dwarf at its center surrounded by an "atmosphere," which is the rest of the (giant) star. The core provides energy to the star by accreting more atmosphere, but the huge luminosity of such stars (on order of 10,000 times that of the sun) rapidly removes the energy until the core grows so massive that it collapses and forms a neutron star. This collapse releases a sudden burst of energy, which ejects the outer shell of the core and the atmosphere—the rest of the massive star. It had already been shown years ago by S. Chandrasekhar that a white dwarf could be no more massive than about 1.4 times the mass of the sun without collapsing to a point, so the collapse of the core is inevitable. This theoretical analysis was dramatically confirmed in 1987 when a giant star in the LMC became a supernova, designated SN 1987A. The reason that white dwarfs can only be so massive is that they are supported almost entirely by the pressure of degenerate electrons. This is the same pressure that effectively gives atoms finite sizes instead of the electrons being bound arbitrarily tightly. Planets are thus held up by electron degeneracy pressure in the guise of the near-incompressibility of solid matter. But near-incompressibility is not good enough, and at high enough pressures the electrons become relativistic, which signals a slight weakening in their ability to hold up stars, and they fail to do so beyond the above Chandrasekhar mass limit (an entirely theoretical calculation that follows directly from quantum mechanics and special relativity). The core does not, however, collapse to a point but stops when the nucleons are pushed close enough together to themselves become degenerate, which in direct parallel with atoms is when the density reaches that of nuclear matter instead of ordinary matter. This is our neutron star. There is a mass limit for neutron stars just

as for white dwarfs, but it is harder to calculate because electrons stay electrons under compression but nuclei begin to change into exotic forms ("strange" particles) under such circumstances in ways not yet fully understood. Most estimates give limiting masses around twice that of the sun, and most astrophysicists assume that the neutron star would then collapse to become a black hole.

This understanding of how pulsars are created is adequate for most of the observed pulsars but cannot be complete. The problem is those found in globular clusters. The globular clusters are thought to be among the oldest objects in the galaxy, and massive stars must last only a short time (if the sun lasts billions of years, a star 10,000 times brighter must burn up within millions). So the pulsars formed by the above mechanism (a "type II" supernova; one in which spectral evidence for hydrogen is present—presumably the outer envelope) would have been formed early on and would by now be spinning very slowly, not *faster*, than most pulsars. Thus, surprisingly, there must be another mechanism as well by which these exotic-seeming objects can be made. That there is a second mechanism is further supported by the fact that so many of these fast pulsars are in binary systems, while the "ordinary" pulsars seen scattered about are virtually all single objects. Supernova do, however, occur in systems like globular clusters, namely, the elliptical galaxies. These systems lack the gas and dust with which to manufacture new massive stars and in this respect resemble closely the globular clusters, except for being much larger and containing vastly more stars.

It is thought that the second mechanism for making supernovae (again surprising to have two different sources of such exotic events) is the evolution of white dwarfs' binary systems. Binary star systems are common, and most old stars become white dwarfs, so such systems are not uncommon. If the two orbit one another close enough to transfer matter, one white dwarf loses matter and expands (because it is no longer weighted down by the lost mass), while the other gains matter and shrinks; again, we can have a white dwarf pushed over the Chandrasekhar mass limit. These explosions are probably the "type I" supernovae (no hydrogen detected; white dwarfs have "burned" all their hydrogen already). Calculations suggest that the collapsing white dwarf might be entirely incinerated (it has not been supporting a massive atmosphere all this time and consequently should have more internal fuel left). But under the right circumstances, a neutron star might be left, and in this sort of evolutionary sequence the remnant should be rapidly spinning. A popular alternative is instead that an old pulsar has been captured and is spun up by accretion of matter from the companion. This model requires that the magnetic fields of pulsars decay away (a topic of current debate). It also makes it difficult to explain why some of the millisecond pulsars are single objects (although the eclipsing binary pulsar has a very light companion of 0.02 solar masses, so accretion or some other process could remove most or all of the companion's mass).

V. NATURE OF PULSARS

Pulsars are thought to be formed in supernova events. There are actually only a few pulsar–supernova associations. Many young supernova remnants do not contain detectable pulsars, and most pulsars are not in supernova remnants. The first is partly a selection effect because the remnants themselves are intense radio sources, and only the brighter pulsars can be seen against such a background. Even if the pulsar is bright, its beam might not sweep the earth. The reason for the second is that the remnant expands and fades over a period of $\sim 10^4$ years, whereas the pulsars typically live on for a few million years. On the other hand, the weakly magnetized (so-called millisecond) pulsars survive on time scales leading back to the very formation of our galaxy.

The fastest pulsars provide the strongest test of the rotating neutron star hypothesis. Early theories involving white dwarfs as the pulsar object were ruled out by the short time scales involved; such stars are simply too large to behave coherently for periods much less than 1 sec.

All pulsars are observed to be slowing down. Again, this observation is consistent with a rotating object for which the source of free energy is the kinetic energy stored in rotation. The energy output can then be calculated by estimating the moment of inertia of the neutron star (essentially its mass times the square of its radius) and using the observed rate of increase of period to determine how rapidly this energy is declining. The rate of change of the periods spans a large range, but a change of period of $\sim 10^{-15}$ sec/sec is representative. Again, for the weakly magnetized millisecond pulsars, the corresponding period changes are more like 10^{-20}. A pulsar with a period of 1 sec (again typical; the pulsars noted above are the fastest) has a stored energy of 10^{46} ergs, and therefore the energy loss rates are $\sim 10^{31}$ ergs/sec. These figures are always much larger by a comfortable factor ($\sim 10^5$) than the energy output in radio waves. Most of the pulsar energy output seems to be invisible and is thought to be in the form of a relativistic magnetized "wind" flowing away from this pulsar. However, the entire Crab nebula supernova remnant radiates power comparable to that lost from the pulsar, suggesting that it is the wind from the Crab pulsar that lights up the nebula.

The pulsating X-ray sources behave quite differently; they have quite large spin-up and spin-down rates, so that the period of a given source may increase over about 1 year

and then decline. The spin-up is thought to be caused by angular momentum carried in by the accreted matter. The spin-down mechanism is less certain but is usually attributed to electromagnetic coupling of the neutron star, assuming it has a strong magnetic field just as the radio pulsars do, to an orbiting disk of matter. Indeed, the rotation periods of many of these binary X-ray sources are much longer than what a radio pulsar would ever spin down to in the age of the galaxy (about a 25-sec period).

The glitch phenomenon has proved difficult to explain. Most theories concentrate on some change in the physical body of the neutron star itself. A "starquake" that released internal stresses and allowed the neutron star to shrink slightly would explain the spin-up seen in the Vela glitches, but these events happen too frequently. The star cannot endlessly shrink. Present models argue instead for an interior core, which is decoupled from the crust and rotates more rapidly. The glitches might then be abrupt braking phenomena that transfer angular momentum from the core to the crust and spin up the latter. The strong magnetic field, however, would lock the crust and core into corotation, so these models impose special conditions on this field (e.g., it would have to reside entirely in the crust and be excluded from the core). The Crab pulsar also displays these events, but with much smaller amplitude and much more frequently. However, PRS $1509-26$ (also associated with a supernova remnant) has an exceptionally large spin-down rate of 1.5×10^{-12} sec/sec and yet has shown no glitches whatsoever. A few slow old pulsars (PSR $1641-45$ and PSR $1325-43$) have experienced single glitches of amplitude intermediate between those of the Crab and Vela.

With these and a few other exceptions, pulsars have extremely stable slowing-down rates, and if their pulse rates are corrected for spin-down, they rival the finest atomic clocks. The location of a pulsar in the sky can be determined to high accuracy (comparable to that for visible stars) using the above clocklike properties. Owing to the earth's orbital motion about the sun, the pulsar clock appears to run fast and slow at different times of the year. This variation can be removed only if the pulsar is located at a very specific location in the sky. The amplitude of this period variation gives the declination, and the phase gives the right ascention. In general, there are no stars found at these well-determined locations that seem likely visible counterparts (except for the Crab, Vela, and PSR $0540-693$, which display optical emission).

VI. THE PULSES

Pulsars would not sound like the good clocks that they are if they could be heard. There are large pulse-to-pulse variations, and occasional pulses may even be missing. Several hundred individual bursts of radio emission added together on top of one another form a generally stable integrated pulse profile, but this profile serves more as a "window" through which a variety of puzzling pulse variations are witnessed. In general, the individual pulses bear little resemblance to the integrated profile, often as sharp spikes appearing more or less randomly within the window. In some cases, however, the behavior is quite coherent, with a smaller pulse (subpulse) appearing at one edge of the window and steadily working its way across to disappear at the other edge. These drifting subpulses are found in perhaps 10% of the pulsars, with PSR $0809+74$ being a classic example. Another pulsar (PSR $0826-34$) is a remarkable example in that it has an extremely wide profile within which can be seen four or five subpulses at once. These subpulses are spaced about 30° apart (defining 360° as one pulse period) and move back and forth together. The interpretation is that we are looking nearly along the spin axis of a neutron star that is also magnetized nearly along the spin axis, and therefore, we are almost constantly in the pulsar emission beam. Both of these drifters also display nulling (as do some nondrifters), a phenomenon in which the pulsar disappears, becoming entirely undetectable for 10 to even 1000 periods before reappearing abruptly. One pulsar, PSR $0904+77$, has never been seen since its first detection. If not a spurious observation, it is an extreme example of nulling.

The phenomenon of nulling shows that pulsar action can cease, and indeed, it must if all pulsars slow down. The inevitable consequence should be an accumulation of older and slower pulsars, whereas observationally the average pulsar has a period of $\sim$0.5 sec and yet the slowest pulsar (PSR J1951 + 1122 at present) has a period of only 5.1 sec. The pulsating binary X-ray sources, on the other hand, have periods of 700 sec or longer. Apparently, pulsars as such must rapidly disappear once they slow below 1 sec or so. Pulsars with a 1-sec period have slowing-down rates of $\sim10^{-14}$ sec/sec, implying that they will live for only $\sim10^{14}$ sec $=3\times10^6$ years. Since most of the time is spent as a slow pulsar, the typical pulsar lifetime is still expected to be of the order of a few million years regardless of how fast the pulsar was rotating at birth.

The slowing down of the fastest pulsars is much less than average, a seeming contradiction. It would take $\sim4\times10^8$ years for the 1.558-msec pulsar to become even a 3-msec pulsar and extraordinarily long for it to become a 4-sec pulsar. In the magnetized rotating neutron star picture, it is the strong magnetic field that couples the rotating star to its surroundings. Thus, an unusually weak magnetic field would account for the small slowing-down rate. Indeed, it is only when a rapid pulsar has a weak magnetic field that we can ever hope to observe it, unless

it is seen shortly after birth. The Crab pulsar is 1000 years old and spins at 33 msec. If we scale the lifetime it could have had (with its apparently strong magnetic field of $\sim 4 \times 10^{12}$ G) with a spin rate of 1.5 msec, we find that it would double its period in only a few years. Thus, only in a fresh supernova could one hope to see a strongly magnetized millisecond pulsar. Such events are rare; centuries can pass between known supernova events in our own galaxy, and it is not evident that the surrounding dense cloud of debris would permit the pulsar to shine through early on. Supernovas are more frequently seen in other galaxies (there are so many of them), but associated pulsars have not been detected (possibly they are simply not bright enough to be seen at such huge distances—megaparsecs). Unexpectedly, supernova SN 1987A in the nearby LMC has yet to show any evidence that it contains a pulsar, even 13 years after the event. Owing to their long lifetimes, the millisecond pulsars are largely a class that have accumulated in the galaxy, whereas the strongly magnetized field pulsars are a young population, and we do not see the old ones because they would have spun down and become too faint (even if no other effect of aging entered: cooling, possible decay of the magnetic field, etc.). Ironically, pulsars are more readily observable if the magnetic field is not too strong, even though presumably pulsars require strong fields to operate in the first place.

VII. THEORY OF PULSARS

A large amount of circumstantial evidence, as shown in the preceding section, now exists on the behavior of pulsars. Yet the very fact that we can detect them at all is poorly understood. The radio luminosity is $\sim 10^{28}$ ergs/sec for a typical pulsar, and yet a neutron star is only ~ 10 km in radius, which would require it to be an extremely efficient antenna. (If much more energy than that in the radio spectrum went into making the system glow in the visible spectrum, some should have been seen as dim stars.) The simplest picture of how so much radiation might be produced in such a small region is the "bunching" hypothesis, which assumes that the energetic electrons producing the radio waves are bunched together. In a simple geometry one might consider a spherical bunch of N particles. At long wavelengths the electrons radiate as if they were fused into a single particle of charge Ne, giving emission N times more intense than N separate electrons radiating independently. At wavelengths much shorter than the electron spacing, the electrons would radiate more or less independently despite the bunching. Thus, bunching gives a natural qualitative account of why pulsar emission is intense to low frequencies and declines rapidly at high frequencies.

It was early on suggested by P. Sturrock that the pulsar mechanism had something to do with an exotic combination of physical mechanisms wherein electrons are accelerated in the huge electric fields generated by the neutron star rotating through its own magnetic field. These electrons are constrained to follow these curved magnetic field lines and consequently emit curvature radiation, analogous to the emission of synchrotron radiation by electrons moving in curved trajectories around magnetic field lines (a process thought to be unimportant in pulsar magnetospheres because the electron would almost instantaneously lose such energy and it would not be replaced). Indeed, γ-rays are detected from the Crab and Vela pulsars possibly produced by this mechanism. An even more exotic process possible in such huge magnetic fields is the conversion of γ-rays into electron-positron pairs. It is thus possible that a single electron would emit a γ-ray and thereby produce a second electron (and positron). Unfortunately, it was not understood why bunches might form. Moreover, popular models of the time assumed that the magnetosphere was entirely filled with plasma pulled from the stellar surface by the huge electric fields. As a result, the magnetic field lines were thought to be largely shorted out by this plasma, which reduced the acceleration of the electrons to levels unpromising for the above mechanism. Another problem was that the bunching was assumed to result from some instability, and, given the small scale of a neutron star system together with the high velocities of relativistic velocities, such instabilities had to be extraordinarily fast developing.

A considerable theoretical effort was invested in analyzing this magnetized rotating neutron star model. At first it was thought that the simplest possible model, a dipole magnetic field aligned with the spin axis, would emit particles from the polar caps like a pulsar and be a valuable test bed for radiation models. Indeed, most radiation models have assumed acceleration of electrons from the polar caps as a starting point (often attempting to include pair production). It was later realized that such particle emission could only be a transient phenomenon, because only electrons would be lost from the system. The resultant electrostatic charging of the neutron star eventually would halt any further loss of electrons. Positive particles that might neutralize this charging should also be emitted, but they would be trapped very close to the star on the closed equatorial magnetic field lines. Although the model seemed disappointingly unlike a pulsar, it is nevertheless interesting because the neutron star ends up surrounded by an equatorial torus of positive charge and two clouds of electrons over the polar caps, with vacuum elsewhere. It is unusual to find plasmas that consist only of electrons because the self-repulsion of the electrons should

blow the plasma apart. However, laboratory workers have succeeded in producing just such plasmas by trapping large numbers of particles in magnetic Penning traps. These nonneutral plasmas have the remarkable property of behaving like liquids, with a sharp boundary separating a constant density of particles from the vacuum surrounding them. However, there is no reason to expect pulsars to be aligned rotators; that was merely a mathematical simplification, and many wondered if an inclined rotator might be fundamentally different.

A promising new approach takes advantage of what seemed to be a disadvantage, the plasma voids in the magnetosphere. Here, electrons can indeed be accelerated to the energies envisioned by Sturrock. But such a process would simply fill up the magnetosphere of an aligned rotator. For an inclined rotator, however, large amplitude electromagnetic waves form high above the neutron star polar caps. As a consequence, the magnetospheres of inclined rotators can never entirely fill up because the plasma is driven away by the radiation pressure from waves created by the neutron star itself. Thus, electrons can be accelerated by strong electric fields in the void regions created by the electromagnetic waves. Consider a downward-moving electron. It radiates γ-rays, which, if close enough to the star, form pairs. In the strong field, the positron is immediately stopped and ejected, while the new electron is essentially a companion of the original electron and in the strong electric field is almost immediately accelerated to the same energy (the energy is now limited by that rate at which the γ-rays are being emitted). But two electrons quickly become 4, 16, 256, etc., and a highly charged downward-moving bunch is created from *a* single initiating charge. Of course, the upward-moving positrons do the same thing, giving a breakdown in the huge electric fields that create upward- and downward-moving sheets of charge. Thus, it is possible that the bunching (now something of a misnomer because it is not ambient plasma that is being bunched at all) is an intrinsic part of the discharge process. Much more needs to be worked out before such a simple picture can be tested and compared to the abundant data on pulsars.

In all of these models one expects a variation in the apparent (projected) direction of the magnetic field as the emission region swings past the observer. A large number of pulsars indeed show a systematic rotation in the polarization direction of their radiation during each pulse, with the rate being fastest at the center of the pulse (which need not be the point of maximum brightness). These observations are broadly consistent with theory assuming a simple tilted dipole magnetic field, and attempts have even been made to deduce the actual angles between spin axis and observer plus magnetic field for some pulsars. Unfortunately, the rotation rate gives only one number, whereas there are two unknown angles to be solved for, so some additional assumption must be made.

The simplest expectation would be that the radio emission is polarized parallel to the magnetic field lines. (Curvature radiation from relativistic electrons followed curved magnetic field lines would have this behavior.) In fact, one observes in many pulsars abrupt 90° changes in polarization at certain points in the pulse, so if the radiation was originally along the field lines, it must suddenly become orthogonal to them, which is not a property of curvature radiation. The attractive idea of relativistic electrons somehow bunching and radiating as they move on curved magnetic field lines is therefore incomplete. It may be that the more or less static nonneutral plasma thought to be trapped near the neutron star complicates the emission mechanism in interesting ways.

Returning to the supernova origin hypothesis, the supernova event rate seems broadly consistent with that necessary to maintain pulsars against their apparent death rate, with there being considerable uncertainty in each. (For example, only $\sim$0.1% of the pulsars in our galaxy have been detected, and the supernova rate is actually an average over supernovas in other galaxies apparently similar to our own.) These rates are thought to be about one per 50 years. It is not believed that the observed pulsars centered on supernova remnants are chance associations (particularly the Crab pulsar, which seems to be exciting the nebula). Supernova remnants occupy $\sim$1% of the sky in a 10° wide band along the Milky Way. The pulsars are similarly distributed, showing that they are produced mainly by stars in the disk of the Milky Way and that they are seen at distances that are large compared with the thickness of this disk. A few of the approximately 700 observed pulsars could therefore be accidentally in line with a remnant. However, the fact that the pulsars seen in these remnants are also very fast ($<$100 msec) reduces significantly the chances of accidental association. Presently, three of the six pulsars faster than 100 msec are in supernova remnants. Since the millisecond pulsars live so long, there is little hope of still finding the smoking gun.

The supernova association is also complicated by uncertainty in the pulsar beam geometry. Presumably, one observes a one-dimensional slice through the beam cross section it sweeps across the earth, but there is no information about the extent of the beam above and below. If the beam is roughly circular, there must be a large number of unseen pulsars, because the beams are observed to be so narrow ($\sim 10^\circ$ of the latitude if rotating), and statistically only about one in five would be expected to sweep over the earth, about the same ratio as remnants with pulsars to those without.

Extinct pulsars may be resuscitated as γ-ray burst sources long after they fade from view. These sources emit single isolated but intense bursts of γ-rays lasting $\sim$1 sec. There is little concensus on what produces these bursts, with models ranging from an asteroid falling onto the neutron star to rapid accretion of material from a disk. One such event, the brightest ever observed, which occurred on March 5, 1979, has even been localized near a supernova remnant designated N49 in the LMC. This is tantalizing but not necessarily supportive of the idea that the source is an extinct neutron star, because then the remnant should long since have dissipated. It has recently been suggested that some such events might be triggered by the formation of a class of very strongly magnetized pulsars ("Magnetars"), with magnetic fields about 1000 times as strong as typical pulsars. In contrast, the millisecond pulsars are weaker by about this same factor. Alternatively, the association could be accidental, with the γ-ray burst source being actually much closer than average rather than being much farther than average, which would explain the brightness. However, recent observations suggest just the opposite, namely, that the γ-ray bursts may be from huge ("cosmological") distances, which if true would render the above burst a very weak one, bright only because it was so "close."

SEE ALSO THE FOLLOWING ARTICLES

GAMMA-RAY ASTRONOMY • GRAVITATIONAL WAVE ASTRONOMY • NEUTRON STARS • STELLAR STRUCTURE AND EVOLUTION • SUPERNOVAE • X-RAY ASTRONOMY

BIBLIOGRAPHY

Lyne, A. G., and Smith, F. G. (1998). "Pulsar Astronomy," Cambridge Univ. Press, New York.

Manchester, R. N., and Taylor, J. H. (1977). "Pulsars," Freeman, San Francisco.

Michel, F. C. (1991). "Theory of Neutron Star Magnetospheres," Univ. of Chicago Press, Chicago.

Smith, F. G. (1977). "Pulsars," Cambridge Univ. Press, New York.

Quasars

B. M. Peterson
S. Mathur
P. S. Osmer
M. Vestergaard
Ohio State University

GLOSSARY

Accretion disk A gaseous disk surrounding a source that is powered by gravitational accretion. Viscosity heats the accretion disk, causing high temperatures that result in thermal radiation from the disk and allowing dissipation of angular momentum.

Eddington luminosity The maximum luminosity that can be obtained from spherical mass accretion. Higher rates produce higher luminosities, which lead to radiation pressure (on an ionized gas) exceeding self-gravity.

Gravitational lensing Distortion of space–time near large masses that causes deflection of light paths; the mass acts as a lens and can produce multiple images of objects located behind the lens.

Nonthermal emission Emission due to processes that depend on physical conditions other than the temperature of the source, such as synchrotron radiation (relativistic electrons spiraling in magnetic fields) or the inverse Compton mechanism (scattering of photons off relativistic electrons, increasing photon energies).

Schwarzschild radius The radius of the event horizon of a nonrotating black hole ($R_S = 2GM/c^2$).

Very-long-baseline interferometry (VLBI) Technique of aperture synthesis over long radio antenna separations (hundreds of kilometers or more). Angular resolution of milliarcseconds can be achieved by using receivers across the Earth.

Virial theorem For a system in which angular momentum is conserved, the time-averaged relationship between the kinetic energy K and potential energy U is $2K + U = 0$. This can be used to measure the masses of complex stellar systems.

THE TERM "quasar," shorthand for "quasi-stellar radio source," was coined as a description of point-like (or

star-like) radio sources that are out of the galactic plane but could not be clearly identified with extragalactic nebulae (external galaxies) when they were first discovered in early radio surveys of the sky. Accurate radio positions were obtained for a small number of these high-galactic-latitude sources, showing that they were clearly star-like in the optical. Optical spectra of these sources, which showed only very broad low-contrast features, did little to clarify their nature. In 1963, Maarten Schmidt realized that the broad features in the spectrum of the source 3C 273 were in fact the hydrogen Balmer-series emission lines that are seen in the spectra of galactic nebulae, but at an uncommonly large redshift, $z = 0.158$, where $z = (\lambda - \lambda_0)/\lambda_0$ and λ and λ_0 are the observed and laboratory wavelengths, respectively, of the emission line. The emission lines in quasars are much broader than those in any other astrophysical source except supernova remnants: Their widths imply internal velocity dispersions of a few to several thousands of kilometers per second.

I. QUASARS: THE MOST LUMINOUS ACTIVE GALACTIC NUCLEI

Redshifts as large as that measured for 3C 273 had been observed for some of the most distant known clusters of galaxies, but what was extraordinary about 3C 273 was the great luminosity implied by its redshift. If quasar redshifts are a consequence of the expansion of the Universe, the luminosity of 3C 273 is more than 100 times greater than that of a typical giant galaxy. This result was even more alarming when it was realized how compact these sources are. Quasars can undergo significant changes in their brightness on time scales of days to weeks. Since the parts of the emitting region that are varying must be causally connected, light travel time sets an upper limit to the size of the varying region. The problem with quasars is thus immediately apparent: They are objects as small as the Solar System (a few light days or light weeks across), but more luminous than an entire galaxy of stars (hundreds of thousands of light years across).

While quasars were discovered on account of their radio emission, it soon became apparent that quasars also tended to be bluer than stars and that optical searches for blue objects at high galactic latitude might be an efficient way of finding them (Section V.A). Searches for objects with "ultraviolet (UV) excesses" revealed a surprising number of these sources, too many to account for with known radio sources. Indeed, the original "radio-loud" quasars turned out to be outnumbered by their "radio-quiet" counterparts by a factor of 10 to 20; these became known as quasi-stellar objects, or QSOs, although nowadays the terms "quasar" and "QSO" are generally used interchangeably. The extreme properties of the original radio-loud quasars relative to typical QSOs probably contributed to the difficulty in recognizing the similarity of QSOs to the puzzling Seyfert galaxies. These galaxies had been isolated by Carl Seyfert in the early 1940s on the basis of their high central surface brightness (i.e., star-like central nuclei). They share many properties with QSOs, notably the strong, broad emission lines in their UV and optical spectra. The basic difference between QSOs and Seyfert galaxies is their luminosity: The somewhat arbitrary demarcation between them is now taken to be an absolute B magnitude of -23.0 mag; brighter sources are QSOs, and fainter sources are Seyfert galaxies. Collectively, QSOs and Seyfert galaxies are often referred to generically as active galactic nuclei, or AGNs.

Unlike stars and systems of stars, quasars are extremely luminous in every waveband in which they have been observed, as shown in the typical spectral-energy distribution shown in Fig. 1. Broad peaks in the spectral energy distribution occur (1) in the hard X-ray (2–10 keV) region, (2) in the UV region, and (3) in the mid-infrared. Relative to the higher energy bands, the amount of energy radiated per unit bandwidth at radio wavelengths is lower by a factor of about 100 for the radio-loud objects and by a factor of about 10^4 for radio-quiet sources. The various local peaks in the distribution reflect different physical processes, as will be described below.

B. Taxonomy of Active Galaxies

Active galactic nuclei have traditionally been subdivided into a number of subcategories based on differences in various observed properties. As noted above, higher luminosity AGNs are called quasars or QSOs and lower luminosity AGNs are called Seyfert galaxies. There are, in fact, two types of Seyfert galaxies: Type 1 Seyferts are those with quasar-like UV/optical spectra whose major features are a nonstellar continuum and both broad and narrow emission lines. Type 2 Seyfert galaxies lack the broad components of the emission lines; in at least some cases, these are known to be type 1 objects in which the broad components are obscured from our view. There are also broad-line radio galaxies and narrow-line radio galaxies that are the radio-loud analogues of the type 1 and type 2 Seyferts. Finally, there is a subclass of radio-loud quasars whose emission at all wavelengths is dominated by nonthermal emission from relativistic jets oriented close to the line of sight. These objects are referred to as blazars. Among blazars are two major subclasses: BL Lacertae objects, in which the broad emission lines that characterize quasar spectra are too weak to be detectable by normal means, and optically violent variables, which have much more quasar-like spectra but undergo

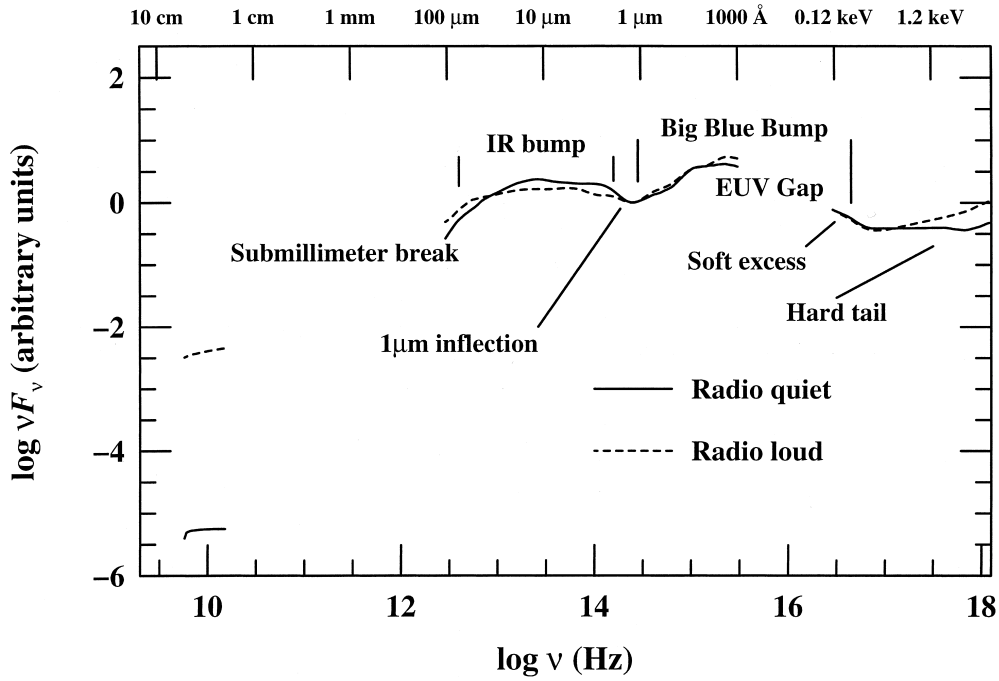


FIGURE 1 Mean spectral energy distributions (SEDs) for radio-quiet (solid line) and radio-loud (dashed line) QSOs. The horizontal axis is photon frequency, and the corresponding wavelengths (or energies in the X-rays region) are labeled on the top axis. The vertical axis is luminosity in ergs per second. Major spectral features referred to in the text are labeled. [Adapted from Peterson (1997), based on data from Elvis, M. *et al.* (1994). *Astrophys. J. Suppl.* **95,** 1.]

strong and rapid continuum variations. At least some of the differences among various classes of AGNs appear to be due to differences in orientation rather than differences in the underlying physics, as discussed in Section VI.B.

II. THE ENERGY SOURCE

A. Supermassive Black Holes

The quasar problem, as noted above, is to try to explain how such an enormous amount of energy can be generated in such a small volume. It was recognized early that gravitational accretion by supermassive black holes was a possible explanation, and indeed this is the only explanation that remains viable after some four decades of research. That QSOs are massive can be inferred from the Eddington limit, which is the minimum mass required for a gravitationally bound, self-luminous object to be stable against radiation pressure. In units appropriate for quasars, $M_{\rm E} = 8 \times 10^5 (L/10^{44}) M_\odot$, where L is the central source luminosity in units of ergs s^{-1}. Thus, a bright quasar with a luminosity of 10^{46} ergs s^{-1} must have a mass greater than $10^8 M_\odot$. An equivalent way to think of this is to define the Eddington luminosity $L_{\rm E} = 1.3 \times 10^{38} (M/M_\odot)$ ergs s^{-1}, which is the maximum luminosity of a source of mass M that is powered by spherical accretion. This is therefore a useful benchmark in considering accretion rates and efficiencies. The mass accretion rate required to sustain the Eddington luminosity can thus be written as $dM_{\rm E}/dt = 2.2\ (M/10^8) M_\odot$ year^{-1}, where M is the mass of the quasar in solar masses. Thus, a modest accretion rate can sustain even high-luminosity QSOs.

The argument that supermassive black holes reside in AGNs continues to get stronger. Probably the best direct evidence is provided by the kinematics of megamasers in the Seyfert 2 galaxy NGC 4268. The Keplerian orbital motions of the individual megamaser spots imply a central source mass of $3.6 \times 10^7 M_\odot$ within the inner 0.1 pc of the galaxy. The light-travel, time-delayed response of the broad emission lines to continuum variations also provides evidence for supermassive central sources: If the kinematics of the emission-line gas are dominated by the central source, then the central mass can be estimated from the virial theorem, i.e., $M \approx RV^2/G$, where R is the size of the line-emitting region, as determined by time delays, and V is the Doppler width of the line. Since the different lines arise primarily at different distances from the central ionizing source, multiple lines in a single AGN ought to show a relationship between line width and time delay $V \propto R^{-1/2}$; for the few AGNs in which the response of multiple lines has been measured, this virial relationship seems to hold and implies masses in the range

10^6–$10^8 M_\odot$. Additional evidence for supermassive black holes comes from X-ray spectroscopy of the iron Kα line at 6.4 keV; in some sources, this line is both redshifted (relative to the AGN itself) and asymmetric. Both of these characteristics can be relativistic signatures that arise within a few Schwarzschild radii ($R_s = GM/c^2$) of a black hole. This interpretation of the line profile characteristics is not yet firm, but since relativistic effects are involved, X-ray observations of the iron Kα line hold the most promise for actually proving the existence of supermassive black holes.

B. Accretion Disks

Fundamentally the luminosity of AGNs seems to derive from conversion of gravitational energy into radiation. We lack, however, a detailed understanding of how this process actually occurs. It is generally assumed that the radiation is emitted from an "accretion disk" that forms around the black hole. Accretion disks are expected to form as a mechanism for dissipating angular momentum, and are known to exist around stellar-mass compact objects. For a 10^8-$M_\odot$ black hole, the radiation from the inner accretion disk will thus be in the far-UV part of the spectrum (Section III.C).

Is there evidence that accretion disks do indeed exist in AGNs? Certainly, the spectral energy distributions of AGNs peak somewhere in the UV to soft X-ray range, as expected for their masses and mass accretion rates. But the observational details do not agree well with this simplest of accretion-disk theories. Expected signatures such as Lyman edges (due to onset of hydrogen opacity shortward of 912 Å) and continuum polarization are not observed. Also, the shape of the continuum in the UV/optical is a poor match to the simple prediction that the specific luminosity per unit frequency should follow the thin-disk prediction, $L_\nu \propto \nu^{1/3}$. On the other hand, in the case of NGC 7469, one of the best-studied variable Seyfert 1 galaxies across the UV/optical spectrum, brightness variations at longer wavelengths follow those at shorter wavelengths. The observed behavior is consistent with a thin-disk structure in which the variations are driven by radiation from an on-axis source (i.e., so-called "reprocessing models"): Variations are seen first at the shortest wavelengths, which arise in the inner, hotter regions of the disk, and later at longer wavelengths, which arise further out. However, it is also widely acknowledged that the thin-disk model is too natïve; the accretion-disk itself is probably immersed in a hot corona that inverse-Compton upscatters UV photons into X-rays (Section III.A.2), which in turn play a role in heating the disk and modifying its temperature structure. Detailed successful self-consistent models of the continuum source are still lacking.

C. Jets and Radio Structure

As noted earlier, quasars were first identified on the basis of their radio emission, which is synchrotron emission (Section III.A.2) from plasma originating in the nuclear regions and fed to extended "lobes" through "jets," which are long linear radio structures. The jets are collimated plasma outflows that are at least mildly relativistic in the most energetic sources. The electrons radiate over the whole electromagnetic spectrum. While most of the energy of a jet is radiated at high energies (detected at even TeV energies), jets have been most thoroughly studied at radio wavelengths on account of their high contrast in the radio and on account of the high angular resolution and sensitivity achievable with radio interferometric techniques.

Radio sources typically have a double-lobed morphology and are considerably larger than the host galaxy at optical wavelengths (i.e., they have sizes of hundreds of kiloparsecs). Details of this morphology are correlated with source luminosity: Lower power sources (Fanaroff-Riley type I, or FR I) are brighter in the central parts of the lobes, and higher power (FR II) sources are edge-brightened. When radio structures are observed in the plane of the sky, emission from the extended lobes is most prominent and these sources are referred to as "lobe-dominated." As the observer's line of sight gets closer to the jet axis, emission from the jet becomes relatively more prominent, and these are referred to as "core-dominated" sources. If one observes very close to the jet axis, the relativistically boosted and beamed emission from the jet source swamps the light from the host galaxy and the inner regions of the AGN; such sources are highly variable and are known as "blazars." In some core-dominated sources, shock fronts are observed within the jet structures, and these seem to propagate at speeds in excess of the speed of light. This apparent superluminal motion is a consequence of relativistic motion within a few degrees of the line of sight, basically because in the observer's reference frame the fast moving shock is moving towards us almost as fast as its radiation. The mechanisms by which jets are formed and collimated are not understood, but they are probably linked to the accretion-disk structure by magnetic fields.

D. Alternative Models

The early history of quasar astronomy featured lively debates on their fundamental nature. Early arguments were focused on the nature of the emission-line redshifts. If these were not of cosmological origin, then the quasars might be closer than supposed, thus greatly reducing their luminosities and concomitant high energy densities.

Of course, the role of beamed radiation from jets was not fully appreciated in the early years, and the quasar energy problem is now regarded as "solved" at least qualitatively through accretion by supermassive black holes. Within the last ten years, the only serious challenge to the black-hole paradigm has been "nuclear starburst" models, which fail to account for some quasar characteristics such as rapid X-ray variability. Nevertheless, a small minority still maintain that quasar redshifts are of noncosmological origin. Evidence cited includes apparent angular correlations between quasar positions and those of nearby bright galaxies and a suggested relationship between them. Most astronomers discount these argument because, for example, of direct detection of the host galaxies of quasars which show that the quasars and their host galaxies do indeed have the same redshifts.

III. OBSERVED PROPERTIES OF QUASARS

As noted earlier, AGNs emit radiation across the entire electromagnetic spectrum, from the lowest detectable radio frequencies through high-energy X-rays. Gamma rays are detected in some blazars, and the highest-energy γ-rays are detected indirectly by the Cerenkov showers that they cause in the Earth's atmosphere.

Typical spectral energy distributions (SEDs) are shown in Fig. 1 for both radio-loud and radio-quiet nonblazar AGN in units showing the energy emitted at a given frequency. An almost constant power per frequency decade is emitted from 100 μm to at least 100 keV. To a first approximation, the SED can be approximated as power law $F_\nu \propto \nu^{-\alpha}$, where $0 \lesssim \alpha \lesssim 1$. Deviations from this simple form are evident in various spectral bands; from high to low photon energies, the notable broad-band spectral features and their probable origin are

1. X-ray emission ($\nu \approx 10^{18}$ Hz). X-rays are predominantly a nonthermal emission that arises in the immediate vicinity of the black hole or in relativistic jets. Soft X-rays are those in the energy range $0.1 \lesssim h\nu \lesssim 2$ keV, those at higher energies are hard X-rays.
2. The "big blue bump" ($\nu \approx 10^{15}$–10^{17} Hz). As noted earlier, this is the feature that is most plausibly identified with the accretion disk. It is observed in the short-wavelength optical (~4000 Å) through the ultraviolet (i.e., wavelengths as short as 912 Å). It is unobservable at shorter wavelengths due to the opacity of our own galaxy to hydrogen-ionizing photons in the extreme ultraviolet (EUV).
3. The infrared bump ($\nu \approx 10^{13}$–10^{14} Hz). AGN emission at infrared wavelengths seems to be dominated by thermal emission from dust in the immediate vicinity of the active nucleus. The thermal dust emission spectrum turns over at short infrared wavelengths, which leaves a local minimum in AGN SEDs known as the 1-μm inflection.
4. The submillimeter break ($\nu \approx 5 \times 10^{12}$ Hz). This feature occurs where thermal emission from dust grains becomes highly inefficient.

The two gaps in the SEDs shown in Fig. 1 are due (1) to the absorption by the neutral hydrogen in our galaxy (the EUV gap) and (2) to absorption by water vapor in the Earth's atmosphere (the IR-mm gap). The observed AGN properties are outlined in more detail below. In the case of blazars, the SEDs show simpler structure and can peak in virtually any part of the spectrum. The emission from blazars arises in relativistic jets that are directed approximately in the direction of the observer.

A. Continuum Emission

1. Gamma Rays

Strong γ-ray emission is detected in blazars only. Gamma-ray emission is strongly correlated with radio emission and radio variability. The γ-rays are almost certainly relativistically beamed emission that arises in jets (Section II.C), and therefore are seen only when the jet is directed towards the observer. The γ-ray emission is also highly variable. Gamma rays are probably the result of inverse-Compton upscattering of lower energy photons. The origin of the seed photons is not known, but might be either the low-frequency synchrotron photons or photons from the accretion disk.

2. X-Rays

X-ray emission is a ubiquitous property of AGNs. The amount of energy emitted in the X-rays is a good fraction of the total bolometric luminosity of a quasar (Fig. 1). As noted above, the X-ray spectrum is usually described as a power law with a spectral index typically around $\alpha \approx 0.9$. Superposed on top of the power law are a hard tail at energies higher than several keV which is thought to be due to reflection from an ionized gas and a soft excess at energies less than 2 keV which may be the high-energy tail of the big blue bump. Emission features are sometimes seen, notably the iron Kα line at about 6.4 keV, but commonly only in lower luminosity AGNs. In some cases, there is evidence of emission lines due to ions of iron, oxygen, and other elements. Due to absorption in our own galaxy, X-ray spectra generally turn over at low energies (~0.1 keV). Absorption features are sometimes seen in soft X-ray spectra, arising in the AGNs themselves.

Spectral features are absent at the highest energies, which completely ionize all abundant species (up through iron). The X-ray flux from quasars varies on time scales of days or hours and in some cases even minutes. This suggests that the X-ray emission originates very close to the central black hole.

The power-law nature of the spectrum indicates that the X-ray emission mechanism is primarily nonthermal. The physical processes involved could be one or more of the following nonthermal processes that can naturally produce a power-law spectrum:

1. *Synchrotron radiation*. Also known as magneto-bremsstrahlung, synchrotron radiation is emitted by relativistic electrons as they spiral around magnetic field lines.
2. *Inverse-Compton radiation*. Relativistic electrons lose energy to photons that scatter off them. The "seed photons" are lower energy (infrared, for example) radiation, perhaps from the accretion disk.
3. *Synchrotron self-Compton radiation*. This combines the other two processes: Radio photons generated by the synchrotron process are upscattered into the X-ray photons by further interactions with the relativistic electrons.

3. UV/Optical

The big blue bump (BBB) dominates the bolometric energy output of nonblazar AGNs. This feature is widely believed to be optically thick thermal emission from a thin accretion disk, generated by viscous heating in the differentially rotating disk annuli. An alternative explanation is that the emission is optically thin free–free emission (bremsstrahlung); while this better accounts for the observed optical-UV SED slope ($\alpha \approx 0.3$), this requires an emitting volume that is probably too large to explain rapid continuum variability.

The classical "thin accretion disk" is geometrically thin but optically thick (i.e., opaque) and thus radiates locally more or less like a blackbody. The temperature structure not too close to the inner radius is $T(r) = 6.3 \times 10^5 \, (dM/dt)_{\rm E}^{1/4} (M/10^8)^{-1/4} \, (r/R_{\rm s})^{-3/4}$ K, where $(dM/dt)_{\rm E}$ is the mass accretion rate in units of the Eddington rate. For a 10^8-$M_\odot$ black hole, the radiation from the inner accretion disk will thus be in the far-UV part of the spectrum.

Each annular region of the accretion disk radiates as a blackbody with the local temperature, and the sum over all the annuli is thought to make up the observed BBB. For the expected disk temperatures ($\sim 10^5$–10^6 K) the emission should peak in the unobservable EUV region, and the observed SEDs appear to do just that (Fig. 1). More complicated models are required to account for the high-energy emission: For example, there is probably a hot corona overlying the disk, though it is probably rather patchy (covering about 20% or so of the disk). The corona inverse-Compton scatters UV-optical radiation from the disk to produce the X-rays. Some of the X-rays, in turn, are absorbed by the accretion disk and increase its temperature locally. The interaction between the disk and corona is a complicated problem in magnetohydrodynamics, not unlike the problem of the solar corona, and is not well understood.

4. Infrared Emission and Thermal Dust Reprocessing

Figure 1 shows an inflection point in the SEDs of nonblazar AGNs at 1 μm, between the BBB and the broad infrared bump, which peaks at $\sim$25–60 μm. Sharp cutoffs between 100 μm and 1 mm are typically observed in radio-quiet quasars. The IR bump is due to thermal emission by warm extranuclear dust. The dust absorbs higher energy photons from the AGN and re-radiates in the infrared. The evidence for a thermal origin is severalfold:

1. Constancy of location is demonstrated by the inflection point in the spectrum. Dust fully exposed to the intense UV radiation of an AGN will be destroyed once it is heated to the sublimation point, which is about 2000 K for graphite grains. The hottest dust will be closest to the AGN and near the grain sublimation temperature. This leads naturally to a high-energy cut-off at 1 μm.
2. In a few AGNs, the infrared continuum apparently varies in response to changes in the UV continuum, with a time delay that is attributable to light-travel time between the UV source (the accretion disk) and the IR-emitting dust. The grain sublimation radius is consistent with the light-travel time measured.
3. The far-IR spectra of AGNs show little or no variability on short time scales, consistent with the cooler dust residing at larger distances from the UV source.
4. The steepness of the submillimeter break, typically a 5- to 6-decade drop in specific luminosity in radio-quiet quasars, implies a thermal origin. The spectral index of the break is sometimes very steep, with $\alpha < -2.5$. Synchrotron self-absorption cannot yield such steep spectral indices, while dust emission can.

In core-dominated radio-loud quasars (i.e., nearly face-on or blazar-type objects), the IR continuum is dominated by the high-energy tail of the synchrotron emission. The far-IR emission is anisotropically emitted, beamed along with the radio emission. In the highly inclined, lobe-dominated, radio-loud quasars, the radio and IR continua are not directly related and a discontinuity is evident in their SEDs.

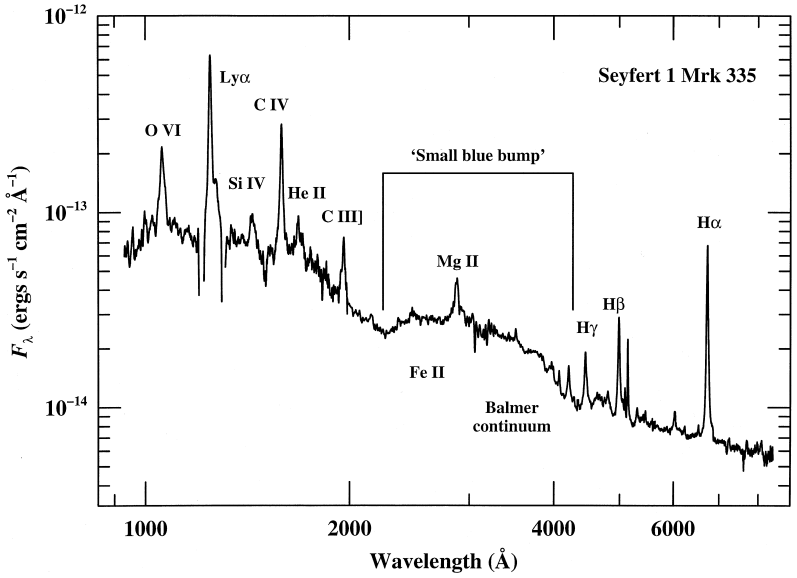


FIGURE 2 The UV/optical spectrum of the Seyfert 1 galaxy Markarian 335. Strong, broad emission lines that cover a wide range of ionization are characteristic of the spectra of active galactic nuclei. Narrow emission lines (such as the [O III] 11 4959, 5007 doublet seen immediately longward of the hydrogen Hb 1 4861 line) are also seen in AGN spectra. The feature labeled "small blue bump" is comprised of blended Fe II broad lines and Balmer continuum emission. [Adapted from Peterson (1997), based on data from Zheng, W. *et al.* (1995). *Astrophys. J.* **444,** 632.]

B. Emission Lines

1. Basic Properties

The predominant features of the ultraviolet and optical spectra of nonblazar AGNs are strong, broad emission lines (Fig. 2). The emission lines are produced by photoionization of the line-emitting gas by high-energy continuum radiation from the AGN.

Emission lines in AGN are classified as narrow or broad based on their line widths, usually as characterized by the full width at half maximum (FWHM) expressed as a Doppler width in the rest frame of the AGN. The narrow lines (FWHM in the range 500 to 1000 km s^{-1}) are emitted by an extended, low-density ($\leq 10^6$ cm^{-3}) narrow-line region (NLR) typically at distances $\sim 10^2$–10^4 pc from the central source. The characteristic velocity dispersion of the NLR gas is similar to or only somewhat larger than that of the stellar component in the host galaxy. In the most nearby AGN, the NLR is spatially resolved, showing a biconical region that is illuminated by the central continuum source.

The NLR is the most distant and extended line-emitting region where photoionization by the central source dominates over other heating processes.

The broad emission lines (FWHM in the range $\sim$1000 to a few times 10,000 km s^{-1}) arise in a compact, higher density ($\sim 10^{11}$ cm^{-3}) broad-line region (BLR). On the basis of simple energetic considerations, the BLR intercepts about 10% of the radiation emitted by the central source. Because the BLR is comprised of high-density gas (so that the recombination time scale is very short, and the gas can therefore respond quickly to a change in ionizing radiation) and is both optically thick and close to the continuum source, the broad emission lines vary in flux in response to variations of the continuum. The emission-line response, however, is time delayed relative to the continuum variations on account of light travel-time effects within the BLR. The characteristic time scale for emission-line response is days to months, depending on (1) which emission lines are observed (since in a given source the highest ionization lines respond more rapidly than the lower ionization lines), and (2) the luminosity of the source, since the BLR size is larger in more luminous sources: The size of the hydrogen Hβ λ4861 emitting region, for example, scales approximately as $R = 30\,(L/10^{44})^{1/2}$ light days, where L is the optical continuum luminosity in ergs s^{-1}.

The BLR is usually assumed to be comprised of a large number of discrete clouds of gas. The relative strengths of various emission lines (see below) indicate that the line-emitting gas is at a temperature of about 20,000 K, but if the total line widths were interpreted as purely thermal broadening, unrealistically high temperatures ($T > 10^9$ K) would be inferred. This suggests that the line widths are due to bulk supersonic motions of individual clouds that each emit line radiation with much smaller widths (about 10 km s^{-1}, the thermal width for more typical nebular temperatures). The smoothness of the broad-line profiles

implies that the number of line emitting clouds must be very large for stochastic clumping to be negligible.

While the BLR kinematics and structure are largely unknown, there is fairly good evidence that the gas is gravitationally bound to the central source; that is, the gas is probably orbiting the central source and is specifically unlikely to represent pure radiatively driven outflow. Outflow is still certainly possible, although either the radial motion of the gas is small compared to the orbital motion (i.e., the gas spirals outward) or most of the line emission comes for gas that is moving very close to escape velocity. If the line-emitting clouds are gravitationally bound to the source (i.e., in "virial motion"), then the central mass can be determined from the virial theorem, as described in Section II.B. The broad emission lines are also a means to study the important but unobservable EUV region as the BLR reprocesses the EUV photons emitted by the hottest parts of the accretion disk into the observable UV and optical lines.

2. Gas Densities, Temperatures, and Metallicities

As noted above, the line emission is primarily a result of photoionization by the central source. Heating of the gas by shocks or particle collisions is less important. This is evident by the relatively low, constant gas temperatures ($T \approx 10^4$ K) and by the large range in ionization levels indicated by the emitted line spectra (Fig. 2). Shock-heated gas, on the contrary, shows a narrow range of ionizations with strong emission from collisionally excited lines with $T \gtrsim 30{,}000$ K. These signatures are not commonly observed.

Line ratios are useful for determining the density and temperature in the low-density NLR gas. Electron densities n_e can be measured using "forbidden" line-doublets arising from different levels of the same multiplet, such as [OII] $\lambda 3729/\lambda 3926$ and [SII] $\lambda 6716/\lambda 6731$. The NLR densities are found to be in the range 10^2–10^4 cm^{-3}, typically with $n_e \approx 2 \times 10^3$ cm^{-3}. Flux ratios of lines from the same ion with very different excitation potentials are temperature sensitive. For AGNs, the best line ratio for temperature determination is [OIII] $\lambda\lambda 4959, 5007/$[OIII] $\lambda 4363$, as all of these lines are relatively strong. NLR electron temperatures measured this way are in the range $(1\text{–}2.5) \times 10^4$ K, with an average value $T_e \approx 1.6 \times 10^4$ K. The BLR density is too high for these line ratios to be useful temperature diagnostics; the critical density for these lines (above which collisional de-excitation of the upper state becomes important and the line emissivity grows only linearly with density) is about 10^6 cm^{-3}, so these "forbidden" lines are very weak ("collisionally suppressed") in broad-line spectra. There are also no good density diagnostics for the BLR, but the most successful photoionization equilibrium models suggest densities in the range $n_e \approx 10^9$–10^{12} cm^{-3} and temperatures of about 20,000 K. The presence of CIII $\lambda 977$ sets a strong upper limit of $T_e \approx 25{,}000$ K, ruling out shock heating.

The most abundance-sensitive line ratios, least affected by the density, temperature, and local radiation field in the gas, involve lines of different elements that are formed under similar conditions, preferrably in the same region. For example, NV $\lambda 1240/$CIV $\lambda 1549$ is representative of the relative C/N abundance in the BLR, while the NV/HeII $\lambda 1640$ flux ratio can provide robust lower limits on the N/He abundance and thus the overall metallicity. This holds great promise for mapping the early metallicity production in the Universe and its cosmological evolution in galactic nuclei by observing quasars at low to very high redshifts (Section VI.D). Studies of C/N, for example, show that the metallicities increase with redshift above the approximately solar values measured in nearby AGN. This may be a natural consequence of the denser and more massive systems present at earlier epochs.

3. Emission-Line Profiles

Active galactic nuclei emission lines have non-Gaussian profiles. The narrow lines have more pronounced bases than a Gaussian with the same half width, and the broad wings tend to be blueward asymmetric (i.e., with more emission short of line center) while the line core is more symmetric. This indicates that the highest velocity gas has a net motion towards us and/or that the line emission is subject to some type of obscuration, such as absorption by dust, that suppresses the red wing. Indeed, dust-sensitive emission-line flux ratios, such as Hα/Hβ, confirm the presence of dust in the NLR. The strength of the asymmetric wing is correlated with the presence of nuclear radio jets. The high velocity wings on the narrow emission lines probably represent material that has been entrained in radio outflows from the nucleus.

The situation is more complicated for the broad emission lines, which have approximately logarithmic profiles (i.e., $F(\Delta V) \propto -\ln \delta V$, where ΔV is the Doppler velocity relative to the line center), with subtle differences seen among objects. The observed profile asymmetries reveal a complex picture but appear somewhat related to the radio properties of the sources. Radio-loud quasars exhibit predominantly redward (i.e., towards longer wavelengths) asymmetries, and radio-quiet quasars show mostly blueward asymmetries, yet these are only statistical trends, not rigid rules. The broad-line asymmetries are not understood but suggest either that the BLR has different kinematics in different types of AGNs or that the BLR comprise multiple, distinct physical components and different components dominate in different types of AGNs.

The low-ionization lines, the Balmer lines included, have emission-line peaks that are approximately at the AGN systemic redshift, defined by the forbidden lines or absorption lines in the host galaxy. The high-ionization lines are blueshifted relative to the low-ionization lines. The velocity shift is observed to be stronger for face-on radio-loud quasars, which are also the most Doppler-boosted sources. There is no definitive picture to explain this, but it indicates that the bulk motion of the low-ionization gas may be different from that of the high-ionization gas in the BLR. Detailed interpretations are strongly dependent on the assumptions of the unknown BLR geometry and structure.

While the line profiles of the highest luminosity sources tend to be fairly smooth and structureless (on small scales), the broad-line profiles of lower luminosity objects can be quite complex. Some emission lines, usually but not always in low-luminosity Seyfert galaxies, have a double-peaked structure, with the two peaks displaced approximately symmetrically around the systemic redshift of the AGN. As noted above, emission-line fluxes vary in response to changes in the continuum brightness. Emission-line profiles are also observed to vary with time, but the line profile changes have not been successfully linked to any other property of AGNs.

4. Continuum/Emission-Line Correlations

Correlations between various emission-line and continuum properties of AGNs are of great potential importance determining the predominant physical properties at work in these sources. Many such correlations have been found. Perhaps the most important and well-known is the "Baldwin Effect," an inverse relationship between the broad-line equivalent width* of CIV λ1549 and the underlying continuum luminosity. This effect is observed over more than 6 orders of magnitude in AGN continuum luminosity. Part of its importance derives from its potential as a cosmological "standard candle." If we can deduce the source luminosity from the observed line equivalent width, then a comparison with the observed flux allows a determination of the "luminosity distance," which in turn depends on various cosmological parameters. It is thus possible to infer the values of cosmological parameters if the luminosity of an object is known from its spectrum. However, the observed Baldwin relationship shows significant scatter, the origin of which is not yet completely identified, although source variability and inclination effects seem to contribute. The origin of the Baldwin effect itself is not understood. Possible explanations include that with increasing luminosity (1) the covering factor (fraction of the continuum radiation intercepted by the BLR) decreases, (2) the ratio of ionizing photon density to electron density (the "ionization parameter" that determines the basic characteristics of the emission-line spectrum) decreases, (3) the metallicity increases, (4) the SED peak shifts to lower energies, and (5) the accretion-disk inclination systematically decreases.

If, as inferred from several considerations above, the central continuum source is responsible for photoionizing the BLR and NLR gas, certain intimate relationships between the ionizing continuum and the emission lines are expected. As most of the ionizing continuum lies in the EUV and soft X-ray regime, this is not currently fully explored, though we can try to infer some characteristics of the EUV spectrum from the observable UV/optical spectrum. Operationally this amounts to searching for correlations among various continuum and emission-line spectral properties. One way to do this is by principal component analysis, which can be used to identify complex sets of correlations that appear in many AGNs. The best known set of these correlations, known as the Boroson–Green Eigenvector 1, reveals that as the strength of the optical broad-line FeII emission increases, the narrow-line [OIII] λ5007 strength decreases, and the Hβ line width and profile asymmetry both decrease. At one extreme of the Eigenvector 1 properties are the narrow-line Seyfert 1 galaxies, which have unusually narrow (FWHM $\lesssim$ 2000 km s^{-1}) Balmer lines and strong Fe II emission. These sources also show very rapid and strong X-ray variability. The physical basis of these correlations is not understood, but it is anticipated that they are driven primarily by one or two fundamental parameters, such as central black-hole mass and black-hole accretion rate. For example, the best current explanation for the narrow-line Seyfert 1 phenomenon is that these are relatively low-mass systems that are accreting at a higher than normal rate.

*The equivalent width of a line is essentially the amount of local continuum emission, measured in wavelength units, that would give the same total flux as the emission line. Equivalent width is thus a measure of line strength, not line width.

C. Absorption in AGNs

Very distant quasars often show resonance absorption lines (e.g., hydrogen Lyα λ1215, C IV $\lambda\lambda$1548, 1550, Mg II $\lambda\lambda$2795, 2802) at redshifts smaller than the quasar emission-line redshift. In most cases, these arise in material that is unrelated to the quasar and is simply gas at lower cosmological redshift that happens to fall along our line of sight to the quasar (see Section VI.B). Some absorption features, however, especially in lower redshift objects, clearly arise in the immediate vicinity of the AGN.

Narrow (FWHM $<$ few hundred km s^{-1}) "intrinsic" resonance-line absorption is found within a velocity $\sim$5000 km s^{-1} of the AGN systemic redshift and is often

superposed on the emission-line profiles. Such features are almost always blueshifted relative to the emission lines and thus represent some kind of outflow from the source. These features are very common in the UV spectra of Seyfert galaxies.

Broad absorption lines (BALs) constitute a more spectacular class of absorption as the line widths can reach several 10^4 km s^{-1}. The absorption troughs are always located blueward of the emission lines and can have P-Cygni profiles,* and in some cases the absorption appears to be completely detached from the broad emission lines, indicating that the absorption does not simply occur in the broad emission-line gas along the line of sight to the continuum source. Spectroscopy and polarization studies suggest that BALs arise in high-density, high-velocity, optically thick equatorial outflows that cover $\sim$10% of the continuum source, as BALs are detected in only $\sim$10% of QSOs (Section VI.B). All known BAL quasars were radio-quiet until the *FIRST* survey, which is based on deep radio observations, began turning up significant numbers with radio emission. The BAL objects yield information on the inner regions of quasars.

High-velocity ($V \gtrsim 3000$ km s^{-1}) absorption lines, BALs included, are only observed in high-luminosity QSOs, never in Seyferts. This indicates that the energetics of the outflow are closely related to the luminosity of the AGN.

Absorption features that arise in the nuclear regions of AGNs are also seen in X-ray spectra. Their presence is strongly correlated with the narrow-absorption features in the UV. The X-ray absorption arises in fairly highly ionized gas referred to as "warm absorbers." Before the launch of the Chandra X-Ray Observatory, the strongest absorption features observed in the X-ray band were the edges of ionized carbon, oxygen, neon, and iron. These demonstrate that there is a large amount ionized matter in the nuclear regions of AGNs, possibly just outside the BLR. The Chandra X-Ray Observatory has now observed absorption lines due to highly ionized oxygen and other species. These lines provide more powerful diagnostic of the physical conditions of matter close to the nuclear black hole than the absorption lines observed in the UV.

D. Dust

Dust is known to be present in AGNs, as noted above, although little is known about its properties or distribution in AGNs. The dust extinction as a function of energy depends on properties such as grain size and composition, but is observed to consistently increase towards higher energies in the UV-optical region. AGN spectra lack the strong graphite absorption at 2175 Å that is often seen in galactic sources. Silicate dust absorption (10 μm) and emission (19 μm) features are also rarely observed, except in a few Seyfert galaxies. Emission-line flux ratios, such as Balmer line Hα/Hβ, show that dust is mixed with the NLR gas. Dust is destroyed inside the BLR by the energetic photons, unless it is strongly shielded, so that, relative to the galactic interstellar medium, the BLR gas is less depleted in heavier elements: An order of magnitude enhancement of elements, such as iron, silicon, magnesium, aluminum, and carbon, is found in the BLR compared to galactic gaseous nebulae containing dust grains.

As discussed in Section IV.B, the central regions of AGNs are thought to be at the center of a dusty torus that obscures the view of the AGN in the equatorial plane of the system. The large columns of cold dust and gas that are sometimes detected along the line of sight may originate in this torus.

E. Radio-Loud/-Quiet Dichotomy

The radio emission from radio-quiet quasars is several orders of magnitudes fainter than that from radio-loud quasars (Sections I and II.C), and their weak radio structures, observed so far, are either amorphous or show signs of small jet-like structures. These structures may be due to weak collimation of the radio plasma. The powerful, large-scale radio structures in radio-loud quasars are, on the contrary, believed to be due to a strong collimation of the radio plasma, probably by magnetic fields.

It has long been a puzzle why the SEDs of radio-loud and radio-quiet classes are so similar over most of the electromagnetic spectrum, yet display such dramatically different radio properties. However, all across the spectrum detailed differences exist. The radio-loud quasars (1) have nonthermal synchrotron emission contributing to their IR emission, (2) are stronger X-ray sources, and (3) have X-ray spectra that are flatter (i.e., spectral index α close to zero). The excess X-ray emission may be linked to the nonthermal radio emission. The radio-quiet quasars tend to have (1) stronger optical FeII emission lines, (2) weaker forbidden narrow lines (such as [OIII] $\lambda\lambda$4959, 5007), and (3) a much higher frequency of occurrence of the BAL phenomenon. How these differences relate more directly to the radio-luminosity and the central radio-generating mechanism is still unknown.

Most QSOs, irrespective of their radio power, are now found to be hosted by elliptical galaxies. This is consistent with the lack of luminosity differences between

*The spectra of stars with massive outflowing winds, such as P-Cygni, have strong emission lines that have self-absorption due to the cooler diffuse gas that is seen projected against the background star. The absorption occurs in the blueshifted gas that is superposed against the star, so the emission lines are strongly self-absorbed in their blue wings.

the two QSO classes (the lower-luminosity Seyferts are generally found in spiral galaxies), but leaves open the issue of why and how radio-loud quasars can produce the powerful, large-scale radio emission, while radio-quiet quasars cannot. This indicates that it is not the host galaxy properties that play a vital role for the generation of powerful radio emission, but rather some fundamental properties and/or structures of the AGN central regions, such as the combined properties of the magnetic field, the accretion disk.

IV. STRUCTURE OF QUASARS

The inner structure of AGNs is far too small to be observed directly: Even in the case of the nearest AGNs, the BLR and accretion-disk structure project to angular sizes smaller than tens of microarcseconds. Milliarcsecond radio structures in jets can be observed using very-long-baseline interferometry (VLBI), and some narrow-line region structure can be observed with Hubble Space Telescope (with angular resolution of about 0.05 arcseconds) or even with ground-based telescopes (with angular resolution generally no better than about 0.3 arcseconds). Thus, much of what we know about the inner structure of AGNs is inferred indirectly from, for example, polarization and variability studies.

A. Interrelations Among Spectral Components

While the UV-optical emission in AGNs is generally believed to arise in an accretion disk, the origin of the X-rays is not understood. As noted earlier, strong, rapid X-ray variability implies that the X-ray emitting region is very compact. Furthermore, certain X-ray spectral features, specifically the iron Kα line and the Compton reflection hump, suggest that about 50% of the X-ray energy is reprocessed (i.e., absorbed and re-emitted without additional energy input) by an optically thick plasma. This has led to heuristic models in which the X-ray emitting region lies on the accretion disk axis and irradiates the disk from above. This suggests that the variations in the X-ray band will induce variations in the emitted spectrum of the accretion disk. In this case, the X-ray variations drive the variations at longer wavelengths, with a time delay due to light travel-time effects. The expectation is that the UV variations that arise in the inner hotter part of the accretion disk should thus lead the longer wavelength variations that arise in the cooler outer disk. This effect has been detected in only one AGN, the Seyfert 1 galaxy NGC 7469, and the interpretation is somewhat problematic because the observed hard X-ray variations are poorly correlated with the UV-optical variations. In any case, the UV-optical variations are closely coupled in time, so closely that variations attributable to disk instabilities (which propagate on the much slower viscous time scale) as seen in accretion disks around stellar-mass collapsed objects can be ruled out. Whatever the agent that causes continuum variability, the signal must surely be propagated at light speed.

Also, as noted earlier, IR continuum variations follow those in the UV-optical on a time scale consistent with light-travel time within the dust sublimation radius. In other words, the IR flux arises in dust as close to the AGN as it can be and still survive, and this dust absorbs AGN continuum radiation and reprocesses it into IR emission. Variations at the shortest IR wavelengths lead those at longer IR wavelengths, consistent with the hottest gas arising closest to the UV/optical continuum source.

B. Dusty Tori and AGN Unification Models

In Section I.B, we briefly described a rich and varied AGN taxonomy. Given the similarities and differences among the various types of AGNs, we need to ask whether all these sources are intrinsically different from one other, or simply appear to be different when viewed in different ways. A long-standing question is why Seyfert 2 galaxies do not have broad lines: Is the BLR actually missing in these sources, or is it somehow hidden from our view? In at least some Seyfert 2s, the latter has been shown to be the case. The polarized UV/optical spectra of these Seyfert 2 galaxies show broad components. The simplest interpretation of this phenomenon is that our direct line of sight to the continuum source is blocked by dust. Light from the central source and the BLR is seen only because it is scattered or reflected; the wavelength-independent continuum polarization observed in the best-studied sources suggests that the scattering is done by electrons rather than by gas. This is the basis of what are now known as AGN unification models. In these models, all the radio-quiet AGNs are intrinsically similar to each other, but their external appearance depends on the observer's viewing angle. The intrinsic differences among various AGNs are then, in the most optimistic case, limited to luminosity (which presumably depends on black-hole mass) and radio-loudness (which may be related to a parameter such as black-hole spin rate). It is assumed that the central regions of AGNs are at the center of a thick torus of gas and dust, as shown schematically in Fig. 3: When one observes the AGN close to the torus axis, one sees a Seyfert 1 spectrum. When one views the AGN close to the torus plane, the nucleus is seen only in scattered light, and a Seyfert 2 spectrum is observed.

A similar unification exists for radio-loud objects. Blazars can be neatly fit into this unification scheme by supposing that they represent AGNs observed directly

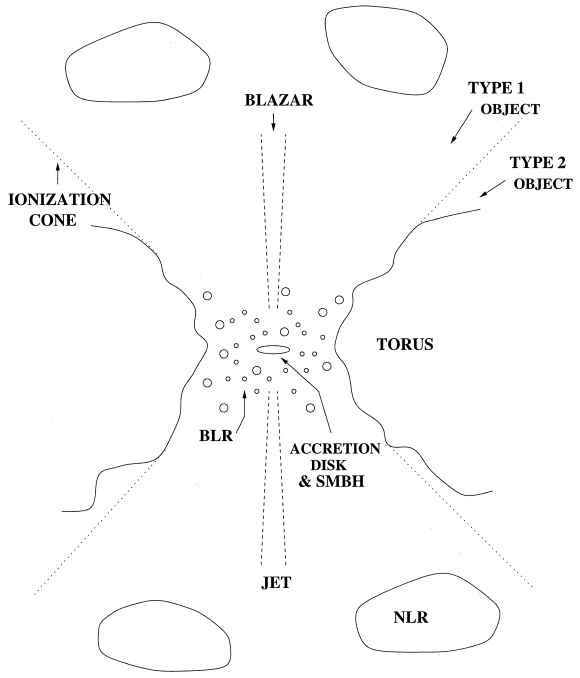


FIGURE 3 A schematic view of the current paradigm for inner regions of an AGN. The main components are labeled in the bottom half of the figure. At the center is a supermassive black hole (SMBH) and an accretion disk that accounts for most of the observed continuum radiation. Surrounding this is a system of fairly high-density clouds, the broad-line region (BLR), that accounts for the strong broad-emission features, such as those seen in Fig. 2. This system is embedded in the center of torus that is opaque in the UV/optical and near IR. The narrow-line region (NLR) is a system of lower density clouds on scales comparable to or larger than the torus. Radiation from the accretion disk can only drive emission lines in clouds that are not shielded from the ionizing radiation by the countinuum source (i.e., those clouds within the "ionization cone" boundaries, which are shown in the figure as dashed lines). In radio-loud objects, there is a jet structure comprised of outflowing plasma, along the accretion-disk axis. Labels in the top half of the figure show the AGN classification that will result based on different viewing angles: a radio-loud AGN with the jet pointed at the observer will likely be classified as a blazar. Sources observed off-axis out with an unobscured view of the nucleus will be type 1 sources; those in wihch the continuum source and BLR connot be observed directly will be type 2 sources.

down the axis of the system so that the observed radiation is dominated by the beamed emission from the relativistic jet. Furthermore, core-dominated radio morphologies represent sources seen close to their axes, but lobe-dominated radio morphologies are those in which the radio axis is closer to the plane of the sky.

Can unification models account for all the diversity seen among AGNs as simple orientation effects? At the present time it is unclear. While it is certainly true that *some* Seyfert 2 galaxies are intrinsically Seyfert 1 galaxies, it is not yet clear that *all* Seyfert 2 galaxies can be explained in terms of the standard unification model described here.

C. Environment of Quasars: Hosts and Neighbors

About 5 to 10% of galaxies contain an active nucleus as described here. What makes galaxies that harbor quasars different? Are all galaxies active at some point in time, or is nuclear activity an unusual phenomenon that occurs only in certain galaxies? On the other hand, does the presence of a quasar affect its host galaxy?

Most luminous quasars are observed to be present in massive galaxies. In both active and nonactive galaxies, the central black-hole mass correlates well with the luminosity of the nuclear bulge component of the host galaxy, though it is not yet clear whether or not the relationship is the same for active and nonactive galaxies.

Seyfert host galaxies are generally spiral galaxies, although it appears that the most luminous quasars reside in elliptical galaxies. Whether or not Seyfert galaxies have a disproportionate number of nearby companion galaxies is also an open issue: tidal interactions with neighboring galaxies can disturb the gravitational potential and result in matter flowing inwards towards the nucleus. It has thus been proposed that interactions between galaxies might be responsible for triggering activity in the galactic nuclei.

Only a decade ago, the key question about AGN environments might well have been "Why do some massive galaxies have black holes in their nuclei and others apparently do not?" Within the last several years, primarily as a result of Hubble Space Telescope observations, it has become clear that most galaxies have massive black holes at their centers, and the question might now be posed as "Why are the black holes in some galaxies active and not active in others?" The answers to these questions are of fundamental importance to understanding the AGN phenomenon. Because quasars are seen in such abundance at high redshifts (Section VI.A), it is widely believed that the formation of quasars is intimately related to the formation of galaxies. Nuclear black holes might form when galaxies form and go through a high-accretion phase during which they are quasars. At some point, the fueling process is arrested, and the quasar fades and eventually becomes a quiescent black hole.

V. FINDING QUASARS

A. General Principles and Historical Techniques

Compared to stars and galaxies, quasars are rare objects at optical wavelengths. Much effort has been devoted to developing effective ways to find quasars so that adequate samples can be assembled for studies of their nature and evolution. The general principle for finding quasars is to

make use of the differences in their SEDs compared to stars and galaxies; quasars are luminous over a much wider range in frequency (Fig. 1).

As mentioned in Section I.A, quasars were first identified because of their strong radio emission. Initial radio surveys produced catalogs of radio-loud quasars, an important class of quasars but one that represents only 5 to 10% of the optically selected population. Interestingly, as the sensitivity of radio telescopes improved, deep radio catalogs now contain many radio-quiet quasars, which have reduced radio emission but are not radio silent. Radio-selected quasars are important both because of their physical nature and because radio emission is virtually unaffected by dust clouds that can block the optical emission of quasars.

At optical wavelengths, quasars were first separated from stars by their "ultraviolet excess," i.e., quasars were strong ultraviolet emitters compared to most stars. This technique is most effective for redshifts up to about 2.2. At higher redshifts, the Lyα emission moves beyond the ultraviolet passband into redder bands and quasars lose their characteristic UV excess.

B. Multicolor Techniques

With the advent of digital sky surveys and computerized, multicolor techniques, it became possible to find quasars efficiently at redshifts greater than 2.2 by using filters that cover a wide range of wavelengths and applying new algorithms to separate foreground stars and distant quasars by the differences in their SEDs. This approach encounters the most difficulty around redshift 3, where the differences in the SEDs of quasars and local stars are less than at higher redshifts. At redshifts greater than 4, the continuum shortward of Lyα is significantly depressed by absorption from intervening clouds of neutral hydrogen, which makes quasar SEDs stand out from those of all stars except those with the lowest temperatures.

C. Slitless Spectroscopy

In the mid-1970s a different approach was developed for finding quasars that made use of low-resolution slitless spectra on which the emission lines of quasars could be detected directly. The technique was especially effective for quasars at high redshifts because Lyα is the strongest emission feature in quasar spectra and can be seen at optical wavelengths for redshifts between 2 and 7. As did the multicolor approach, the slitless spectrum technique benefited from improvements in detectors and digital survey techniques, and it has also been used successfully for objects with redshifts less than 2 because of SED information contained in the data.

It should be noted that all survey techniques for quasars have selection biases, and considerable effort is required to determine the selection efficiency of surveys and the corresponding corrections for incompleteness. However, this is now done with increasing accuracy for modern digital surveys.

D. X-Ray Surveys

The advent of powerful X-ray observatories enables a new and powerful technique for finding quasars based on their X-ray emission. Indeed, it can be argued that the strong X-ray emission of quasars and AGNs provides the best way to distinguish them from stars and normal galaxies. Historically this was of practical use only for the brightest objects, but the deepest X-ray surveys carried out with the *ROSAT* satellite now achieve higher surface densities of objects than the deepest optical surveys published to date. However, high-accuracy positions for the X-ray sources are critical for the identification of the correct source, because confusion problems are severe at such low flux levels. We may expect important new results both on the quasar and lower luminosity AGN populations and on the nature of faint X-ray sources as even deeper surveys with the *Chandra* and *XMM* observatories are carried out and followed up with spectroscopy and SED measurements at other wavelengths.

The optical/UV techniques of finding quasars may fail for quasars surrounded by large amount of obscuring matter. Dust reddens quasars, changing their SEDs and making them fainter in the UV. Such obscuration and reddening is could be present in a large fraction of quasars. Since the hard X-ray flux is relatively insensitive to absorption by dust, X-rays seem to be more useful in finding the "hidden" quasars. As a result, many new quasars have been discovered as a result of X-ray surveys. More recent missions such as *Chandra* and *XMM–Newton* are expected to find thousands of new quasars. In particular, hundreds of high-redshift quasars are expected to be found by the new X-ray telescopes. Also, some objects that are especially prominent in X-rays (e.g. narrow-line Seyfert 1 galaxies) are found more easily in X-ray surveys.

E. Current Status

The number of known quasars has increased rapidly in recent years owing to the development of the techniques described above and their application in large surveys. As of March 2000, the Verón catalog lists 18,104 quasars, BL Lac objects, and quasars. The Anglo-Australian 2dF Quasar Survey now has redshifts for 10,000 quasars in a program that is planned to observe more than 25,000 quasars for the study of their spatial distribution

and other properties. The Sloan Digital Sky Survey (SDSS) will cover 10,000 deg^2 of the sky and be the largest quasar survey ever done; it should obtain spectra of 100,000 quasars.

VI. QUASARS AND THE UNIVERSE

A. Evolution and Luminosity Function

Quasars provided our first view of objects in the Universe at redshifts greater than 0.5 and within two years of their discovery the first object, 3C 9, at a redshift greater than 2. The lookback time to 3C 9 was of order 10 billion years, or 80% of the age of the Universe. Quasars offered great potential as cosmological probes because of their high luminosities and large redshifts.

Schmidt noted in the late 1960s that the fraction of high-redshift quasars in his samples was unusually high. Upon analyzing their distribution in space, he showed that the quasar population was significantly nonuniform and increased significantly with redshift. The effect out to redshift 2 was dramatic: The derived space density was a thousand times greater than the local value. Alternatively, quasars at high redshift were significantly more luminous than their local counterparts. In any case, the evidence for evolution was very strong and provided one of the main arguments at the time in favor of Big Bang cosmology.

Subsequently, major efforts to determine the luminosity function of quasars and its evolution with redshift have been carried out. Large samples covering a wide range in redshift and luminosity and having well-understood selection properties are needed. Current results indicate that the space density of quasars reaches a maximum near redshift 3 and declines steeply toward higher redshift. A straightforward interpretation of the results is that we have detected the time of greatest quasar activity. However, this presumes that there is no significant amount of intervening absorption by dust clouds along the line of sight to the quasars. This is an important question because some quasars and nearby AGNs are known to be heavily obscured by dust clouds in their host galaxies. However, a survey of radio-selected quasars at high redshift, which are not sensitive to dust obscuration, also shows a peak in their space density similar to that of optically selected quasars.

In recent years, the advent of 8- to 10-m class telescopes and the success of programs such as the Hubble Deep Field have provided important new tools for discovering and studying the most distant objects in the Universe. The results for the first galaxies confirmed with slit spectroscopy to have redshifts greater than 5 were published in 1998. Then the first quasars with $z > 5$ were discovered in the following year. At the time of writing, the SDSS has just yielded a quasar with $z = 5.8$. The light from this object was emitted when the Universe was less than a billion years old. We may expect continuing important discoveries in the near future about the nature of the first galaxies and quasars to appear in the Universe.

B. Quasar Absorption Lines

Because of their great luminosity and high redshifts, quasars enable the detection and analysis of cold gas along their lines of sight to the Earth via absorption-line spectroscopy, gas that would in general not be observable otherwise. Furthermore, the lines of sight provide fair samples of the Universe. Thus, the study of quasar absorption lines is a major subject in itself. Absorption lines were found in all quasars with $z > 2$ because (1) the strong ultraviolet resonance lines of abundant ions in the intergalactic medium were redshifted to wavelengths that could be observed with ground-based telescopes, and (2) the number density of lines was found to increase with redshift. Subsequently, the ultraviolet capability of the Hubble Space Telescope has enabled extensive studies of absorption lines in quasars with redshifts less than 2.

There are three general classes of absorption-line systems, depending on the distance of the absorbing gas from the central region of the quasar. In the first class are those intrinsic absorption systems that arise in gas very close to the active nucleus and appear to be some kind of manifestation of the nuclear activity. This includes the BAL quasars discussed earlier (Section III.C.)

The second class, associated absorption-line systems, consists of quasars with narrower lines of elements such as hydrogen, carbon, and magnesium at redshifts close to that of the quasar itself. The lines are thought to be produced in gas that is either associated with the host galaxy or neighboring galaxies.

The third class, intervening systems, refers to the absorption lines produced by clouds of gas distant from the emitting quasar and unrelated to it. These systems, which are numerous, provide fundamental information about the nature and evolution of the intergalactic medium at redshifts from 0 to beyond 5. The intervening systems are themselves grouped into three different categories according to the strength of the absorbing systems.

The strongest absorption, with neutral hydrogen column densities in excess of 10^{20} atoms cm^{-2}, occurs in the damped Lyα systems, in which the Lyα absorption is strong enough to show damping wings in the line profile. This absorption arises when the line of sight traverses a disk galaxy or dense filaments of intergalactic gas. These absorption systems also show lines from heavier elements, including carbon, silicon, magnesium, and iron, among

others. The inferred abundance of metals is typically about one tenth solar, indicating that the gas has undergone less nuclear processing than is observed in nearby galaxies, which are being seen at a later stage of their evolution.

Intermediate absorption systems have a column density of approximately 10^{17} neutral hydrogen atoms cm^{-2}, high enough to show a discontinuity at the Lyman limit of hydrogen. They also have absorption lines from carbon, magnesium, and other metals and are believed to arise in the halos of galaxies. Their metal abundances can be as low as 1% solar.

Finally, the low column density systems are the most numerous and are referred to as the Lyα forest, because of the large number of lines that are seen in the spectrum shortward of the Lyα emission line. They arise in clouds of gas in the intergalactic medium with a range of column densities from about 10^{12} neutral hydrogen atoms cm^{-2} up to the densities of the intermediate systems. Their number density increases steeply with increasing redshift, and by redshift 3, for example, they produce a significant depression of the continuum emission shortward of the Lyα emission. Observations of the Lyα forest are being used in conjunction with theoretical modeling of the intergalactic medium for both cosmological studies and the development of large-scale structure in the Universe.

C. Gravitational Lenses

In 1979, Walsh, Carswell, and Weymann discovered the first gravitational lens Q0957 + 561, a quasar at $z = 1.41$ that showed two images separated by 6 arcsec. Gravitational lensing occurs when two objects are nearly perfectly aligned along the line of sight. The gravitational field of the nearer object bends the light of the background object and produces several effects, including multiple imaging and magnification of the brightness. Gravitational lensing can provide information on the expansion rate and geometry of the Universe and about the distribution of mass in the Universe, particularly dark matter.

D. Implications

The formation and evolution of quasars at high redshift is one part of the larger subject of the formation of structure in the Universe and the formation of galaxies. Quasars at high redshift mark the sites of the active black holes in the nuclei of galaxies. Furthermore, the spatial distribution and clustering properties of quasars are related to the development of large-scale structure in the Universe in models in which quasars (and galaxies) form at sites of high-density fluctuations in the distribution of dark matter. Advances in observational capabilities, which now enable both galaxies and quasars to be observed to redshifts beyond 5, and developments in the theory of galaxy and structure formation are enabling a unified study of all these topics. The large surveys mentioned above such as the SDSS and 2dF will provide very important observational data for these topics.

Quasars are related to cosmology in another way—they are a major source of the radiation that ionizes the intergalactic medium, which contains most of the hydrogen and helium in the Universe. Observations of the absorption line systems described above show that the IGM and the clouds producing the Lyα forest are highly ionized at high redshift. Otherwise absorption from neutral hydrogen would block the continuum light of quasars at wavelengths shortward of the Lyα emission (the Gunn–Peterson effect). A main research question is when did this reionization occur and what caused it—hot stars in galaxies or quasars and AGNs? In this view, reionization is connected with the appearance of the first luminous objects in the Universe. While it had been thought that quasars were the main sources of ionization at high redshift, the rapid decline in their space density at $z > 3$ and the subsequent discovery of numerous galaxies at $z \geq 3$ with rapid rates of star formation have made it apparent that ionization from hot stars is also an important source of ionizing radiation. Their total ionizing flux could exceed that of quasars. The relative contributions of quasars and hot stars to ionization at $z > 3$ are uncertain at present. Deep X-ray and optical surveys and the Next Generation Space Telescope are expected to provide valuable data for this problem.

Quasars are also connected to the chemical evolution of galaxies. The emission-line spectra of quasars show features from elements such as carbon, nitrogen, oxygen, helium, silicon, magnesium, and iron. The general similarity of the emission-line spectra over the redshift interval from 0 to 5 is striking. Although the derivation of chemical abundances from the spectra is a difficult topic, current evidence indicates that the abundances of elements such as nitrogen are solar or possibly greater, even at the highest redshifts. This is an indication that the emission-line gas was enriched by earlier generations of stars.

Finally at low redshifts, recent observations show that inactive black holes are common in the spheroids of nearby galaxies. They provide important information about the endpoints of quasar and AGN activity in galaxies, which occur when no material remains to be accreted by the central black hole. It is plausible that the decline in space density of quasars from redshift 2 to 0 is a consequence of the expansion of the Universe and the conversion of gas in galaxies to stars. The material available for accretion declines with advancing time. The masses and frequency of occurrence of black holes in nearby galaxies provide important constraints about the formation and growth of

the black holes with time and the efficiency of the accretion process.

SEE ALSO THE FOLLOWING ARTICLES

COSMOLOGY • GALACTIC STRUCTURE AND EVOLUTION • ULTRAVIOLET SPACE ASTRONOMY • X-RAY ASTRONOMY • X-RAY, SYNCHROTRON RADIATION, AND NEUTRON DIFFRACTION

BIBLIOGRAPHY

Antonucci, R. R. J. (1993). "Unified models for active galactic nuclei and quasars," *Ann. Rev. Astronomy Astrophys*. **31,** 473.

Blandford, R. D., Netzer, H., and Woltjer, L. (1990). *In* "Active Galactic Nuclei" (T. J.-L. Courvoisier and M. Mayor, eds.), Spring-Verlag, Berlin.

Hamann, F., and Ferland, G. (1999). "Elemental abundances in quasistellar objects: star formation and galactic nuclear evolution at high redshifts," *Ann. Rev. Astronomy Astrophys*. **37,** 487.

Kembhavi, A. K., and Narlikar, J. V. (1999). "Quasars and Active Galactic Nuclei: An Introduction," Cambridge University Press, Cambridge, U.K.

Mushotzky, R. F., Done, C., and Pounds, K. A. (1993). "X-Ray spectra and time variability of active galactic nuclei," *Ann. Rev. Astronomy Astrophys*. **31,** 717.

Osterbrock, D. E. (1989). "Astrophysics of Gaseous Nebulae and Active Galactic Nuclei," University Science Books, Mill Valley.

Peterson, B. M. (1997). "An Introduction to Active Galactic Nuclei," Cambridge University Press, Cambridge, U.K.

Rees, M. J. (1984). "Black hole models for active galactic nuclei," *Ann. Rev. Astronomy Astrophys*. **22,** 471.

Urry, C. M., and Padovani, P. (1995). "Unified schemes for radio-loud active galactic nuclei," *Publ. Astronom. Soc. Pacific* **107,** 803.

Weedman, D. W. (1986). "Quasar Astronomy," Cambridge University Press, Cambridge, U.K.

Star Clusters

Steven N. Shore
Indiana University South Bend

GLOSSARY

Association Loosely gravitationally bound group of coeval stars, including massive (OB) stars and often numerous subgroups of more tightly bound systems. Generally associated with massive molecular clouds and sites of recent star formation.

Hertzprung–Russell diagram (H-R diagram) Plot of loci of stars by surface temperature or color versus luminosity or absolute magnitude.

Horizontal branch Stage of helium core burning. In evolved clusters, this appears as a nearly horizontal grouping of stars (i.e., nearly constant luminosity) in the Hertzprung–Russell diagram.

Main sequence Stage of a star's evolution when energy is generated via hydrogen processing in the stellar core.

OB stars Massive stars, typically more massive than $\sim 10\, M_{\odot}$, which have short mainsequence lifetimes and evolve into luminous supergiants and, probably, Wolf–Rayet stars. Generally found in associations.

Pre-main-sequence star Star in the stage of evolution, before hydrogen core ignition, when energy is still being generated by gravitational contraction and mass accretion.

Stellar populations Differentiation between stars of different ages, determined from their metal abundances and distribution. Population I stars are young stars confined to the disk of the galaxy and of metal abundances near the solar value; Population II stars are the oldest stars in the galaxy and reside primarily in the galactic halo, with metallicities much less than the Sun's.

STAR CLUSTERS are gravitationally bound, presumably coeval groups of stars. Clusters can be morphologically distinguished between open or galactic and globular on the basis of both the overall geometry and the stellar density. Globular clusters, which reside primarily in the halo of the galaxy, are associated with an older population and are typically an order of magnitude more metal poor than the disk. They contain upwards of 10^5 members with typical radii of the order of 1 to 10 parsacs (pc). Open clusters are usually of much lower mass, containing of the order of 100 to 1000 members, and are among the recently formed stars in the galaxy. They span the age and metallicity range of the Population I stars. The OB associations are looser groups of massive stars and contain the most recently formed stars. These are thought to be a possible

extension to the lower mass open-cluster sequence. The populous blue clusters are observed in a number of external galaxies, notably the Magellanic Clouds. The use of clusters for the study of stellar evolution and metals production in galactic history is facilitated by the fact that they display a mass range that is similar to that observed for disk stars and appear to form stars sufficiently rapidly that there is no great spread in their initial chemical composition. These objects provide the basic test of models for stellar evolution theory and also fundamental distance calibrators.

I. HISTORICAL INTRODUCTION

The first observations of a cluster can be found in the earliest compilations of the constellations by the Greeks. The Hyades, Pleiades, and Praesepae are all mentioned. These are the most visible northern open clusters, the first two located in Taurus and the last in Cancer. That these stars are in fact of higher density than the rest of the field was initially recognized by Galileo in 1610, in the first use of the astronomical telescope. In the *Siderius Nuncius* he mentions the Pleiades specifically, showing that it has a considerably larger population than can be seen with the naked eye and larger even than the neighboring galactic plane. A more complete collection of cluster data can be found in the first telescopic surveys by Halley in the 17th century and Messier about a century later. These revealed the existence of several classes of clusters—specifically the globular and compact open varieties. For a time, it was generally believed that all such objects were nebulous, in agreement with the Kant—Laplace nebular hypothesis, which appeared to predict the preplanetary stage these objects represented. The resolution by Herschel and later observers of many of these nebulae into stars for a time precluded the general acceptance of the idea of star formation from gaseous matter, but the spectroscopic observations of Huggins and Secchi in the 1860s assisted in distinguishing clusters from nebulae.

In the first years of the 20th century, observations of the globular clusters at Harvard under S. Bailley and later H. Shapley showed that there is a distinct class of variable star associated with the clusters. These are the RR Lyrae stars, a group of horizontal branch stars with periods of $\sim\frac{1}{2}$ day and that display (like many pulsating variables) a distinctive period–luminosity relation. In addition, cepheid variables were also observed and these enabled Shapley to determine that the globular clusters form a spherical distribution about the galactic center and that we are displaced from that center. The current value is 8.5 kpc.

In the 1930s, R. Trumpler discovered that the most distant of the clusters also appeared to be more reddened than the local stars. This led to the discovery of interstellar dust and further served to correct the determination of the distance scale. The discovery of globular clusters in the halos of external galaxies has continued at a rapid pace in recent years, extending their use as distance indicators.

Observations of the colors, motions, and brightnesses of cluster stars had become sufficiently advanced by the mid-1940s to reveal that the stars are in fact coeval. M. Schwarzschild and F. Hoyle, and later A. Sandage, exploited this property of the members to test models for stellar evolution. The argument proceeds as follows. If the stars are formed at the same time and have different masses, then the most massive ones should evolve the most rapidly toward the red giant phase. By using the luminosity and temperature of the brightest main-sequence star, one should therefore be able to find a corresponding mass for the most evolved hydrogen-core-burning star from which the age of the cluster follows. E. Salpeter extended this to the determination of the initial mass function (IMF) by determining the relative number of stars of each mass that are needed to synthesize the morphology of the H-R diagram of the cluster. Fitting a power law shows that many more low-mass than higher mass stars are formed in these clusters. The same is true for the globular clusters, although there is a narrower mass range.

The globular clusters were shown by W. Baade to have the same population mix as, and similar metallicities to, the stars found in the halo of the galaxy and M31, the Andromeda galaxy. Specifically, when a sample of high-velocity field stars (presumed to be the stars that have the most extended distribution and thus are moving statistically faster) is obtained, their overall properties closely match those of the globular clusters. The distributions in space, around the disk of the galaxy, are also quite similar. From the low metallicities and the extended spatial distributions, Baade and subsequent investigators argued that the globular clusters were formed in a rapid sequence of events in the early history of the galaxy, thus forming a sample of the stages before the formation of the disk stars.

Observations from space began for these objects in the late 1960s with the launch of OAO-2, which performed UV photometric observations in many of the globular clusters and individual stars in open clusters. Additional work by International Ultraviolet Explorer (IUE) has shown that a number of globular clusters possess UV-bright cores. These have been resolved using the WFPCZ on the Hubble Space Telescope (HST) revealing a population of blue stragglers and post-AGB stars (see below). EINSTEIN observations have revealed the existence of a class of bright X-ray sources, at first thought to be central block holes and now believed to be low-mass binaries, in $\sim$10% of the globular clusters. None of the open or globular clusters appears to be a γ-ray source.

II. GALACTIC OPEN CLUSTERS

A. Introduction

The young stellar population of the galaxy, Population I, whose metal abundance is about the same as that of the Sun, appears to form only loose clusters, consisting of some hundreds of members, which have ages ranging from only a few hundreds of thousands of years to about the age of the oldest globular clusters, of the order of 10^{10} yr. Among the youngest are NGCs 2244, 2264, and 6530 (all with ages of less than a few tens of millions of years), and among the oldest are NGC 188, NGC 2506, Mel 66, and M 67, which are as old as the oldest disk stars (about 10 Gyr). The mechanism of formation appears to be the same as that of the associations; that is, the stars are formed from a parent molecular cloud in which many stellar masses are present. If we take the mass of the typical cluster to be several hundred solar masses, this places a lower bound to the mass of the parent cloud. The formation time for the stars in the cluster appears to be quite short, of the order of the contraction time for a star of a few solar masses, though we return to this point below.

The stars in the cluster all have a common proper motion and thus were formed in a gravitationally bound system. The mass of the cluster can be estimated as follows. There is a simple relation between the total kinetic and gravitational potential energies of a group of bound objects in a gravitational field, called the virial theorem, which is that

$$2T + \Omega = 0 \tag{1}$$

where

$$T = \tfrac{1}{2} N m_* \sigma^2$$

is the mean kinetic energy of the N stars in the cluster, which are all assumed to have about the same mass m_*, and σ is the velocity dispersion, and

$$\Omega = -GN^2 m_*^2 \langle R \rangle^{-1}$$

is the mean gravitational potential energy, where $\langle R \rangle$ is the mean radius of the stars from the center of the cluster. The agreement between the calculated mass required to bind the cluster and the observed masses of the stars (inferred from their luminosities and spectral types) shows that the clusters were formed as gravitationally bound systems.[1]

[1]The total energy of the cluster is given by

$$E = T + \Omega < 0,$$

which shows that the clusters are bound (that gravity wins out). It should also be added that the virial theorem holds for any gravitationally bound cluster, even of galaxies.

B. Isochrones and the Interpretation of H-R Diagrams for Clusters

The fact that the stars were all formed at approximately the same time means that the cluster represents a snapshot of stellar evolution for the masses of the component stars. Assume that we have two stars of different mass, $M_1 > M_2$, which start their life at the same time. The more massive star will evolve faster owing to the higher rate of core nuclear burning and thus will become a red giant in a shorter time than the lower mass star. If we know that the stars on the main sequence, the hydrogen-core-burning stage of evolution, have no more than some mass, say $M_{\max}$, then we can place a lower limit on both the age and the mass of the evolved star. If we know, for example, that M_2 is a main-sequence star and M_1 is a red giant, then a lower limit on the mass of M_1 can be determined. By fitting stellar evolutionary tracks to the entire ensemble of cluster stars, one can determine the number of stars in a given mass range that were formed at the time of formation of the cluster, the age of the cluster, the initial chemical composition, and the distance to the cluster through the knowledge of the masses and absolute luminosities of the component stars.

In order to employ isochrones, however, the cluster's distance must be determined. The raw data are simply the visual apparent brightness (or V magnitude) of the star plotted versus its color, or $B - V$ (the more negative $B - V$, the bluer is the star; the lower the V magnitude, the brighter is the star). Figure 1 shows the color–magnitude or H-R diagram for a moderately young open cluster, M11. Note the well-populated main sequence and the scattering of a few red giants to the right of the figure. Figure 2 shows the old disk cluster M67. Here, the subgiant branch is well populated, and the turnoff point is at about the age of the sun.

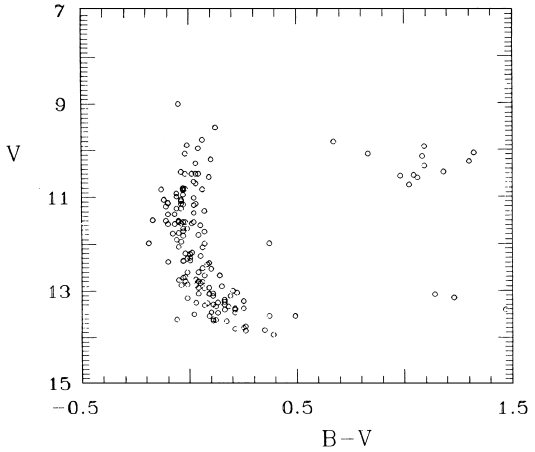


FIGURE 1 Young open cluster, M11. (Courtesy of B. Anthony-Twarog.)

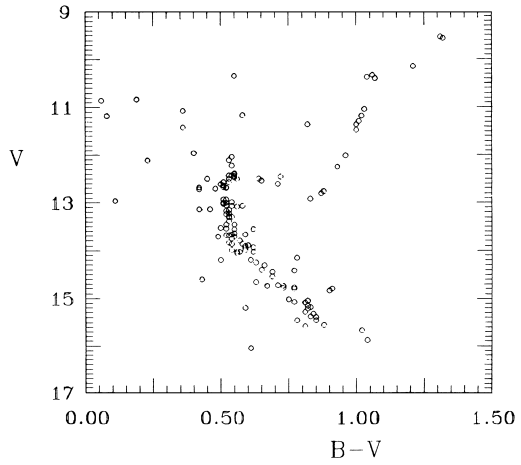


FIGURE 2 Old open cluster, M67. (Courtesy of B. Anthony-Twarog.)

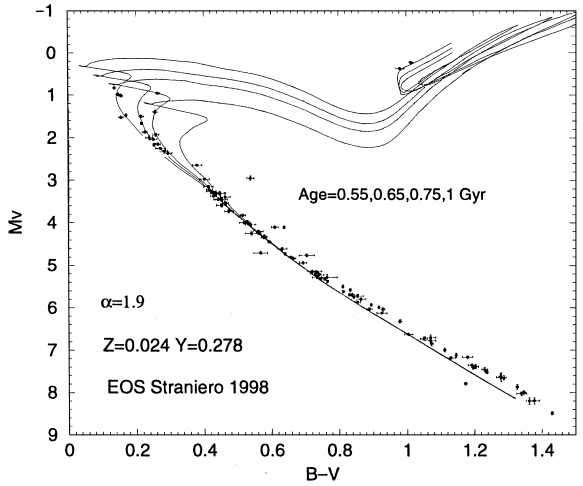


FIGURE 3 Theoretical isochrones for the Hyades color transformed to match photometry data; absolute magnitudes based on Hipparcos parallax measurements. The models used the Staniero equation of state (1998). The mixing length parameter (ratio of convection mixing length to scale height) is 1.3 in these models with the curves labeled by age (the top-leftmost curve is the youngest). Notice how the giants and subgiants constrain the age as well as the main sequence turnoff point. [From the Monthly Notices of the Royal Astronomical Society, **320,** 66–72, 2001, with permission.]

An additional use of isochrones concerns the determination of physical properties of envelope and core convection. In the stars more massive than $\sim$1.5 $M_{\odot}$, the core is dominated by carbonnitrogenoxygen (CNO) processing on the main sequence and is consequently completely convective. The envelopes are radiative, so that there is a turbulent interface between the two regions. Overshooting of the turbulent cells from the core can promote mixing of fresh, hydrogen-rich material into the nuclear processing region, with the consequent increase in the mass of the chemically helium-rich core with time. This produces, at the end of the main-sequence stage, a more rapid than expected contraction of the core and a gap in the H-R diagram of the cluster immediately after the main sequence. Several clusters, notably M67 and NGC 188, have been used for the study of the interiors models, but the definitive answer must await the development of better interiors models. The post-main-sequence gap, as this feature of the diagrams is known, is one of the few probes we have, using the aggregate population of the clusters, of the detailed properties of the interiors of stars more massive than the Sun. This gap can be clearly seen at $V = 13$ in Fig. 2, the color–magnitude diagram for M67. An example of isochrone fitting to the Hyades is shown in Fig. 3.

C. Initial Mass Function

Since all of the stars were formed from the same composition and at the same time in the same place in the galaxy, one can use this important feature to trace the chemical history of the galaxy. The older clusters in the disk have been shown to have a slightly lower metallicity than the Sun, whereas the ones just now being formed have a slightly higher or about the same value of metal abundance. This is one of the key observations indicating that the metallicity of the disk of the galaxy has been secularly changing in time. In addition, it appears that for the lifetime of the disk, during which the open clusters have been formed, there has been little change in the so-called initial mass function. This is the population distribution of the stellar masses.

The actually measured quantity for a cluster is the number of stars in a given range of color and visual magnitude. Stellar atmosphere models are used to determine the transformation between color and effective (surface) temperature, and this provides the bolometric (total) luminosity. The stars are then placed on a "theoretician's" H-R diagram (luminosity versus effective temperature) and traced back to the main sequence. Thereafter, their masses (assuming no mass loss has substantially altered their abundances or masses) can be specified, and the number per unit mass can be determined. This is the initial mass function of the cluster.

The observational picture is still only approximately known, but there is strong evidence that the IMF for cluster stars is the same as that for field stars and that for at least the past 10^{10} yr it has remained stable, with few high-mass and many low-mass stars being formed. Simple power law fits to the distribution show that

$$\Phi(M)\,dM \sim M^{-\alpha}\,dM,$$

with α being 2.35. There is still considerable uncertainty in the exponent ($\pm$0.5). A more accurate fit for field stars is a log-normal distribution. However, the basic result, that the massive stars are considerably rarer than the low-mass

ones like the Sun, is unchanged by this detail. The details of the origin of the mass function are, presently, unknown. Qualitatively, it is attributed to a stochastic fragmentation cascade to some minimum stable mass. It is also likely that it entails a kind of fragmentation–coagulation phenomenon.

The lower mass of star formation is also an important parameter for stellar evolution theory. The lower end of the main sequence, representing the minimum mass at which stable nuclear reactions can be initiated in the stellar core, appears to be at $\sim 0.1\ M_\odot$. The minimum mass from fragmentation is of the same order, so the census of this population in clusters is a very important task.

Another check on stellar evolution theory provided by open clusters is that there are white dwarf stars present in many of the older clusters. These are extremely compact remnants of stellar evolution, formed from stars of about the mass of the Sun. They are no longer generating luminosity by nuclear processes, so their presence indicates their cooling time and rate of formation. Again, it is important to be able to observe these in systems of known age, since then both the ages of the white dwarfs and the masses of the progenitor stars can be specified. There are no pulsars associated with any cluster presently known, nor are clusters generally associated with supernova remnants. In general, the massive stars appear to have a connection with associations and not the open clusters, and this may be an additional clue to the formation mechanism for such systems.

In some clusters, notably the Pleiades (among the younger group of open clusters), there is a sequence above the main sequence, in the lower mass star region of the H-R diagram, that may be an indication of a continuing star formation in the cluster. The blue stragglers may also represent such a group. The bounds on the time scale required for star formation to proceed in these systems is thus of the order of the premain-sequence contraction time for a very low-mass star to the main-sequence lifetime of a roughly $3\ M_\odot$ star. This is from about 10^7 to 10^8 yr.

D. Rotation in Open Clusters

Studies of stellar rotational velocity in clusters have shown that there is also a distribution of angular momentum among young stars, such that the angular momentum J is larger for the more massive stars. Although at present we do not know why this is the case, the study of stars in clusters greatly aids the determination of this important parameter for star formation theory. The lower mass stars are typically slow rotators, owing to loss of angular momentum during the main-sequence phase through stellar winds, but in some clusters there is an excess of rapid rotation on the lower main sequence.

One of the few relations that appears to hold for rotation has been determined by A. Skumanich. It is that the rate of chromospheric activity appears to decline with age and that, if this is due to the spin-down of the stars owing to angular momentum loss via a stellar wind, then the rotation frequency should vary as $\Omega \sim t^{-1/2}$, where t is the age. For the youngest clusters, like the Pleiades, the low-mass stars also display strong stellar coronas, suggesting that the enhanced dynamo activity presumed to be the origin of the X-ray-emitting region is connected with the statistically higher rotation rate of these stars due to their age.

Recent observations of rapidly rotating lower-main-sequence stars in the Pleiades and possibly other young clusters may be an indication of a long time (in comparison with the age of the upper-main-sequence stars) of formation for the cluster members. In associations, this appears to be short in the subgroups of the systems but in galactic clusters may take as long as 10^8 yr. Much remains to be done in relation to this problem.

E. Binary Frequency in Open Clusters

There is still considerable uncertainty about both the frequency of occurrence of, and distribution of mass ratio in, binary systems in clusters. The problem is that few long-term surveys of clusters have been undertaken to determine the presence of small-amplitude velocity variations in main-sequence stars. About half of the stars in the field are binaries, and the mass ratio for many of these systems is nearly unity. (It is still uncertain what the true value is.) There appears to be a dearth of binaries among the Pleiades stars, but the claimed excess in IC 4665 has been shown to be incorrect. There may be some clusters for which the binary statistics are unusually high or low compared with this value, but this is a problem for the future.

F. Star Formation in Clusters

The youngest clusters, such as the Trapezium, σ ori, and NGC 2264, are still imbedded within their parent clouds. Many of the individual stars in these young clusters still have dust shells and are contracting toward their initiation of hydrogen core burning. In a few such young clusters, one can see stars evolving both toward (the lower mass stars) and away from (the highest mass ones) the main sequence, thus providing a check on the pre-main-sequence contraction time and the rate of stellar evolution on the main sequence. There is also some evidence, from the fact that the lower mass stars evolve more slowly throughout their lifetimes, that the star formation in open clusters may take place over a time scale comparable to the age of the cluster. In some clusters, like the Pleiades and NGC 752, there is evidence, albeit weak at the moment, that

there are several epochs of star formation. The spread in the rotational and luminosity properties of the lower main sequence is larger than would be expected for a sample of stars formed simultaneously.

Some clusters display a population of stars that lie above the main-sequence turnoff points of the massive stars, the so-called blue stragglers. These may be stars that have been the product of mass transfer between the members of close binary systems, in which one of the stars has been moved above the turnoff point due to mass transfer from a companion that is evolving toward the red giant phase. However, there is little evidence for binarity among these stars. An interesting alternative is that they have somehow become completely chemically mixed and are evolving as homogeneous configurations. The evolution of such stars, rather than being toward the cooler parts of the H-R diagram, is almost along the main sequence toward slightly smaller and hotter configurations as the stars age as a result of the global increase in the mean molecular weight of the gas in the interior. Normally, stars massive enough ($\sim$1.5 $M_\odot$ or greater) to burn hydrogen via the CNO cycle would have convective, completely mixed cores whose mean composition would be considerably heavier than the stellar envelope (which is radiative). As a result, as the cores contract under the increase in the mass of the constituent species, the released gravitational binding energy is capable of lifting the envelope. This causes the star to increase, overall, in radius and move toward lower surface temperatures (the mechanism responsible for producing red giants). However, the mechanism whereby the mixing is initiated and maintained remains conjectural. The blue stragglers remain one of the important puzzles in the otherwise largely coherent picture of cluster evolution. They may even be additional evidence for multiple epochs of star formation.

G. Abundance Studies

An important use of galactic clusters is the study of the age dependence of the metallicity of the stars in the galactic disk. There are several elements for which it is especially important to know the absolute, rather than statistical, ages. One such element is lithium. It is well known that lithium is destroyed by nuclear processing in solar-type stars. The element is mixed into the core by convection and totally consumed by proton nuclear processes. Thus, it is a useful constraint on the rate of mixing in stellar envelopes as well as a valuable potential age discriminator in field stars. Studies of intermediate-age clusters show that the time scale for this mixing to occur is on the order of 1 Gyr and is also consistent with the change in the rotation rate for solar-type stars due to angular momentum loss via stellar winds.

We have no direct determination of the deuterium abundance in cluster stars, but this would also serve as a useful constraint on both the rate of consumption of deuterium via a mechanism similar to that of lithium destruction and also on the primordial deuterium abundance.

The abundance of heavy metals appears to vary among cluster stars, but this is likely due primarily to processes connected with mixing and diffusion that are intrinsic to the stellar envelope. As yet, there is no clear measure of the intrinsic dispersion in initial abundances among stars in a cluster. The abundances overall follow the trend of the field stars—the youngest clusters are a bit more metal rich than the Sun, whereas the oldest clusters are considerably more metal poor.

H. Variable Stars in Clusters

Several classes of pulsating variable stars are observed in open clusters. The main-sequence population comprises the δ Scuti stars, which are $\sim$1.5 $M_\odot$ and are late A and early F stars. These have periods of about an hour and amplitudes that are typically several hundredths of a magnitude. They are the same type as the field population, but their properties can be best studied, as for all pulsators, when their distances can be independently determined and their absolute magnitudes assigned. Perhaps the most important aspect of clusters, however, for the purposes of both stellar evolution and cosmology is that a few clusters have δ Cepheid or classical Cepheid variables as members. These stars, which are about 3–5 $M_\odot$, are evolved stars that pulsate with periods of several days to several months and have large amplitudes (typically near a magnitude). They have a regular period–luminosity relation from which their absolute brightnesses can be determined given their period of variation, a comparatively easily measured quantity. They form the anchor of the extragalactic distance scale. It is, in part, the fact that the Cepheids occur in clusters that has led to one of the most interesting questions of stellar evolution theory. When one determines the masses of the stars, they typically have evolutionary masses that are greater than indicated by pulsation-theoretical calculations. Although a number have been shown to be members of binary systems, thereby altering their placement in the H-R diagram, not all have been shown to be such. Thus, again, clusters provide one of the most important handles on the absolute properties of these stars.

Flare stars and T Tau stars are associated with the youngest clusters. The flare stars, which are chromospherically active M dwarfs, may be more frequent because of greater dynamo activity in the newly formed systems. These stars are rapidly rotating systems and may not yet have spun down to their more sedate rotation rates due to

breaking by stellar wind-driven angular momentum loss. One serious complication in this observational picture is that at present there are few surveys of such activity in clusters.

I. Magnetic Fields in Cluster Stars

Additional important information provided by galactic clusters is the occurrence of magnetic fields in main-sequence stars. Only a few clusters have so far been studied in any detail for direct observation of magnetism, but there is some indirect information provided by the presence of Ap stars. These are stars with anomalously high (compared with the average Population I star and with the Sun) abundances of heavy metals—especially the rare earths and silicon—which possess strong magnetic fields. They occur in clusters having ages greater than $\sim 10^7$ yr, suggesting that the magnetic fields are present but that there is some lag time after formation for the development of the abundance anomaly. These stars are massive, typically greater than 2 $M_\odot$.

For the lower mass stars, the statistics for the presence of chromospheric activity and X-ray emission form a useful tool with which to study the development of dynamo-generated magnetic fields in stars. The youngest systems, like the Pleiades, have already well-developed coronal activity indicators, and it seems that most stars in the youngest clusters on the lower main sequence have some X-ray emission associated with them. Indeed, in agreement with the fact that the mean rotational velocity decreases with time, the oldest clusters show weaker dynamo activity among the lower main sequence stars than do the young clusters. Again, because the ages of these stars can be determined independently from the H-R diagram for the cluster, the study of the aggregate properties is far more valuable than the detailed work on field stars for which such information is only statistically available. There is only weak evidence at presence for the decay of the strong magnetic fields of the upper main sequence stars with age, although the time scale for such a phenomenon should be of the order of 10^9 yr or fewer due to diffusive losses. This remains one of the most important areas of research in open cluster populations.

III. ASSOCIATIONS

A. OB Associations

The most massive stars in the galaxy do not appear to form in the tight configurations that open clusters represent. Instead, they form dynamically fragile associations of several thousand solar masses. Some, like the Sco–Cen association, contain open clusters (NGC 6231). Others, like Orion, contain tight multiple-star systems of the Trapezium variety, which contain ~100 stars but which in total have masses about like those of very small open clusters. A few show evidence not only of having formed over a considerable period of time, of the order of 10^7 yr, but also of forming in sequences (e.g., Cep OB3 and Ori OB1). Many OB associations, these two groups being prototypical, are still associated with their molecular clouds and show evidence of continuing star formation.

The physics governing the division between cluster and association is not well understood although it is clear that a continuous hierarchy exists that connects star formation on different scales. One suggestion is that the formation of massive stars stops the continued formation of the lower main sequence objects and, by altering the thermal and dynamic structure of the environment, brings the system's star-forming activity to a halt. A related point is that only the most massive molecular clouds may have sufficient mass and stability to allow for the formation of OB associations, whereas the average cloud may serve as the nursery for the lower mass objects. If this is the case, the OB associations should show at the same age as the clusters a systematic preference for the upper main sequence stars, a test that has yet to be completed. The problem is that there are few open clusters known from optical observation to be of sufficient youth with which to test this idea. Instead, infrared satellite: observations of clusters with IRAS and ISO should be used to determine the mass function for the stars forming currently in clouds of different mass.

Another important feature of the associations is that they can evaporate. Due to the gravitational interactions of a bound star in the group with the others in the vicinity, there can occasionally be a large enough kick from a close encounter that the star is sent at escape velocity or higher out of the system. The main reason is that the associations are rather fluffy, from the gravitation point of view—quite distended even considering the masses of the individual stars. The escape of stars from the clusters produces a contraction of the cluster core, increasing the rate of encounters and gravitationally "heating" the stars in the association. The contraction of the core of the association therefore increases, with a further increase in the rate of escape of the stars. This evaporative dissipation of the cluster, first discussed by S. Chandrasekhar, L. Spitzer, and J. Oort in the 1940s and 1950s, is the main mechanism for the dissolution of the associations.

It is likely that the evaporation time scale for the associations is always short, of the order of the main-sequence time for the B stars, and may be important for understanding the formation mechanism for these systems. The Ori OB1 and Cep OB3 associations have among the best determined dynamic ages, and these agree quite well with the inferred stellar ages from evolutionary models.

B. R and T Associations

If the open cluster or association is still connected with dust, it is sometimes called an R association, for *reflection*. Dust has the property of scattering, as well as absorbing, starlight. The presence of B and A stars in an association, in the absence of massive stars, does not provide sufficient UV radition to ionize the environment. A result is that the neighborhood dust survives, without the formation of an emission nebula, which then scatters the starlight and forms an extended filamentary diffuse nebula. These R associations are among the youngest clusters in the galaxy, having ages a bit greater than those of the OB associations.

Finally, if the association contains T Tauri stars, which are pre-main-sequence objects, it is sometimes called a T association. The purpose of separating these groups was to call attention to the star-forming activity associated with them. However, in light of the recent observations by ISO that star formation is present in the molecular clouds connected with many or most of the OB associations in the galaxy, even if it were optically invisible to us, there seems little reason to continue this designation. The T Tau stars are well emerged from their parent cloud material, and the indication of ongoing star formation is better provided by the wealth of point IR sources observed to be associated with many of the OB associations in the galaxy.

IV. GLOBULAR CLUSTERS

A. Introduction and Basic Properties

The most striking property of globular clusters is their overall optical appearance. They form nearly spherical, compact systems that are enormous aggregates of stars. It is this feature that was first selected as their distinguishing characteristic. The morphologically distinguishing features of individual clusters are the degree of central concentration, the so-called Shapley classes (on a scale of I to XII, depending on decreasing concentration), and the ellipticity. The latter, though originally looked to as providing evidence of the rotation of the cluster as a whole, actually appears due to both the tidal interaction with the galaxy as a whole and the formation of nonaxisymmetric internal velocity and spatial structures.

The frequently employed model for the overall mass distribution, the so-called King model, assumes that the globular cluster is an isothermal sphere of stars. This assumption means that the velocity dispersion is assumed to be homogenous and isotropic. The stars behave as if they are particles in the mutually generated gravitational potential well moving around at uniform temperature. The rate of escape of stars from such a system is seen to be lower that the rate of escape from the open clusters because of the relative compactness of the potential and of the mass of the cluster. The central portion of such spheres is called the core radius, a measure of the distance over which the surface brightness of the cluster falls to about half its central value. The Shapley classes are not well correlated with this more quantitative measure of concentration, for which reason the former is preferred as a characterization of the culster morphology.

B. Integrated Properties of the Cluster System of the Galaxy and Comparison with External Galaxies

An important feature of the globular clusters is that they can be studied as if they were extremely bright stars and, because of their compactness and brightness, can be observed to very large distances. This includes the halos of external galaxies. The cluster system in our galaxy appears to consist of two components, the halo and disk systems, which are distinguished by their metal abundances and spatial distributions.

For the halo system, the mean integrated absolute visual magnitudes of the disk and halo clusters are about the same, $\langle M_V \rangle = -6.9$, to within $\sim$20%. For external galaxies, only a few systems have so far been observed. These include most of the galaxies in the Local Group (the cluster to which the galaxy belongs) and NGC 5128 = Cen A, a nearby active and morphologically extremely peculiar galaxy.

There exists in the Large Magellanic Cloud a population of young, globularlike value clusters, the so-called populous blue clusters. These appear to be still in a mode of active star formation. Clusters having ages as young as 0.01 Gyr are observed. Recent observations of several blue clusters show that they can have ages as great as several gigayears but, again, that they are not as old as the glactic globular clusters. Similarly, these have been observed in the even more metal poor Small Magellanic Cloud. They typically have metallicities at most 10% of the solar value, comparable with the most metal-rich galactic clusters. This latter point is likely the most important clue to their origin, since the Magellanic Clouds, being of lower mass than the galaxy, also have an overall lower abundance of metals.

For M31, the clusters appear to have an excess of blue evolved stars for the same metallicities (determined from integrated spectra), but much remains to be done on the populations of clusters in general. Their luminosity distribution is not markedly different from that of our galaxy. For M33, the mean magnitude is about the same as in our galaxy, and the population of the LMC-like blue clusters is lower, more in agreement with our experience locally. The same appears to be true for the clusters about the active

peculiar elliptical galxy NGC 5128. The cluster system of M87, the central giant elliptical galaxy in the Virgo cluster, has also shown that many of the clusters have a higher metallicity than those in our galaxy and that there is a substantial population of brighter systems than we possess. There appears to be no radial gradient in the abundances. More detailed information on external galatic systems will have to await further high-resolution space observations. Strongly interacting starburst galaxies, such as Arp 220, show such compact blue clusters in currently triggered star formation. And the Galactic Center also shows massive ongoing star formation in dense compact clusters analogous to R 136 in the LMC H II region 30 Doradus.

C. Metallicities and Space Distribution

The globular clusters are characteristic members of the old population, Population II, of the galaxy. They are distributed in a halo around the disk, although some are present in the disk. Their metallicities range from the lowest values observed in the oldest open clusters to $\sim 10^{-3}$ the solar metallicity (which is ~ 0.02 by mass). The usual means of quoting metallicity for these clusters, as with the open galatic clusters, is in terms of the Fe/H ratio. This is normally quoted as a differential measure relative to the Sun, as $[\mathrm{Fe/H}] = \log(\mathrm{Fe/H}) - \log(\mathrm{Fe/H})_{\odot}$.

An important observational fact, still not well understood, is that there are no galatic globular clusters with the characteristics of the Population I stars and that they do not now appear to form in the disk. Their distribution can be separated into two fairly distinct groups, differentiated on the basis of metallicity. One is distributed approximately spherically around the disk, centered on the galactic center, and concentrated toward the bulge of the galaxy. These are the most metal poor clusters, having values of Z, the metal fractional mass abundance of metals, ranging from $10^{-4}\ Z_{\odot}$ to about $10^{-1}\ Z_{\odot}$.

The metallicity scale for the globular clusters is not as well established as that for the open systems, since they are intrinsically fainter and only the more evolved stars can in fact be individually analyzed. These tend to be stars on the red giant branch, which may have undergone internal mixing processes and so may not be representative of the entire cluster abundances. Furthermore, the metallicity determined from integrated cluster spectra, in which the entire system is treated as if it were a single star, are severely affected by the morphology of the cluster H-R diagram. Specifically, the population of the blue end of the horizontal branch can cause spurious abundance results when photometric indices are used. The lines from these stars are extremely strong and the stars are far brighter than the main sequence so that they also wash out the contribution from these fainter, although more numerous, stars. However, despite the uncertainty, it appears that even the most metal-rich globular clusters, like 47 Tuc, barely if at all reach the current abundances of the disk. Dynamical evidence from halo stars points to accretion of stars from galactic collisions with the Milky way. The same is indicated by the bimodal globular cluster distribution, which separates into a group that is more confined toward the Galactic plane, such as ω Cen, while most are more broadly distributed and follow the halo mass profile.

Observations of the distribution of the cluster suggest that the break point in the flattening of the cluster system occurs at $[\mathrm{Fe/H}] = -1$ to -1.5 (depending on the metallicity scale employed). The spatial distribution of the more metal-rich globulars appears flattened, with a scale height (distance above the Galactic plane) of the order of 0.5 kpc. These are far fewer in number than the more metal poor systems and appear to have the metallicity of the oldest field stars in the disk, the so-called old Population I distribution. They too are concentrated toward the galactic center.

D. H-R Diagrams and Aggregate Properties

Two important observational features dominate the study of these clusters. One is the fact that the morphology of the H-R diagram appears to depend on the metallicity of the stars. The stage of the helium core burning is represented in clusters by a horizontal branch (HB), which lies about a magnitude or so above the main sequence and stretches across the diagram from red to blue, connecting with the nearly vertical red giant branch on the cool end. The lower the mean metallicity of the cluster, the bluer the horizontal branch is. This is reflected in the $B/(R+B)$ ratio for the HB stars. Since the presence of RR Lyrae stars is a function of the morphology of the HB, the clusters of lower metallicity have a larger population of these variables. In addition, there appears to be a relation between the HB morphology and distance of the cluster from the galactic center such that the lower metallicity stars and the bluer HB clusters lie farther from the center. The relation is a weak one, but it nonetheless appears to indicate that the objects of higher metallicity lie closer to the galactic disk.

There is one other metallicity-dependent parameter for globular clusters. The brightness of the tip of the giant branch relative to the HB is also a function of the mean cluster metallicity, although it may depend as well on the initial helium abundance. One serious problem in understanding this is that there is a considerable spread in metal abundance, especially of CNO, among the stars on the giant branches of several clusters. The study is hampered by the fact that only the brightest stars, in the nearest

TABLE I Metallicity of Select Globular Clusters

NGC	Other name	[Fe/H]	$B/(B+R)$	$\log(Z/Z_\odot)$
104	47 Tuc	−1.1	0.0	−0.4
5272	M3	−1.6	0.5	−1.2
5904	M5	−1.1	0.8	−0.8
6205	M13	−1.4	1.0	−1.2
6341	M92	−2.1	1.0	−1.7
7078	M15	−1.8	0.8	−1.4

clusters, can be studied presently. Nonetheless, there is a spread of upward of an order of magnitude in the abundance of CN in ω Cen and in 47 Tuc, the two brightest such clusters.

The integrated spectra of the clusters can also be observed. The most important parameter here is ΔS, the difference between the spectral type provided by the hydrogen and the metallic lines (especially Ca II H and K). This can be derived for both individual stars (e.g., halo high-velocity stars) and for the integrated light of the clusters. It is an excellent indicator of metallicity. The metallicities of globular clusters are summarized in Table I.

E. Ages

The globular clusters are the oldest members of the galactic system that we can study in detail. Their ages are inferred to be about 11 to 13 Gyr, with no large age spread even with the large observed metallicity dispersion. It is clear that the epoch of formation of these objects lasted only for a brief interval, in contrast to the ongoing activity of star formation in the disk of the galaxy. Both the optically observed color–magnitude diagrams and the UV properties of the stellar populations of the clusters give similar ages. It appears that the disk increase in abundance was very rapid indeed, taking no more than a few gigayears.

F. Luminosity Functions

The mass function for globular clusters is heavily weighted toward the low-mass stars because of evolutionary effects. The turnoff points for most of the clusters are well below 1 $M_\odot$, leading to functions that can be easily compared with only the solar neighborhood in the disk of the galaxy. Until the introduction of charge-coupled devices (CCDs), the determination of this population was hampered by the faintness of the stars; however, recently a few clusters have been studied to lower than 0.3 $M_\odot$. One of the most thoroughly studied is M13, the Hercules globular cluster. It has been taken to $M_V = +9.5$. The IMF is steeper above 0.65 $M_\odot$ than for the galactic disk, flattening out for the lower mass ($\leq 0.5\ M_\odot$) objects compared with the field. The intrinsic width of the main sequence is quite narrow, of the same order as observed for the most populous open clusters, less than 0.1 magnitude, and there is no evidence for a sizable population of equal-mass binary stars. This has been seen in a number of open clusters as well. The low spread, even in light of the rotation of stars (see discussion for the open galactic clusters), is in part due to the extreme age of the systems (these stars would have slowed in their rotation by now) and the fact that lower mainsequence stars are intrinsically slow rotators anyway. There is also no evidence for a second epoch of star formation in the clusters, which would agree with the narrow spread in their various evolved members across the H-R diagram.

G. Initial Mass Function

The globular clusters now consist of only lowmass stars, but this is likely due to their epoch of observation and formation. If they were formed in the early stages of the evolution of the galaxy and all of the stars within them were born at the same time, then we would not see only stars lower in mass than the sun due to stellar evolution. We therefore cannot use them as direct confirmation of the stability of the IMF with time, since all of the massive stars have disappeared.

H. Variable Stars

The primary variable stars in these clusters are the RR Lyrae type, which have periods of about 0.3–0.6 day. Originally called cluster variables, they have been extensively cataloged in the past several decades. The light curves fall into two types: the *ab*, which have sharp rising portions and slow declines, and the *c* type, which is more nearly sinusoidal. The *ab* type tend to be longer periods and have larger amplitude. First recognized by Bailly in the late 19th century, they are apportioned differently in different globular clusters. This is the Oosterhoff dichotomy. The bluer the horizontal branch, the larger is the ratio of *c* to *ab*. There appears to be some single parameter, as yet unknown, that controls both the population of the HB in general and the population of instability strip in particular. Both metallicity and initial helium composition have been implicated. The RR Lyr stars have nearly the same absolute magnitude, about +0.7, and thus allow the distances to the clusters to be determined independent of parallax measurements. Since the light of the cluster is dominated by the HB and the red giants, these colors and synthetic population models can be used to determine the distances to galaxies in which the clusters can be located. This is yet another use of the globulars as distance indicators on the cosmological scale, one that will be increasingly important with the advent of the HST.

I. X-Ray Emission

About 10% of the known globular clusters in the galaxy are X-ray sources. Not all of them, however, show the bursting activity with which the class is usually identified. The burst sources are due to accretion onto a neutron star of matter from a low-mass companion, which may have formed a binary by means of capture in the cluster's lifetime. There is no evidence from X-ray data for the presence of diffuse gas in globular clusters. In every cluster observed with ROSAT and *Chandra* the XR emission has been resolved into stellar sources that are cataclysmic binaries.

J. Binary Stars and X-Ray Sources

Binary stars have been directly detected in globular clusters by light variations and direct radial velocity measurements. These are mainly WUMa-type common envelope systems on the main sequence, but novae and other cataclysmics are found. Millisecond pulsars, that are produced by spinup of binary neutron stars, have also been detected, particularly in 47 Tuc.

Since the formation of neutron-star-containing binary systems involves supernova explosions, it is probable that they are formed from the remnants of low-mass stars, which have been pushed past the Chandrasekhar limiting mass for stable white dwarfs rather than from recently formed massive objects. The presence of such stars in the clusters is also strong support for the mechanism of neutron star binary formation for SN type I in the galaxy.

Simulations argue, however, that such systems should be quite rare in globulars. The lifetime for a binary against tidal disruption from a close encounter with another star is so short that there should be no longer period (days or weeks) separations permitted in these systems. The best model for the formation of any close binaries still appears to be tidal capture of the companion. Blue stragglers support this. They are likely merger remnants resulting from capture and subsequent tidal dissipation of angular momentum.

K. Diffuse Gas and Star Formation

There is no compelling evidence for diffuse interstellar gas in globular clusters. Sensitive searches have failed to turn up extended emission, and there are no known instances of planetary nebulae in any cluster, although in at least one case there is a chance superposition of a foreground nebula with M15. So there is again strong support for the contention that such clusters cannot and do not undergo star formation in the present epoch. Finally, the evidence for blue stars in the cores of these clusters comes from the HB objects and not from new massive stars. The clusters contain red giants that should have strong stellar winds, of the order of $10^{-7} M_{\odot}\ \mathrm{yr}^{-1}$. However, the fact that the material does not accumulate within the cluster also suggests strongly that it is swept out on time scales short compared with the cluster lifetime. Such a mechanism as the ram pressure-induced stripping of the clusters as they pass through the interstellar medium in the plane of the galaxy has been successful in explaining the absence of this extended material. Material may also be blown out by supernovas or removed via a galactic shock due to the tidal potential of the plane. The gravitational binding energy of globulars is such that if the gas is heated by supernovas and other hydrodynamic heating sources, it will simply be blown out of the cores of the clusters and lost, unlike clusters of galaxies in which the matter can be retained to much higher temperatures. Again, considerable work remains to be done on this important problem in the chemical history of the galaxy.

V. CONCLUDING REMARKS

At present, from IRAS and ISO observations and related work on the chemical evolution of the galaxy, it is certain that stars arise in clouds of considerable mass—from hundreds to millions of solar masses. Very likely, they also arise in association with other stars, perhaps by stimulated processes such as supernova-induced collapse of clouds or spontaneously. Clusters hold the most important key to an understanding of the formation of stars and ultimately of the chemical input for all of subsequent galactic history. The question of the evolution of the mass function for star formation is best answered with cluster data, since we have the additional certainty that the stars are essentially simultaneously formed (on the time scale of the galactic lifetime). This has direct implications in cosmology, the interpretation of the spectra of distant galaxies at earlier epochs in their histories.

The comparison of galactic and extragalactic clusters and associations is likely to yield an important test of stellar evolution: the dependence of stellar population properties on the environment. This will have to await high-resolution observations from the HST and related future large telescope projects. CCD imaging has greatly improved knowledge of the lower main sequence of old galactic and globular clusters, and it is likely that in the next few years age determinations will be routinely performed for globulars by fitting main-sequence turnoff points. The helium abundance determination for evolved clusters can also be greatly improved by such observations, an important cosmological parameter. Finally, the observation of spectra for individual cluster stars in the lower main sequence will also improve as telescope

imaging technology and solid-state detectors are improved. The revision of the metallicity scale for stellar populations and of the understanding of the origin of H-R diagram morphology for globular and open clusters remains an intriguing prospect for the next decade.

SEE ALSO THE FOLLOWING ARTICLES

BINARY STARS ● GALACTIC STRUCTURE AND EVOLUTION ● NEUTRON STARS ● STARS, MASSIVE ● STARS, VARIABLE ● STELLAR SPECTROSCOPY ● STELLAR STRUCTURE AND EVOLUTION ● SUPERNOVAE ● X-RAY ASTRONOMY

BIBLIOGRAPHY

Bailyn, C. D. (1995). "Blue stragglers and other stellar anomalies: Implications for the dynamics of globular clusters," *Annu. Rev. Astron. Astrophys*. **33,** 133.

Blaauw, A. (1964). The O associations in the solar neighborhood. *Annu. Rev. Astron. Astrophys*. **2,** 213.

Castellani, V., Degl' Innocenti, S., and Prada Moroni, P. G. (2001). "Stellar models and the Hyades: The Hipparcos test," MNRAS **320,** 66.

de Zeeuw, P. T. *et al*. (1999). "A Hipparcos census of the nearby OB associations," *Astron. J.* **117,** 354.

Freeman, K., and Norris, J. (1981). Physical properties of globular clusters. *Annu. Rev. Astron. Astrophys*. **19,** 319.

Goudis, C. (1982). "The Orion Complex: A Case Study of Interstellar Matter," Reidel, Dordrecht, Holland.

Harris, G. H. (1970). "Atlas of Galactic Open Cluster Color-Magnitude Diagrams, Vol. 4," David Dunlap Observatory.

Hut, P. *et al*. (1992). "Binaries in globular clusters," PASP **104,** 981.

James, K., ed. (1991). "The Formation and Evolution of Star Clusters," Astronomic Society of the Pacific, San Francisco.

Lebreton, Y. (2000). "Stellar structure and evolution: Deductions from Hipparcos," *Annu. Rev. Astron. Astrophys*. **38,** 35.

Lewin, W. H. G. (1980). X-ray burst sources in globular clusters and the galactic bulge. *In* "Globular Clusters" (D. Hanes and B. Madore, eds.), p. 315, Cambridge Univ. Press, New York.

Philip, A. G. D., and Hayes, D. S. (1981). "Physical Parameters of Globular Clusters: IAU Colloquium 68," Davis, New York.

Pilachowski, C. A., Sneden, C., and Wallerstein, G. (1983). The chemical composition of stars in globular clusters. *Astrophys. J. Suppl.* **52,** 214.

Sandage, A., and Roques, P. (1984). Main sequence photometry and the age of the metal-rich globular cluster NGC 6171. *Astrophys. J.* **89,** 1166.

Vanden Berg, D. A., Bolte, M., and Stetson, P. B, (1996). "The age of the galaclic globular cluster system," *Annu. Rev. Astron. Astrophys*. **34,** 461.

Zinn, R. (1985). The globular cluster system of the galaxy. IV. The halo and disk subsystems. *Astrophys. J.* **293,** 424.

Stars, Massive

Steven N. Shore
Indiana University South Bend

GLOSSARY

CNO cycle Thermonuclear fusion process that converts hydrogen into helium by reactions with carbon, nitrogen, and oxygen.

Eddington limit Greatest luminosity that a self-gravitating, luminous object can achieve while still dynamically stable. It is the luminosity at which radiation pressure can balance gravitational acceleration.

Eta (η) carinae Prototype luminous blue variable. One of the intrinsically brightest stars in the Galaxy, known as an eruptive optical variable. Surrounded by a dense nebula formed from previous eruptions, it has often been cited as a prospective supernova candidate.

Luminous blue variable (LBV) Massive supergiant that undergoes aperiodic brightness increases of over several magnitudes, with a strong stellar wind; also called S Doradus variable and Hubble–Sandage variable.

P Cygni lines First observed in the luminous blue variable P Cygni. These lines have blue-shifted absorption and strong redward-shifted emission. Absorption is formed in the expanding part of the flow in the line of sight to the photosphere while the emission is from the extended envelope. These profiles are indications of mass loss. The absorption extends to the terminal velocity, which is about the same as the escape velocity from the star.

Units Solar luminosity ($L_\odot = 4 \times 10^{33}$ erg sec^{-1}); solar mass ($M_\odot = 2 \times 10^{33}$ g); solar radius ($R_\odot = 7 \times 10^{10}$ cm); $M_\odot$ yr$^{-1} = 6.3 \times 10^{25}$ g sec^{-1}.

Wolf–Rayet star Highly evolved stars with extremely strong stellar winds and hydrogen-deficient envelopes. The surface abundances of He, C, N, and O have been severely altered from their normal (solar) values because of mass loss. Observed as members of binary systems, their companions may be either normal or compact stars.

THIS REVIEW of recent work discussed the evolution of stars with masses greater than 10 $M_\odot$. Such stars are the major source of mechanical and radiative energy for the Galaxy and intrinsically the rarest stellar objects. They are the progenitors of type II supernovae, Wolf–Rayet stars, and the luminous blue variables. They possess strong stellar winds and provide an important abundance

input of heavy metals to the Galaxy through this mass outflow.

I. INTRODUCTION

The study of the most massive stars is an astrophysical problem belonging to this century, which was the first to recognize their existence.

The measurement of stellar masses began with the discovery of eclipsing binary stars and the determination of stellar spectral types. It was realized by the middle of the last century that stars come in a wide range of masses, ranging from about 10% of the mass of the Sun to more than 50 $M_\odot$. The most massive star for which a direct determination of the mass is possible, Plaskett's star, consists of two main sequence objects with about 50 $M_\odot$. Through the observation of detached binaries, it has been possible to obtain a fair understanding of the mass luminosity relation for the main sequence, the first stage of stable stellar evolution corresponding to core hydrogen burning.

Extended stellar atmospheres were first observed in the strong emission line Wolf–Rayet stars, in the Be stars, and in P Cygni at the end of the 19th century. The first explanations in terms of mass outflow were proposed in the 1930s by C. Beals. However, it was not until the late 1960s that rocket ultraviolet spectroscopy showed that all massive stars lose mass via stellar winds. Even stars that do not show strong optical emission lines show the signature of outflows in the ultraviolet on the profiles of the C IV, Si IV, and N V lines. The inclusion of this information in stellar evolution models, beginning in the 1970s, dramatically changed the picture of the life cycle of a massive star.

In 1953, Hubble and Sandage discovered the class of variables that now bears their names. These stars are the brightest stars in a galaxy and are best observed in the nearby spirals M 31 and M 33, both members of the Local Group. These stars typically have luminosities in excess of $10^5\, L_\odot$ and vary irregularly on time scales of decades, which makes them exceedingly difficult to follow. Because they are generally faint optically, they have been difficult to study until the advent of solid-state detectors, which permit their spectra to be measured with high precision. As early as 1970, models of massive supergiants showed that radiation pressure produced vibrational instabilities that might lead to mass ejection. The extremely rapid increase in computer capabilities has produced significant advances in these calculations.

The theory of radiative stellar winds has also developed rapidly. Radiative driving of mass outflows was first suggested in the 1920s by E. Milue, although it remained unwed to hydrodynamics until after the development of stellar wind theory by E. Parker and others in the 1960s. The first analytic theories were developed in the early 1970s by L. Lucy, J. Castor, and their collaborators. These have been substantially extended through the use of numerical models, which include upward of hundreds of thousands of atomic transitions in the calculation of radiation pressure. Time-dependent calculations are now becoming feasible for massive stellar envelopes.

The explosion of SN 1987A in the Large Magellanic Cloud has suddenly and dramatically brought this field into a sharper focus. Information gained from the multiwavelength light curves, the fact that the star was a blue supergiant at the time of the explosion, and the fact that the nucleosynthesis can actually be studied in the ejecta all mean that models can be better constrained than ever before. More than any single event in the history of the study of the stellar evolution, this supernova has redefined the questions asked by theorists about the origin and development of massive stars.

II. MASS LOSS FROM STARS

Stellar masses are not fixed, although they change for single stars only on long periods. Low-mass stars, like the Sun, lose material through modest winds, about $10^{-14}\, M_\odot$ yr^{-1} while in the hydrogen core burning (main sequence) phase. Turbulence and magnetic field dissipation appear responsible for such mass outflows. Red giants support stronger outflows, likely by the same mechanism for the lower mass stars. Such a star, losing mass at the thermal rate, will have a wind of about $10^{-7} M_\odot\ \mathrm{yr}^{-1}$. Such winds are common for convective stars in the red giant phase.

For luminous stars, the intensity of the envelope radiation field may be sufficiently great that radiation exerts a pressure on the background gas. The reason for this is physically simple. Photons carry not only energy $E = h\nu$ but also momentum, $p = h\nu/c$, where h and c are Planck's constant and the speed of light, respectively, and ν is the frequency. If a photon is scattered, its energy may be unchanged—except in the Compton effect or if there is resonant redistribution of the energy in the line transition responsible for the scattering—but the momentum vector changes direction. Any absorption or scattering of a photon transfer momentum from the photons to the matter.

To calculate the effective radiative acceleration, it is necessary to take into account the effects of absorption and scattering by stellar line and continuum transitions. The available lines are quite numerous and densely distributed in the ultraviolet. In this wavelength region, the radiation field for a massive hot star is intense, and if there are many lines near the peak of the photospheric distribution of intensity, the radiative acceleration will become quite

large. The precise definition of the radiative acceleration is therefore the following:

$$g_{\rm rad} = \frac{1}{P}\frac{d}{dr}P_{\rm rad} = \frac{\pi}{c}\sum_i \int \kappa_{\nu i} F_\nu \, d\nu, \qquad (1)$$

where the sum is taken over all ionic species in the stellar envelope at a given depth with fractional abundance X_i. The density of the envelope is ρ, κ_ν is the opacity, and F_ν is the monochromatic flux. This factor may be quite large compared with the predicted acceleration because of electron scattering, since the opacity is wavelength dependent. The most effective acceleration occurs when the peak of the photospheric flux occurs in a wavelength regime of very large line opacity. In such cases, the radiative acceleration may dominate over gravity. The envelope is then unbound and flows away from the star.

If electron scattering, also called Thomson scattering, is the only opacity source, the opacity is said to be gray, that is, independent of frequency. The radiation pressure felt by the scatterers in ensemble is expressed as follows:

$$\frac{d}{dr}P_{\rm rad} = \frac{L\kappa\rho}{4cr^2}. \qquad (2)$$

Here L is the luminosity and κ is the mean opacity per gram of material.

There is a critical luminosity at which the radiation pressure precisely balances the gravitational acceleration. Called the Eddington luminosity, this is given by the following:

$$L_{\rm Edd} = \frac{4\pi cGM}{\kappa} \approx 3\times 10^4 \left(\frac{M}{M_\odot}\right) L_\odot \qquad (3)$$

for electron scattering ($\kappa = 0.4$ cm^2 g^{-1}). Above this luminosity, the effective gravitational acceleration at the surface of the body is negative, so that the matter is actively accelerated outward by the radiation. The factor $\Gamma = L/L_{\rm Edd}$ measures the departure of the gravitational acceleration, g, from that of the mass alone, thus $g_{\rm eff} = g(1-\Gamma)$. For a given mass, the scattering due to electrons provides an upper limit on the stable mass, and any increase in the opacity will produce a greater tendency of the stellar material to be lost by the star.

The equation of motion for a steady-state, spherically symmetric stellar wind is given by the following:

$$\begin{aligned} v\frac{dv}{dr} &= -\frac{1}{\rho}\frac{dP_{\rm gas}}{dr} - \frac{GM}{r^2} - \frac{dP_{\rm rad}}{dr} \\ &= -\frac{1}{\rho}\frac{dP_{\rm gas}}{dr} - g_{\rm eff}. \end{aligned} \qquad (4)$$

If the gas is isothermal, the equation of state is $P_{\rm gas} = \rho a_{\rm s}^2$, where $a_{\rm s}$ is the sound speed. The radiative acceleration is a function of depth in the atmosphere, so the flow accelerates outward. If at some point in the envelope the effective gravity is zero, the wind becomes supersonic. It is impossible for a static solution to be achieved once the flow speeds exceed the sound speed, so the matter has no alternative but to leave the star and reach a terminal speed at some distance that is large—generally a few times larger than the escape velocity. Such an outflow is called a radiative stellar wind because the driving force is radiation pressure.

In steady state, the condition of the continuity equation is that $\nabla \cdot \rho v = 0$. The mass loss rate for spherical (radial) flow is given by the following:

$$\dot{M} = 4\pi r^2 \rho v. \qquad (5)$$

The mass loss rate is fixed by the conditions at the sonic point, near the stellar photosphere. For low-mass, main sequence stars, radiation pressure plays no role in the structure of the envelope and thus is incapable of producing any serious mass loss. From the cool stars, such as red giants, despite their luminosities, the radiation field does not produce a large acceleration. This is because of the wavelength distribution of the absorption lines with respect to the continuum and the fact that the average momentum carried by the photons is quite low. Only if the surface gravity is very low, for example, for supergiant stars, which are very evolved, will significant radiative driving take place. It is generally assumed that dust plays the primary role in coupling the radiation to matter in the atmosphere. However, for the hot massive stars, the luminosity is carried by the ultraviolet photons, which encounter large line opacities, and the accelerations may be very large. In fact, predicted mass loss rates range from $10^{-10}\ M_\odot$ yr^{-1} for 5 $M_\odot$ stars to about $10^{-5}\ M_\odot$ yr^{-1} for 40 or 50 $M_\odot$ stars. For these massive stars, the effect of the mass loss may be quite severe. This property of the most massive stars creates most of the difficulty in present evolutionary calculations.

The winds for massive stars reach terminal velocities that are of order 3000 to 5000 km sec^{-1}. When this matter mixes with the interstellar medium, it dissipates the transported energy and momentum. In the lifetime of the star, the combined effects of radiation and wind amount to nearly as much energy being put into the medium as in a supernova event. We return to this point below.

III. MASSIVE STELLAR EVOLUTION AND NUCLEOSYNTHESIS

A. Formation

The most massive stars in galaxies appear to form in groups. This is virtually all we can say about the origin

of these objects at present. The associations in which they form range in mass from several hundred to several thousand solar masses, often heavily weighted toward the upper mass end. In general, stars form in a variety of clusters and associations, with a mass function that is skewed toward the lower mass end. This distribution in mass, known as the initial mass function (IMF), may be a universal property of star formation, sampled differently in different regions of star formation, but this is still controversial.

The distribution of the very massive stars in other galaxies, such as the Large Magellanic Clouds (LMC), makes it clear that these stars often form in tightly bound, compact subgroups within more extensive complexes. One of the best examples is 30 Doradus, an enormous H II complex near the bar of the LMC. The triggering mechanism is not well understood, but one point is certain: Once a cloud begins to form massive stars, it is permanently changed. Such stars may induce star formation throughout the cloud, leading to its disruption on short (10^6 yr) time scales.

B. Main Sequence Stages

During most of their lifetimes, massive stars are dominated by their stellar winds. Their masses are never constant, but are constantly decreasing as their luminosities change. On the main sequence they burn hydrogen into helium via the CNO process, during which phase they possess massive convective cores and radiative envelopes. The stellar luminosity of the mast massive main sequence stars is very close to the Eddington limit for most of their lives. Maeder (1987) gives the following mass–luminosity relations for the massive stars on the main sequence:

$$\log \frac{L}{L_\odot} = 2.55 \log \frac{M}{M_\odot} + 1.27 \qquad (6)$$

for $15 \leq M/M_\odot \leq 40$

and

$$\log \frac{L}{L_\odot} = 1.84 \log \frac{M}{M_\odot} + 2.42 \qquad (7)$$

for $40 \leq M/M_\odot \leq 120$. It is possible that the most massive stars, above about 60 $M_\odot$, do not have a real main sequence stage in the sense of being globally static. Their surfaces may be merely opaque layers in their wind. As a result, the atmosphere is likely dynamic and more extended than a stationary layer. This, if correct, alters the age determination from fitting the main sequence isochrones to clusters that are computed with conventional assumptions.

Stars more massive than about 25 $M_\odot$ lose as much as, or more than, 10% of their mass while still in the hydrogen core burning phase of evolution. There is a rapid increase in this rate for the stars more massive than 25 $M_\odot$, reaching upward of 30% for 80 $M_\odot$ stars. The main sequence lifetime for these stars is more difficult to assess than for lower mass objects. Because their mass continually decreases even during the hydrogen core burning stage of evolution, their lifetimes are longer than would be anticipated from constant-mass models:

$$t_{\mathrm{MS}} = 1.3 \times 10^8 \left(\frac{M}{M_\odot} \right)^{-0.86} \mathrm{yr} \qquad (8)$$

for stars with masses $15 \leq M/M_\odot \leq 60$. The greatest uncertainty in these figures is the rate of mass loss while the star is on the main sequence. This depends critically on the prescription used for calculating the relation between the stellar luminosity and the mass loss in the stellar wind.

Winds are observed for luminous stars. The question is not whether they occur, but what drives them and how the driving depends on the stellar properties, such as mass, radius, and luminosity. If driven by turbulence or some other form of direct mechanical input, the dependence on the luminosity is through the surface temperature and radius of the star. On the other hand, for radiatively driven winds, the driving depends on the envelope opacity and consequently the metal abundances. Presently, there is considerable uncertainty associated with these choices and the reader is referred to Maeder and Chiosi (1994) for a tabulated listing of currently proposed theoretical and empirical laws for M as a function of stellar parameters.

The main sequence lifetime is also affected by core convection. There are still uncertainties in the physics of stellar convective energy transport, particularly concerning the overshooting of convective cells at the boundary of the CNO-processed core and overlying radiative envelope. Mixing of envelope material, which is hydrogen rich, into the helium core extends the lifetime of the star in the hydrogen core burning stage and increases the core mass relative to the envelope. It also extends the period of subsequent helium core burning and thus alters the entire subsequent lifetime of the star. While population statistics of stellar distributions near the main sequence on the Hertzsprung–Russell diagram (HRD) seem to require some mixing during this early stage of evolution, the mechanism for such mixing is an open problem.

As helium builds in the core, the star begins to evolve toward the red giant branch, with increasing luminosity and larger radius. Here again, the effects of mass loss play an important role. As the surface gravity drops and the luminous output of the nuclear source increases, the stellar wind becomes more powerful. The combined effect of core contraction and increased mass loss produces nearly constant luminosity evolution across the HRD. The star's trajectory is then cutting across isochrones computed for

constant mass models with more than 30% of the mass lost by the time of carbon core ignition. The core of a massive star evolves almost independently from the mass of the envelope, so the nucleosynthetic details are less sensitive to mass loss prescription than are the surface properties.

C. Post-Main-Sequence Evolution: Helium to Iron Core Phases

Helium ignites in the stellar core under nondegenerate conditions while the star is still near the main sequence. In rapid succession, He is converted to oxygen via carbon through the successive capture of He nuclei. Some of the O is converted to neon, $^{16}O(\alpha, \gamma)^{20}Ne$. Most of the carbon and oxygen converts to magnesium and then silicon via oxygen burning. Finally, the core builds to sulfur and then to iron. Throughout this sequence, core homogeneity is maintained by convection. Table I, gives some representative numbers for the relevant time scales and main sequence luminosities.

During these successive core nuclear burning stages, the inner envelope is also processed. The energy released through the core reactions increases the temperature at the base of the envelope, and an onion skin effect is generated, with progressive motion of the nuclear processed material outward to a larger mass fraction of the star. The entire process is almost independent of mass above about 20 $M_\odot$, and at the stage at which the Fe core is produced, about half of the mass has been converted from hydrogen into heavier elements.

D. Surface Abundance Changes on Evolutionary Time Scales

It would be incorrect, however, to think that the core evolution is without effect on the surface properties of the star. As the more energetic, and short-lived, phases of nuclear processing are reached, the luminosity of the star increases, thereby powering progressively more intense mass loss. This strips off the outer layers of the star and appears to bring nuclear-processed matter to the stellar surface. The precise mechanism for this is presently unknown but is assumed to be caused by turbulent mixing. An important point to note here is that regardless of the details, CNO-processed material is observed to appear at the photosphere in many evolved massive stars. There is strong support from observations for the idea that at least some effect of the mass loss can be detected at the surface of the star long before it has stripped down to the zones that have actually been burned deep in the interior. It is also likely that an attendent dynamical mixing process, perhaps powered by shear flows in the core boundary layer, occurs.

Another very important aspect of the evolution is that, as the mass loss rate increases with decreasing surface gravity, the envelope begins to strip off the progressively processed layers, and eventually the redward trend of motion of the star on the HRD is halted and reversed. Stars more massive than about 50 $M_\odot$ are predicted never to enter the red giant phase at all. Instead, they turn around at surface temperatures of about 8000 K and become Wolf–Rayet stars.

TABLE I Evolutionary Time Scales for 15 to 120 $M_\odot$ Stars[a,b]

$M/M_\odot$	$\log(L/L_\odot)$	$\tau_{\text{H core}}$	$\tau_{\text{He core}}$
15	4.25	11.58	2.34
20	4.60	8.46	1.89
25	4.85	6.62	1.22
40	5.34	4.53	0.86
60	5.70	3.71	0.71
85	5.98	3.25	0.76
120	6.23	2.81	0.84

[a] From Maeder, A. (1987). *Astron. Astrophys.* **173,** 247.
[b] Times given in years $\times 10^6$.

E. The Wolf–Rayet Stars

The Wolf–Rayet, or WR, stars come in three varieties, depending on the dominant species in their optical spectra. Called WN, WC, or WO for nitrogen, carbon, or oxygen, they form a sequence of both luminosity and temperature in the hot, high-luminosity part of the HRD. The WR star subclasses are further subdivided into temperature classes on the basis of line strength and excitation. For instance, WN7 stars are cooler than WN3 stars. The Of stars, which show weak emission in optical nitrogen and ionized helium lines, overlap with the WN7/8 stars and may form a continuous sequence.

Because some WR stars are binaries, we have some idea of their masses. Their masses range from as low as 7 $M_\odot$ (for HD 94546) to upward of 50 $M_\odot$ (for HDE 311884). For a few eclipsing systems, such as CQ Cephei and GP Cephei, the masses are better determined and lie between 10 and 20 $M_\odot$. On the basis of their luminosities, only lower limits would be possible by assuming mass loss at the Eddington rate. The WN stars often have normal, main sequence, massive companions. Since the WR is often the lower mass star in the system, it is likely that it started out as the most massive and evolved more rapidly. The eclipsing binary V444 Cygni is an excellent example of such a system: its WN star is about 10 $M_\odot$ while its companion is an O6 III star of about 30 $M_\odot$. The WC stars

are also observed in binaries, but here the companion is frequently a compact object, most likely a neutron star. A representative star in this class is HD 50896. The WO stars, a rare group, are the most luminous and hottest of the WR stars. They show extreme stages of ionization in their optical spectra, such as O VI, and are likely the most denuded of the sequence.

It appears that the Wolf–Rayet stars form the sequence of stripping in which the nitrogen enhancement is the least extreme, with the carbon and oxygen revealing progressively deeper layers of nuclear-processed material. The mechanism for this process is not well understood, but the binarity may be a clue. In the current scenario, WR stars are produced in massive close binaries by tidal interaction limiting the maximum radius to which either component can expand.

As evolution drives a star toward a larger radius in a close binary system, the mass cannot remain bound to the star—independent of the presence of a stellar wind—beyond a critical radius, called the Roche limit. This radius depends on the period, total mass, and mass ratio of the components of the binary. On further expansion, once the star reaches this surface, some of the stellar envelope is lost from the system as a whole and some is accreted onto the companion. The loss of mass and angular momentum from the system drives the stars closer together and accelerates the stripping process. In systems consisting of two massive stars, wind collisions produce X-rays that have been observed to show orbital modulation. The mass-loser strips farther down into its interior than would have been possible through radiative wind action alone, and a close binary with a WR component is produced. Many details still remain to be understood about this model before its application to the origin of the Wolf–Rayet stars can be declared successful.

The WR stars are often surrounded by thin shell-like nebulae. These ring nebulae, formed by the action of the stellar wind on the interstellar medium, are known as bubbles. They are less dusty, and more energetic, than the winds from the luminous blue variables (LBVs) (see Section V). It is likely that these shells structure the environment around the core into which the supernova ejecta plow during the explosion of the remnant of the WR stage.

F. The Final State Question: What Does a Presupernova Star Look Like?

Is a Wolf–Rayet star the precursor to a supernova? This is still a controversial matter. It appears that the underluminous historical supernova, Cas A, which exploded apparently without notice in the 17th century, was one of these stars. It was probably a WO on the basis of the observed abundances in the optical filaments and the lack of hydrogen in the ejecta. While exceedingly energetic, the shocks from WO stars do not have much overlying matter as they transit the stellar envelope, so the ejected material cools rapidly without producing an optically bright blast wave. The normal Type II supernova arises from a less evolved stage of the star, for instance a red or blue supergiant. In these stars, the envelopes are massive and remain hot and luminous for much longer, nearly a year, before they become optically thin and cool rapidly. The long plateau observed in the normal SN II light curve is generally ascribed to this envelope structure. Thus, the stage that precedes a supernova is set up by the rate of stellar mass loss, so there is still considerable work to be done on the driving mechanisms for that phenomenon before definitive answers can be given to the questions concerning the presupernova stages of evolution. At least one type of γ-ray burst may result from hypernovae, in which up to 10^{52} erg is released by the shock propagating through the WR wind.

The precise appearance of the stellar surface at the stage of Fe core formation is still a hotly debated question. Because of the role played by mass loss in the final stages of stellar evolution, the star may end up either as a blue or red star at the end of its life. The Wolf–Rayet stars are candidates for the final, presupernova state. However, it should be remembered that the precursor of SN1987A in the LMC was a B3 supergiant, and model calculations show that for a 15 to 25 $M_\odot$ star the iron core forms while the star is a red supergiant. Clearly, the end stages of massive stars are a fruitful area for future work.

The formation of successively heavier nuclei by fusion cannot continue past the formation of Fe. The binding energy of this nucleus is the largest possible for a stable isotope, so that subsequent processing can only lead to destruction of the nucleus. This involves endoergic reactions, robbing the star of its gravitational energy and producing the rapid disintegration of the nuclei. Cooling proceeds via neutrino emission, which is rapid enough that the star cannot quasistatically adjust to the energy losses and begins to dynamically contract. The formation of Fe is, therefore, the beginning of the end for a massive star. After this stage, only dynamical collapse and the consequent supernova explosion are possible.

IV. EFFECTS OF MASSIVE STARS ON THE GALAXY

The energy input from stars occurs in two forms: (1) radiatively through the emission of light from their photospheres and (2) mechanically through the action of stellar winds and supernovae. During the course of its life, a star

of about 50 $M_\odot$ puts in roughly equal amounts of energy to the interstellar medium from winds, radiation, and its final demise in a supernova explosion. The principal difference between these modes is the time scale.

A. Winds and H II Regions

The winds from the massive stars account for the formation of many of the expanding rings observed around OB associations in the Galaxy. With typical mass loss rates of $10^{-5}\ M_\odot\ \mathrm{yr}^{-1}$ and wind speeds of about 3×10^3 km sec^{-1}, the average OB star puts some 3×10^{37} erg sec^{-1} into the interstellar medium, about $10^4\ L_\odot$. Over its lifetime, this amounts to about 10^{51} erg. The shell thus generated by the expansion of the wind compresses the interstellar gas, sweeping it up and slowing down in the process, In the Wolf–Ravet stage, this momentum input can increase by as much as an order of magnitude.

The hottest, most massive stars ionize the medium surrounding them, forming so-called H II regions. These regions of ionized hydrogen (and other elements) are initially static, with equilibrium radii determined by the rate of input of Lyman continuum photons with energies above 13.6 eV. The radius of such a region, first determined by Strömgren (hence called a Strömgren sphere) is given by the following:

$$n_e^2 \alpha R_{\mathrm{H\ II}}^3 = N_{\mathrm{LyC}} = n_0 \kappa \int_{\nu \mathrm{LyC}}^{\infty} \frac{L_\nu}{h\nu}\, d\nu, \qquad (9)$$

where α is the recombination rate, n_e is the electron density, and N_{LyC} is the number of Lyman continuum photons. Within this radius, every atom will be ionized, and $R_{\mathrm{H\ II}}$ is limited only by the exhaustion of the Lyman continuum photons. The initial expansion varies as $R_{\mathrm{H\ II}} \sim t^{1/3}$ in the Strömgren sphere stage.

Such a structure cannot remain in equilibrium for long, and with the continuing input of radiative and mechanical energy from the exciting stars, the boundary of the H II region begins to expand. This is augmented by the input of the winds from the massive OB stars. The outer portion of the H II region quickly becomes supersonic and develops a compressional zone. As matter is swept up by the shock, it is first compressed and then ionized, and the hot confined gas of the H II region begins to rarify. If driven by radiation pressure or mass loss from the stars, the radius of the shell expands in time as $R_{\mathrm{H\ II}} \sim t^{1/2}$, while if driven by an impulsive event, the initially adiabatic stage expands as $R_{\mathrm{shell}} \sim t^{2/5}$, and the later momentum-conserving (snowplow) stage expands as $R_{\mathrm{shell}} \sim t^{1/4}$. Material from the stellar wind mixes with the interstellar gas during this phase. Some of the matter, if it comes from Wolf–Rayet stars, will have been processed and it is likely that the most massive red supergiants will also enrich the medium in CNO products.

B. Induced Star Formation

Should the star formation begin within a molecular cloud, the winds and H II regions can either destroy the cloud by heating it up through radiative and mechanical processes or they can break free of the cloud. This latter type of star formation event is called a champagne flow, analogous to that commonly observed at weddings. The sudden release of pressure by the breakout of the shock from its environment causes a rapid outflow of material from the hot H II region surrounding the OB stars. This flow further drives a shock into the molecular cloud via momentum conservation and compresses the already dense material of the cloud core. Such an event may initiate collapse from gravitational instability, at least locally to the OB stars. Thus, the massive stars are not only capable of destoying the cloud environments in which they have formed, they can also serve as agents for propagating star formation through the cloud.

The formation of the massive stars is well traced by the radio and infrared flux emitted by the H II regions that surround them. Even the densest parts of molecular clouds are not optically thick longward of about 60 μm, so that far infrared (FIR) and radio photons freely escape the cloud. The argument that permits these to determine the rate of massive star formation is as follows. Every massive star that forms an H II region will be surrounded by a thermal, radio-emitting plasma with a temperature of about 10^4 K. The rate of radio emission depends on the rate of ionization, which in turn depends on the luminosity of the central stars. This in turn depends on the mass of the stars. Each radio photon is associated with the ionized medium, while each recombination eventually leads to a Lyα photon through radiative cascades. These photons, which are trapped in the optically thick H II regions, eventually collide with dust grains in the cloud and the H II region and are absorbed. The dust reradiates this energy at equilibrium in the FIR. Thus, there is an expected correlation between L_{FIR}, the far infrared luminosity, L_{radio}, and ψ, the star formation rate per unit mass.

Such a correlation is observed in regions of active star formation in the Galaxy and nearby galaxies, although it is still not completely certain what the implied rates of star formation mean. For some galaxies, called starburst galaxies, the rate of massive star formation can reach more than $10^3\ M_\odot\ \mathrm{yr}^{-1}$. This enormous rate cannot continue for very long before molecular cloud material is consumed, but the triggering mechanism is not currently understood. Such systems may be the result of collisions between galaxies and are often seen in galactic systems undergoing merger.

V. THE LUMINOUS BLUE VARIABLES

The discovery of the outburst of η Carinae by John Herschel (1847) was of singular importance for the study of the most massive stars. During the nearly 10 yr following its discovery in outburst, this star was the brightest object in the Galaxy, becoming one of the brightest stars in the sky after Sirius and Canopus, both of far lower intrinsic luminosity. At its peak, η Car reached a luminosity in excess of $10^6\ L_\odot$. The total radiative output during the 30 yr following the eruption may have been as large as 2×10^{49} erg. The outburst was accompanied by the ejection of some debris, which is now just visible in the core of the nebula surrounding η Car, itself the product of a previous eruption some thousands of years ago. Presently, the star has a surface temperature of about 20,000 K, a luminosity of order $7 \times 10^5\ L_\odot$, and a mass loss rate of about $10^{-1}\ M_\odot\ \mathrm{yr}^{-1}$.

η Car was not the first such variable to be observed; P Cygni had also undergone a similar brightening in the 17th century, but it was the first since the advent of telescopic observation.

While for many years η Car has been assumed to be a protostar, both because of its surrounding nebulosity and its location in the galactic plane, ultraviolet spectroscopic observations by T. R. Gull, K. Davidson, and N. Walborn in 1982 showed that the filaments are enriched in nitrogen. This is the signature of a highly evolved massive star and focused attention on the end state problem for such objects. Similarly, nitrogen-enriched filaments have been found around the galactic LBV star AG Car.

Herschel's description of the event is most important for its comparison with the Hubble–Sandage variables. First described for the Local Group by E. Hubble and A. Sandage in 1953, these stars are the most luminous stars in their parent galaxies. They were first recognized and studied in M31 and M33. They have since been observed in several other nearby systems. Several have been known in the Large Magellanic Cloud as S Doradus variables, after their prototype, which is an A supergiant in the LMC. These stars show large-amplitude (upward of 5 magnitudes) variations on decade-long time scales. They are also variable, with smaller amplitude, on time scales of months to years.

During the course of an eruption, the spectrum and light continuum of the star change dramatically, although multiwavelength observations show that the *bolometric* luminosity remains constant. The optical variations are produced by flux redistribution due to enormous increases in the UV line opacity from the increased mass loss rate. While normally appearing as an OB supergiant, once the ejection phase occurs, the star develops a massive wind that becomes optically thick in the Lyman continuum. The effective temperature, given by $T_{\mathrm{eff}} = (L/4\pi\sigma R^2)^{1/4}$, drops rapidly. Here, R is the shell radius and σ is the Stefan–Boltzmann constant. The bolometric luminosity remains constant, so the peak of the radiation from this pseudophotosphere shifts from the ultraviolet to the near infrared through the visible, which produces the marked increase in the brightness at optical wavelengths. The shell specturm develops strong, windlike P Cygni line profiles with expansion velocities of 200 km sec^{-1} up to 3000 km sec^{-1}, and the star begins to look like a late A supergiant. This has been well observed in this past decade with the outburst of R 127 in the LMC and is well known for S Dor and several of the M31 variables, AE And and AF And, and also in M33. In a few cases, the shell may become cool enough to look like an M supergiant, with a temperature as low as 3500 K. Variable A in M33 appears to be one of these stars. As the shell dissipates on the turnoff of the massive wind stage, the photosphere gradually recedes backward in the matter, and the photosphere is again revealed as an OB supergiant. We do not presently know what turns off the ejection event. Figure 1 shows the location of some of the LBVs known in the Galaxy, LMC, SMC, M31, and M33.

The cause of the large mass loss episodes is not well understood. Massive stars are pulsationally unstable and may display impulsive mass loss because of the amplitude of the pulsations. Many of the brightest stars in the Magellanic Clouds, called hypergiants by C. de Jager, show long-term fluctuations with amplitudes of about one magnitude and time scales of months to years. These variations are, however, far smaller than the eruptions that characterize the Hubble–Sandage variables.

One possible picture is that the radiative winds of massive stars are unstable to small changes in the mass loss rate. The temperature of the envelope depends on the opacity in the wind. If the mass loss rate begins to increase, perhaps because of some change in the luminosity of the star or pulsational instability of the envelope, the density increases, causing the temperature in the outer parts of the wind to drop. Lines of lower than normal ionization, such as Fe II, strengthen and produce an increased radiative acceleration. This in turn causes the density to increase still more, leading to an instability that appears to saturate at about $10^{-3}\ M_\odot\ \mathrm{yr}^1$. The ejection events in P Cyg and η Car are well reproduced by this mechanism, although the seat of the initial event is not known.

The resultant light variations of the star are thus well in agreement with the observations. As discussed above, the bolometric luminosity of the star remains constant as the mass loss increases. The exception is η Car, which appears to require an additional energy source to power its emission, possibly mechanical due to wind–environment interaction. In contrast, as the envelope opacity increases,

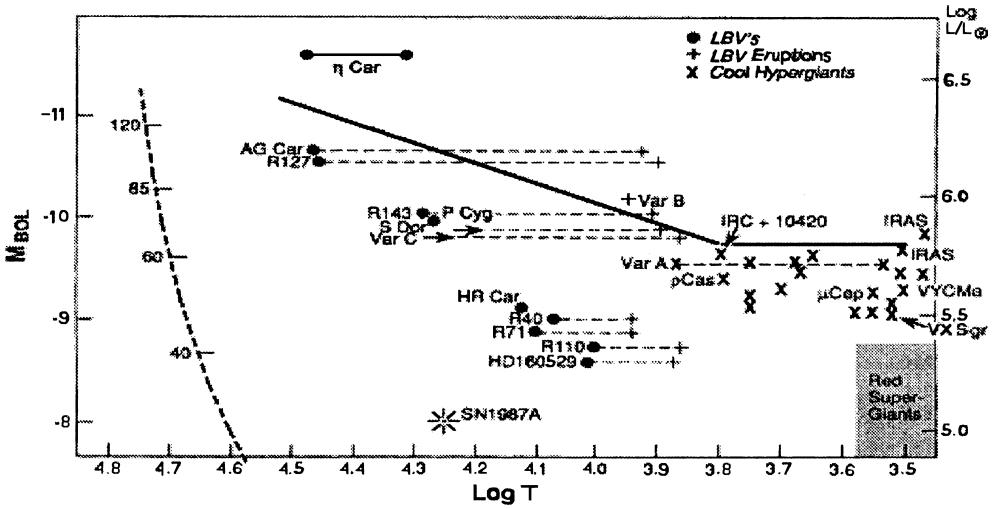


FIGURE 1 Location and trajectories of luminous blue variables in Local Group galaxies. The main sequence (left) is labeled by stellar mass. The bold line marks the red limit for massive stars (Humphreys-Davidson limit). Normal red supergiants are located in the shaded box, lower right. [This figure originally appeared in the *Publications of the Astronomical Society of the Pacific* (Humphreys and Davidson, 1994, *PASP*, 106, 1025), copyright 1994, Astronomical Society of the Pacific; reproduced with permission of the Editors.]

the peak of the envelope radiation shifts into the visible and finally into the infrared. As the mass loss slackens off, the envelope clears and the star settles back into a fainter visual, but brighter ultraviolet, state. The behavior of several LBVs in the Galaxy and the Large Magellanic Cloud has been followed with ground-based optical spectroscopy, in the ultraviolet by the International Ultraviolet Explorer satellite and HST, and now in X-rays with *Chandra*; it appears to conform to this expected behavior.

Several LBVs have been discovered to be surrounded by massive cold shells. R 127 = HD 269858 f in the Large Magellanic Cloud began an outburst during the 1980s and is surrounded by shells from several previous eruptions. The galactic stars He3-519 and AG Car also have such shells. η Car has a complex nebulosity, formed from the ejecta from several previous events. Imaging with the wide field camera on the Hubble Space Telescope has revealed several complex jetlike and lobe structures on the subarcsec scale in the inner nebula, possibly the products of the last eruption in the 19th century. All of these shells have about 10^{-3} to 10^{-2} $M_{\odot}$, a considerable amount of matter for single ejection events. The winds from these stars must have increased substantially during these events, which took place some 10^4 yr ago. In nearly all cases, the ejecta appear bipolar, as do the ionized rings of SN 1987A. The ordering appears to be due to an aspherical wind, equatorially enhanced and slowed presumably by angular momentum conservation. The structures are reminiscent of planetary nebulae.

Yet, despite these findings, the triggering mechanism for some of the outbursts is still problematic. The variations in η Car appear to defy description. The luminosity of this star was so great during the outburst in the last century that it is in excess of what is likely possible for the known bolometric luminosity of the star. Yet another mechanism may be at work here. It may be linked to the 5.5 yr binary that lies imbedded in the ejecta.

VI. CODA

Rare, bright, and physically problematic still seems the best characterization of the most massive stars. Their formation and death are not well understood, and there are still considerable uncertainties associated with the stages in between. They bring into sharper view the most extreme physical processes of nucleosynthesis, mass loss, and stability found in steller environments. They will likely remain true challenges to observers and theorists for decades to come.

SEE ALSO THE FOLLOWING ARTICLES

BINARY STARS • GALACTIC STRUCTURE AND EVOLUTION • NEUTRON STARS • PULSARS • SOLAR PHYSICS • STAR CLUSTERS • STARS, VARIABLE • STELLAR SPECTROSCOPY • STELLAR STRUCTURE AND EVOLUTION • SUPERNOVAE

BIBLIOGRAPHY

Abbott, D. C., and Conti, P. S. (1987). The Wolf–Rayet stars. *Annu. Rev. Astron. Astrophys.* **25,** 113.

Arnett, W. D. (1997). "Supernovae and Nucleosynthesis," Princeton, NJ, Princeton Univ. Press.
Bernasconi, P. A., and Maeder, A. (1996). *Astron. Astrophys*. **307,** 829.
Boland, W., and van Woerden, H., eds. (1985). "Birth and Evolution of Massive Stars and Stellar Groups," Reidel, Dordrecht.
Cassinelli, J., and Lamers, H. J. G. L. M. (1999). "Theory of Stellar Winds," Cambridge, UK, Cambridge Univ. Press.
Castor, J., Abbott, D. C., and Klein, R. (1975). *Astrophys. J.* **195,** 157.
Davidson, K. R., and Humphreys, R. M. (1997). *Annu. Rev. Astron. Astrophys*. **35,** 1.
Davidson, K., Moffatt, A. J., and Lamers, H. J. G. L. M. eds. (1989). "Physics of Luminous Blue Variables: IAU Colloquium 113," Reidel, Dordrecht.
Franco, J., Terlevich, R., Lopez–Cruz, O., and Aretxaga, I., eds. (2000). "Cosmic Evolution and Galaxy Feedback: Structure, Interactions, and Feedback," Astronomical Society of the Pacific, San Francisco.
Humphreys, R., and Davidson, K. (1994). *Publ. Astron. Soc. Pacific* **106,** 1025.
de Jager, C. (1980). "The Brightest Stars," Reidel, Dordrecht.
Kudritski, R. P., Pauldrach, A., and Puls, J. (1987). *Astron. Astrophys*. **173,** 293.
Langer, N., and El Eid, M. F. (1986). *Astron. Astrophys*. **167,** 265.
Maeder, A. (1987). *Astron. Astrophys*. **173,** 247.
Maeder, A., and Chiosi, C. (1994). *Annu. Rev. Astron. Astrophys*. **32,** 227.
Maeder, A., and Meynet, G. (1994). *Astron. Astrophys*. **287,** 803.
Nota, A., Livio, M., Clampin, M., and Schulte–Ladbeck, R. (1995). *Astrophys. J.* **448,** 788.
Schaller, D., Schaerer, D., Meynet, G., and Maeder, A. (1992). *Astron Astrophys. Suppl*. **96,** 269.
Shore, S. N., and Sanduleak, N. (1984). *Astrophys. J. Suppl*. **55,** 1.
Shore, S. N., Altner, B., and Waxin, I. (1996). *Astron. J.* **112,** 2744.
Sziefert, T., Humphreys, R. M., Davidson, K., Jones, T. J., Stahl, O., Wolf, B., and Zickgraf, F.-J. (1996). *Astron. Astrophys*. **314,** 131.
Walborn, N., and Fitzpatrick, E. (2000). *Publ. Astron. Soc. Pac*. **112,** 50.

Stars, Variable

Steven N. Shore
Indiana University, South Bend

GLOSSARY

Hertzsprung–Russell diagram (H-R diagram) Plot of surface, or effective, temperature versus luminosity for a star. Observationally, the locus of the observed stellar population in a color–magnitude plane. Theoretically, the description of the path covered by the surface of a star with time.

Instability strip Portion of the HR diagram, spanning the temperature range from about 3000 to 8000 K and from the main sequence to the supergiants, in which most periodic stellar pulsational instabilities are detected. The strip is characterized by a period–luminosity (P-L) relation, with increasing period corresponding to higher intrinsic brightness.

Overstability Growing, but periodic, instability characterized by a complex frequency.

Population I and II Characterized mainly by age and metallicity differences. Population I is the young, stellar galactic component, with metallicities about the same as the Sun's and dynamical confinement to the plane of the Galaxy. Population II, found in the galactic halo and globular clusters, consists of older, low-mass stars with below solar metal abundances.

Secular instability A nonperiodic, or steady, growth of the pulsation with time.

Units Solar luminosity, $L_\odot$, 4×10^{33} ergs^{-1}; solar mass, $M_\odot$, 2×10^{33} g; solar radius, $R_\odot$, 7×10^{10} cm; mean solar density, $\langle \rho_\odot \rangle$, 1.4 g cm^{-3}.

VARIABLE STARS are stars that vary in photometric output over any period of time. The manifestation of periodic behavior is rare among stars, resulting from resonance phenomena in the interior. The primary mechanism for intrinsic stellar variability is pulsation, which may be either radial or, in rapidly rotating stars, nonradial. Magnetic fields also play a role in structuring the variations. Mass loss and shock phenomena are also associated with large amplitude pulsators, which may affect stellar evolution. This article deals exclusively with pulsating variables.

I. INTRODUCTION

Like most astronomical discoveries, the first observations of stellar variability were accidental. Fabricius noted, in

1596, that the star *o* Ceti showed a range of variability from invisibility to second magnitude. In honor of this, he named the star Mira, the wonderful. This discovery served as confirmation of the view, first forcefully promulgated in the Latin west by Tycho Brahe following the "new star" in Cassiopea of 1572, that the heavens are not static.

A major step was achieved in the late 18th century with Goodricke's discovery, announced in 1782, of the variations of β Persei, also called Algol. The variations were observed to be periodic and stable, and Goodricke explained this by the model of a close, unresolved binary star system undergoing eclipses. This is the first explanation for stellar variability, and one which served as an effective stimulus to search for other binary stars. The statistical reality of binarity and the application of Newtonian mechanics made a potent combination. In 1784, Goodricke discovered the variability of δ Cephei (δ Cep); in the same year, Piggot discovered the variability of η Aquilae (η Aql). The eclipse model was applied at the time to both of these stars, whose periods are similar to that of Algol. Both δ Cephei and η Aql are now known to be pulsating variables, unrelated to the Algol class of binaries, but the general explanation of the variations did not become clear for more than 150 years after this discovery. Table I, adapted from John Herschel's 1847 edition of *Outline of Astronomy*, gives a list of some of the earliest discoveries of what are now known to be intrinsically variable stars.

During the next century, the search for stellar variability grew in importance to astronomy. The work of Argelander at Bamberg, the establishment of the *Bonner Durchmusterung* star atlas, and the international efforts to construct the *Carte du Ciel*, a photographic all-sky atlas, stimulated the discovery and cataloging of variable stars.

Perhaps the most important group working at the end of the 19th century was that at the Harvard College Observatory. S. Bailey, in observing globular clusters, discovered a class of periodic variables that showed the same characteristics as the field variable RR Lyrae, and the search for such stars was extended to the southern hemisphere with the establishment of the Bruce telescope at the Harvard Southern Station in Peru.

TABLE I Early Discoveries of Intrinsic Stellar Variables

Star name	Discoverer	Year
o Ceti	Fabricius	1596
ψ Leonis	Montanari	1667
κ Sagittarii	Halley	1676
χ Cygni	Kirch	1687
δ Cephei	Goodricke	1784
η Aquilae	Piggot	1784
R CrB	Piggot	1795
α Herculis	Wm. Herschel	1796

The most important discovery in this period was connected with observations of variable stars in the Large and Small Magellanic Clouds, most of the work being performed at Harvard by H. Levitt. She realized, around 1905, that these stars were generally of the same type as the variable δ Cephei. The extraordinary discovery was the relation between the period and the apparent brightness of the stars, the so-called period–luminosity relation. While the zero point for this relation was not then known, the relative brightness of these stars, with the same period as δ Cep and other variables in our galaxy, led to a singularly important cosmological breakthrough—the realization of the great distance that must characterize the Magellanic Clouds.

In the meantime, theoretical work on the cause for variability continued. Russell pointed out several paradoxes resulting from the assumption of binarity as the cause for the variations of many stars. Eddington suggested the pulsation model for stellar variability and derived the linearized, adiabatic pulsation equation. Subsequent work by Rosseland and Cowling further clarified the nature of the instabilities. The culmination of this epoch of study was the encyclopedic review article by Ledoux and Walraven, in 1958, which still serves as an important starting point for all discussions of stellar variability. The role played by stellar convection zones in driving pulsation and producing the radial velocity variations in Cepheids was understood by Cox around 1963. The first machine calculations of nonlinear, nonadiabatic pulsation models were accomplished in the mid-1960s. Around the same time, the first large-scale grids of stellar interior models were produced. Schwarzschild and Harm discovered the thermal instability of double shell sources in 1965.

This article deals exclusively with pulsating variable stars, those which are mechanically and thermally unstable because of intrinsic internal structural processes. Several other classes of variable stars are dealt with in the articles "Supernovae" and "Binary Stars" (which includes a discussion of the RS CVn stars, novae, and eclipsing binary stars).

II. OBSERVATIONS OF STELLAR VARIABILITY

Photometric and radial velocity variations are the basic data for the study of stellar variability. Light curves are the most obvious manifestation of variations, but because of the problems associated with interpretation, they cannot

be deemed sufficient to pinpoint a mechanism responsible for the observed changes. Spectroscopic observations, specifically radial velocity variations, when coupled with color and light variability, serve to identify the causative agent behind most stellar variability.

The detailed behavior of velocity and light changes of a Cepheid variable or RR Lyrae star are usually very different from those observed in an eclipsing binary, although both will show both light and radial velocity changes. Most variable stars show asymmetric light curves, usually steepest on the ascending branch. Many are not strictly periodic, some are genuinely aperiodic. There is often a lag between maximum brightness and maximum radial velocity. Most telling is that there are often both color and spectral changes, which are gradual through the photometric cycle, indicating alterations in the physical structure of the stellar atmosphere.

The presence of radial velocity variations means, for a pulsating star, that the radius of the star is physically changing with time. The scale of this change is the time integral of the radial velocity curve. Changes in the sign of the radial velocities take place at extrema of the radius changes, so that they can be thought of as a bounce or recontraction. If the pulsations are strictly periodic, the radial velocity curve will be as well. The relation between the phase of maximum light and maximum radial velocity is most important in specifying the driving mechanism for the pulsation. One application of the radial velocity variations, in conjunction with the light curve, is the determination of temperature by the Baade–Wesselink method. This technique assumes that at two phases for which the colors are the same, the temperatures are also identical. From the comparison of luminosities at these points with the radial velocities, the radii can be obtained and the amplitude of pulsation can be specified. Additional information is available using the Barnes–Evans calibration of surface brightness versus intrinsic red color [$(V - R)_0$ color, in magnitudes]:

$$F_V = 3.956 - 0.363(V - R)_0 \qquad (1)$$

This provides an angular diameter, ϕ (in arc seconds), for the star using the apparent (reddening corrected) visual magnitude V_0:

$$F_V = 4.221 - 0.1V_0 - 0.5 \log \phi. \qquad (2)$$

The radius thus obtained can be compared with that derived from the radial velocities.

For Cepheid variables, there is an empirical, as well as theoretical, period–luminosity–color relation that relates the period of pulsation of the star to its absolute magnitude and surface temperature. In combination with the Barnes–Evans relation, this can help place the star in its evolutionary state, and it also serves as a fundamental calibrator for the extragalactic distance scale (see discussion in Section IV).

III. BASIC THEORY OF STELLAR PULSATION

Stellar pulsation is a phenomenon resulting from departures from strict hydrostatic and thermal equilibrium. Such a state can be reached either by internal, that is to say evolutionary, avenues or by the presence of external perturbers. For instance, the tides observed on Earth are the result of the periodic forcing on the oceans and crust, on a rotating planet, by the gravitational pull of the Sun and Moon. On the other hand, the relaxation of the planet following an earthquake is internally driven.

Pulsation, when it appears on a global scale as ordered standing wave motion, is essentially a resonant phenomenon. Driving is mechanical, due to convection, which couples the temperature gradient to mass transport by buoyancy. In stars like the Sun, this results in a broad spectrum of acoustic modes, trapped in the envelope. Should the overlying mass be large enough as a restoring force and the thermal and mechanical time scales in a layer be of the same order stable pulsation—organized on the scale of the envelope—can result. The conditions are met only in limited parts of the Hertzsprung–Russell (H-R) diagram, the instability strip.

For stars, the problem is that the precise nature of the mechanical and thermal couplings of different parts of the interior is difficult to specify. Stellar interiors are from uniform or isothermal. There are steep temperature gradients near the surface, which is defined by the photosphere, and chemical composition changes because of the processing of the interior mass by nuclear reactions near the core. In addition, the ionization of the gas is a function of depth, as is the opacity, and therefore both the heat capacity and radiative losses are strongly depth dependent.

Pulsation serves as one of the few direct probes of the state of the interior of a star. The way in which an instability, a temperature and pressure variation at some zone in the stellar envelope, manifests itself at the stellar surface from which point it can be studied by an external observer is a complex resonance phenomenon that is not completely understood. Helio- and asteroseismology, both developed on the analogy of terrestrial seismology, use the acoustic spectrum to probe the depth dependence of the sound speed, much like a trumpet bell.

There are several time scales associated with a stellar interior. One is the characteristic time it takes for a zone of some physical size to achieve pressure equilibrium, the sound crossing time. This is given by $t_s = l/a_s$, where a_s is the sound speed that is a function of both the equation of

state, through the ratio of specific heats, and the temperature and is therefore strongly depth dependent. Another is the time scale for radiative losses, which depends strongly on the opacity and thus on the equation of state, density, and temperature. Finally, and most critically, there is the time scale for a parcel of matter to oscillate in a gravitational field, given its local density. This last time scale, the so-called free-fall time scale, is given by the following:

$$t_{ff} = (G\langle\rho\rangle)^{-1/2}, \tag{3}$$

where $\langle\rho\rangle$ is the mean density of the star. For a star with a mean density of 1 g cm^{-3}, this time scale is about 10 min, essentially the collapse time for matter from any distance to the center of mass. Stars have large density gradients and this mean density is heavily weighted toward the core, so in general, to account for the degree of central concentration, one can take the pulsation time scale to be related to the free-fall time through the following:

$$Q = P(\langle\rho\rangle/\langle\rho_\odot\rangle)^{1/2}, \tag{4}$$

where $\langle\rho_\odot\rangle$ is the mean density of the Sun, about 1.4 g cm^{-3}, and P is the pulsation period. The value of Q can be calculated from any interior model once the period has been determined from the pulsation equation.

In the first approximation, a star can be assumed to be spherical. While rotation may play a role in the internal structure, resulting in polar flattening and equatorial expansion to an oblate spheroid, in general this is not important. The equation of hydrostatic equilibrium can be written as a balance between the outward pressure gradient and gravitation:

$$\ddot{r} = -\frac{1}{\rho}\frac{\partial p}{\partial r} - \frac{Gm}{r^2}. \tag{5}$$

Here r is the radius, ρ is the density, p is the pressure, and m is the mass interior to a point r.

Energy loss is assumed to be due to radiation, which takes place diffusively because of the large optical depth of the stellar envelope:

$$L = \frac{16\pi acr^2T^3}{3\kappa\rho}\frac{\partial T}{\partial r}. \tag{6}$$

Here, L is the luminosity, a and c are the Stefan constant and speed of light, respectively, κ is the opacity coefficient (a function of the density and temperature), and T is the temperature.

The energy balance of the stellar interior is given by the change in the entropy of the matter as a result of internal energy generation and radiative losses. The second law of thermodynamics:

$$T\frac{dS}{dt} = \frac{dE}{dt} - \frac{P}{\rho^2}\frac{d\rho}{dt}, \tag{7}$$

when supplemented by the equation of state $p = p(\rho, T)$, becomes

$$T\frac{dS}{dt} = \frac{p}{\rho(\Gamma_3 - 1)}\left[\frac{d}{dt}\ln p - \Gamma_1\frac{d}{dt}\ln\rho\right] = \varepsilon - \frac{\partial L}{\partial m}, \tag{8}$$

where S is the entropy; $\Gamma_1 = (d\ln p/d\ln\rho)_{\mathrm{Ad}}$ and $\Gamma_3 - 1 = (d\ln T/d\ln\rho)_{\mathrm{Ad}}$ are the adiabatic exponents for the pressure and temperature, respectively; and ε is the rate of energy generation per unit mass. This latter quantity, like the opacity, is a function of the density and temperature. The opacity is, in addition, dependent on the composition. Dramatic temperature dependence results from atomic absorption lines.

Finally, the explicit assumption of sphericity of the star enters in the equation for the mass interior to any radius:

$$\partial m/\partial r = 4\pi r^2\rho. \tag{9}$$

It should be noted that all of these equations will be more complicated in the event that the star is distorted from sphericity, because of the coupling in the momentum equation between the various components of the gravitational acceleration on a distorted star and possible effects of rotation in both the momentum and energy equations.

In this enumeration of the equations of stellar structure, there are two time dependent equations, for the acceleration and the energy generation, and two equations of constraint, the diffusion approximation for the energy loss and the mass conservation equation. Thus, we would expect that the final equation will be third-order in time, and that there will be both stable and unstable solutions for many possible combinations of the parameters of energy generation and opacity. In fact, one point is already manifest in these equations—there are three critically important parameters in the pulsation characteristics of a star: the ratio of specific heats, the opacity, and the rate of energy generation. These govern the stability of the star.

Since the radius is changing with time in a pulsating star, the radial coordinate is not the best one to employ for the analysis of the instability. Instead, the Lagrangian approach is used, in which a mass zone is chosen and followed in its oscillations, rather than using the radial coordinate. In this way, the interior model can be calculated in equilibrium and perturbed, and thus the perturbation is only a function of time. For example:

$$r = r_0\left(1 + \frac{\delta r}{r_0}\right) = r_0(1 + \xi), \tag{10}$$

where r_0 is now the equilibrium position of a mass zone $q = M(r)/M$, M being the stellar mass.

The calculation of pulsational properties proceeds in a deductive way. One assumes an interior structure and asks what the response will be for the mass if this configuration is infinitesimally perturbed. The usual argument is that if the mass is intrinsically unstable any perturbation, no matter how small and irrespective of its origin, will be sufficient to begin pulsation. Consequently, it is assumed that all physical parameters, T, p, and ρ, are perturbed from equilibrium values by a small component.

The combined effect of the perturbations of the equations of stellar evolution is that a single equation can be obtained for the temporal variation of the pulsation radial amplitude:

$$\frac{\partial^3 \xi}{\partial t^3} - 4\pi \dot{\xi} r_0 \frac{\partial}{\partial m}[(3\Gamma_1 - 4)P_0] - \frac{1}{r_0^2}\frac{\partial}{\partial m}\left[16\pi^2 \Gamma_1 P_0 \rho_0 r_0^6 \frac{\partial \dot{\xi}}{\partial m}\right] = -4\pi r_0 \frac{\partial}{\partial m}\left[\rho_0(\Gamma_3 - 1)\delta\left(\varepsilon - \frac{\partial L}{\partial m}\right)\right], \quad (11)$$

which is also known as the linearized wave equation (although not really a wave equation). This equation, first derived by Eddington, is one of the best studied tools of stellar interiors. There is an exact solution for the adiabatic version of this equation, which gives secular instability if $\Gamma_1 \leq \frac{4}{3}$ (for a perfect gas, this ratio is $\frac{5}{3}$). The effects of partial ionization, which occurs in convection zones, decrease the star's stability. In consequence, the convection zones, already mechanically unstable, tend to drive the envelope pulsation.

The reason for dwelling for so long on the derivation of the pulsation equation is that there are many classes of instability. The pulsation is driven from different regions of the star depending upon the state of the stellar interior. For instance, as first emphasized by Eddington, a star can be thought of as a self-gravitating analog to a steam engine. If the star is to be pulsationally unstable, or what he called overstable, then the driving region must heat up on compression and overcool on expansion. Thus, the restoring force for the pressure rises on compression, and the rate of energy dissipation is most efficient at the lower temperature, so the matter overcools. The combined effects of the heat capacity and the opacity throttle the rates of heating and cooling, and thus there are specific zones within the star that are identified with the initiation and maintenance of the oscillatory behavior of the envelope. If the radiative time scale is short enough, the pulsations are strongly damped. In contrast, adiabatic pulsation is possible if Γ_1 approaches $\frac{4}{3}$, the limit of stability for a polytropic equation of state.

There are few stars that are driven into instability by the nuclear burning, unless that burning is not central. For example, in the case of stars in the transition stage between main sequence and red giant, hydrogen burning occurs in a thin shell zone surrounding the hot but inert helium core. The temperature dependence of the CNO reaction is not, though, by itself, sufficient to induce the instability observed in stars in this stage of their lives. Instead, once the helium has undergone consumption in the stellar core, the double shell source that surrounds the red giant core is unstable because of the extreme temperature sensitivity of the He and CNO processes combining to produce the observed luminosity. This double shell source instability, first recognized by Schwarzschild and Harm in 1965, is one of the few, and still best studied, examples of nuclear-driven instabilities.

The primary cause of stellar pulsation is that the convection zones of the envelope, which correspond to ionization zones, are thermally unstable. When the zones expand and cool, the ionization drops and so does the opacity. When compressed, they heat up and the ionization goes up thereby increasing the heat capacity (the value of γ drops). Detailed nonlinear modeling of both Cepheids and RR Lyrae stars confirms this expectation. Because of the large energy associated with the ionization of neutral (He I) and ionized helium (He II) and because these zones lie fairly deep in the stellar envelope, they serve as the sources for pulsational driving in most classes of variable stars. The explanation lies in another aspect of the driving of the instability. If the mass overlying the driving zone is too large, the mechanical impulse of the instability in the convection zone will be damped before it can raise the surface layers substantially and the star will be stable or of very low amplitude. Should the mass above the zone be too small, however, the inertia of the zone will be small and it will not be able to drive the compression sufficiently on fall-back to continue the pulsation. In addition, if the convection zone is too near the surface, the layers will be optically thin and the energy of the pulsation will be too easily dissipated by radiative processes. Thus, a delicate balance exists between the depth of the pulsational driving zone and the conditions in the zone, and this serves as the simple explanation as to why most stars do not vary.

Intuition would lead one to expect that, when the star is smallest, the luminosity would be greatest since the decrease in surface area should be more than compensated by the increase in the surface temperature. The observations show, however, that maximum luminosity usually occurs at maximum expansion velocity, the phase lag being about 0.25 of the period. This effect is produced by the presence of the convection zones, mainly the surface hydrogen ionization region, which stores the wave of luminosity

coming from the driving He II zone deep in the interior and slowly releases it on expansion. This nonadiabatic behavior is largely due to the variation in both the opacity and specific heat ratio in this zone as a function of phase in the pulsational cycle.

Nonradial pulsation will result if the star is not spherical, for instance because of rapid rotation or the presence of a close perturbing companion in a binary system. Such pulsations are often nearly adiabatic, involving only the stellar atmosphere and, consequently, of short period and low amplitude. In the case of the sun, for example, there is no organized resonant radial pulsation, but the envelope shows a rich spectrum of modes because of the mechanical instability of the convection zone. Nonradial modes are driven by gravity rather than thermal instabilities and are often more sensitive to the equation of state than to the opacity (the β Cep stars may be an exception to this).

IV. CLASSIFICATION OF INTRINSIC VARIABLE STARS

Classes of variable stars are generally named after the first discovered, or prototypical, member of the class. Subsequent to the recognition of a given phenomenology, it sometimes appears that the namesake is not the best representative of the class, but the name persists. As in all taxonomy, there is a dichotomy between the "splitters" (those who relish the proliferation of subclasses for every departure from the precisely defined properties of a class) and "lumpers," and in this article, the latter presentation is followed. In this article, a broad separation is made between the radial and nonradial pulsators, driven in large measure by the differences in the theoretical models required to understand their behavior.

Variable stars are named according to a standard international rule. The stars are named in order of discovery. The first star discovered in a given constellation is named R (the first letter that had not previously been used in star catalogs during the 19th century), if not already named with some other designation (such as Greek or Bayer names). The single letters are exhausted with Z, and the sequence restarts with RR through ZZ, then goes from AA through QZ before starting as V335 and continuing. The central clearing house for all variable star designations and properties is maintained by the International Astronomical Union in Moscow, which publishes the *General Catalogue of Variable Stars*.

A. Radial Pulsators

Most radial pulsators are confined to a nearly vertical strip on the Hertzsprung–Russell diagram, called the instability strip. In the range of temperature represented by the strip, the stars have thermally and mechanically unstable envelopes.

1. Delta Scuti Variables

The Delta Scuti (δ Sct) stars are main sequence and slightly evolved, intermediate mass stars, A-type stars that show small amplitude (0.003 to 0.9 magnitudes) pulsations in a period range from about 0.01 to 0.2 days. Their luminosities range from 10 to 100 $L_\odot$. They are primarily driven by the surface convection zones, and only a small amount of mass is involved with the pulsational activity of the star. Several classes of stars coexist in the H-R diagram with these stars, the δ Del stars, Am stars, and Ap stars (see Section IV.B.3). The maximum radial expansion velocity closely matches the phase of maximum light to within about 1/10th of the period. These variables are numerous but difficult to detect because of their short periods and small amplitudes; however, because many are found in clusters, their evolutionary status is well understood. They form the lower luminosity (main sequence) end of the Cepheid instability strip.

2. RR Lyrae Stars

The RR Lyrae (RR Lyr) stars are identified with a unique stage in the life of a low-mass star, the horizontal branch, helium core burning period of evolution. As such, they show a small range in luminosity, the most variable property being their effective temperature as a function of mass. The locus of stellar masses on the horizontal branch is determined by helium core mass. The RR Lyr stars range in spectral type from A3 to A6 and have absolute magnitudes of about $0.^m5$, about 10^2 $L_\odot$. They are low-mass stars, 0.5 to 0.7 $M_\odot$. The pulsation is driven by instabilities in the He II convection zone. The RR Lyrae stars display several different light curves, which are correlated with their periods. The RR(b) stars show both fundamental and first overtone pulsations, the typical period ratio being about $P_1/P_0 = 0.75$; the prototype of this class is AQ Leo. For those RR Lyr stars with only a one observed period, Bailey originally divided the "cluster variables" into two main subclasses. The RRab stars have periods ranging from 0.3 to 1.2 days and amplitudes typically from 0.5 to 2 magnitudes. They are characterized by "sawtooth" (steep ascending branches) light curves. The prototype of the class, RR Lyr, is a member of this subclass. The RRc stars show small amplitudes, generally less than 0.8 magnitudes, short periods from 0.2 to 0.5 days, and essentially sinusoidal light curves. As in the δ Sct and δ Cep stars, the maximum expansion velocity corresponds to maximum light.

The RR Lyr stars are best studied as members of globular clusters, where they appear in large numbers because of the sizable population of the horizontal branch in these evolved stellar systems. They can be classified according to a metallicity index, first proposed by Preston, which compares the spectral type assigned from the metal lines, especially Ca II, to that obtained from the hydrogen Balmer lines. This ΔS index is a useful indication of metal abundance and shows that the RR Lyr stars are low-mass Population II stars, related in their pulsational properties to the W Vir and RV Tau stars, but of very low mass.

3. Delta Cepheid Variables: Classical Cepheids

These are the prototype variable stars and the first class for which the period luminosity relation was discovered. They are generally moderately massive stars, being above 2 or 3 $M_\odot$, evolved, and in the stage of helium core burning. They reside in the instability strip, in fact defining the position of the strip on the H-R diagram, and have single modes of pulsation with periods between about 1 day and 2 months.

Most δ Cep stars have periods ranging from 1 to about 140 days, with amplitudes ranging from 0.01 to 2 magnitudes. In general, their variations are greater at shorter wavelength and longer period. The stars are usually spectral type F at maximum, and G or K at minimum. Their luminosities range from $5 \times 10^2\ L_\odot$ to about $2 \times 10^4\ L_\odot$, depending on their mass and evolutionary status. The classical Cepheids are Population I stars, occasionally residing in open galactic clusters from which their approximate masses of 3 to 10 $M_\odot$ are obtained. Several have been found to be members of binary systems, notably SU Cas. Their radial velocity variations are characteristic: with only a small phase shift (about 0.1 in phase), the maximum expansion velocity corresponds to maximum light. δ Cep light curves are usually distinctive, being steep on the ascending branch with little structure to the curve.

A subset of the δ Cep stars, called Cep(B) or bump Cepheids, show multiple periodicity, usually with $P_1/P_0 = 0.7$ and P_0 between 2 and 7 days. TU Cas and V367 Sct are typical of this subclass.

Finally, the DCepS subclass shows amplitudes that are always less than 0.5 magnitudes and almost symmetric light curves. Their periods rarely exceed 7 days. The well-studied star SU Cas is typical of this subclass. These stars seem to be pulsating in the first harmonic (hence, the smaller amplitudes and shorter periods), while classical Cepheids appear to be pulsating in the fundamental mode.

The Cepheids arise from a high-mass population, and as such they do not represent a unique stage in the life of a star. Following hydrogen core exhaustion, stars greater than about 3 $M_\odot$ will cross the instability strip at least once. The higher mass stars, greater than about 7 $M_\odot$, will traverse the strip several times, including the post-helium-core burning stage, each time with increasing luminosity; as many as 5 crossings have been calculated for stars above this mass.

Theroetical models provide the following calibration for the Cepheid period–luminosity relation:

$$P_F \sim \left(\frac{L}{L_\odot}\right)^{0.83} \left(\frac{M}{M_\odot}\right)^{-0.66} T_{\text{eff}}^{-3.45}, \qquad (12)$$

where P_F is the fundamental period and T_{eff} is the effective temperature. Turning this relation around, one can derive a mass for the Cepheid from theoretical models, called the pulsational mass, if the luminosity, surface temperature, and period are known. Few masses are well known, but the pulsational mass determinations are quite sensitive to the temperature determinations for these stars, while the evolutionary masses seem to agree with the few stars for which masses have been obtained. For example, for SU Cyg, the Cepheid mass is about 6 $M_\odot$, which agrees with the evolutionary mass.

Feast and Catchpole (see Heck and Caputo, 1999) have recently rederived the period-luminosity relationship from *Hipparcos* parallax data:

$$\langle M_V \rangle = -2.81 \log P - (1.43 + 0.1) \qquad (13)$$

where $\langle M_V \rangle$ is the mean absolute magnitude of the Cepheid averaged over its pulsational cycle and P is the period in days. The slope derives from the LMC, the zero point comes from parallax measurements toward about two dozen Galactic Cepheids.

The δ Cep stars are also important as distance indicators, in large measure because of their place in the instability strip; as a result, an enormous effort has, in the past 40 years, been put into the determination of intrinsic properties of these stars, especially the study of their masses and luminosities. The discovery of large numbers of binary stars among the Cepheid variables has been an important clue to these properties. Many of the companions are still on the main sequence and are of suficiently early hot spectral type that they must be moderately massive. They are best studied using ultraviolet spectra, in which the spectrum of the Cepheid variable does not appear. Masses can be determined from a comparison of the Cepheid radial velocity curve, obtained from the optical spectrum, with that for the companion, obtained from the UV. In consequence, the lower mass of the Cepheid can be understood, and the time scale for the evolution of the variable specified within limits. The binaries will also serve as useful calibration of the evolutionary and pulsational mass determinations, once a large enough number of them have been identified.

An additional determination of intrinsic properties is provided by the use of the infrared. For stars of surface temperatures less than about 8000 K, the peak of the spectral distribution falls in the red. Therefore, the use of the IR, wavelengths of 1 to 3 μm, can give a clear indication of the bolometric variability of the star and is not sensitive to changes in the shape of the spectrum. Further, since the infrared is far less sensitive to the effects of interstellar reddening than the optical, the zero point for the magnitudes can be more reliably determined. The slope of the relationship depends on the wavelength at which the variations are observed, but the mean stellar magnitude as a function of log P appears to be invariant between galaxies, independent of stellar metallicity, for classical Cepheids.

4. W Virginis Stars: Population II Cepheids

These are the Population II analogs of the classical Cepheids. They arise from a lower mass population, from about 0.5 to 0.8 $M_\odot$, but being similarly internally structured, they display the same mechanical and photometric properties as the more massive Population I stars. They range in period from 0.5 to 35 days and have amplitudes in the range 0.3 to 1 magnitude. These stars also display emission in hydrogen Balmer lines. There are two subclasses identified among the W Vir stars. The CWA stars, like W Vir, have periods greater than 8 days; CWB stars, like BL Her, have shorter periods. In BL Her, which has also served as the prototype of a subclass of W Vir stars, bumps are observed after maximum on the descending portion of the light curves.

The W Vir stars are fainter than classical Cepheids at the same period by between 0.7 and 2 magnitudes, a fact that has important cosmological consequences. Since the extragalactic distance scale rests in large measure on the Cepheid variables as distance calibrators for relatively nearby galaxies, the period–luminosity relation is used to determine the distance modulus, $m_j - M_j$. Here, m_j and M_j are the apparent and absolute magnitude in the jth wavelength band. With only the period and apparent magnitude in hand, one can determine the distance to a galaxy knowing, from the P-L (or period–luminosity–color or P-L-C) relation, the intrinsic luminosity of the star, hence its distance.

5. Long-Period Variables

a. Mira variables. Mira variables, named after *o* Cet, are emission-line cool giants, generally arising from a population of low to intermediate mass (1 to 2 $M_\odot$) stars. Miras typically have luminosities of about 10^3 $L_\odot$ and temperatures less than 4000 K. They are found among the Me, Ce, and Se stars, often displaying time variable emission lines of both atomic and molecular species. These stars have some of the largest amplitudes observed among pulsating variables, ranging from 2.5 to 11 magnitudes in the optical, but generally considerably smaller in the infrared (at a wavelength of a few micrometers, the amplitude is less than one magnitude). The periods range from 80 to over 1000 days, although many of the longest period systems are quite irregular in their light curve behavior. A characteristic of the light variations is the change in the light curve structure with time. Miras in general do not appear to be rigorously periodic, with amplitudes that do not repeat from one cycle to another at any wavelength and considerable evidence for mass loss and shock processes accompanying the pulsation cycle.

The Miras are perhaps the best studied examples of nonlinear pulsators, having amplitudes large enough to produce the ejection of matter and strongly nonadiabatic behavior of the envelopes. One interesting aspect of these stars is that their pulsation periods are sufficiently long, and their amplitudes large enough, that time-dependent convection may be important in modeling their internal structure. Considerable theoretical attention is currently being paid to these stars.

b. RV Tauri stars. The RV Tau stars are intermediate temperature (F and G) supergiants with luminosities of approximately 10^4 $L_\odot$, highly evolved, and consequently of long period. They are often associated with old Population I or Population II. Their periods range from 30 to 1000 days, and their amplitudes are about 3 to 4 magnitudes.

The RV Tau stars are distinctive because of their light curves, which on the short-period end resemble classical Cepheids, but which for the long-period stars are quite distinctive. The longer period subclass, called RVb stars, shows variable mean magnitudes, with a variation in the amplitude of maximum between successive cycles in a periodic way. The RV Tau star is the prototype of this class. The RVa stars, typified by AC Her, show no variation in the mean magnitude.

c. R Corona Borealis stars. The R CrB stars are hydrogen deficient, helium rich, and often carbon rich, high-luminosity stars, characterized by a semiregular behavior, which is unique among variable stars. They undergo quasiperiodic fadings during which time they can decrease in brightness by more than 1 to about 10 magnitudes on a time scale of 30 to several hundred days. The intervals between these events are irregular; some low amplitude periodicity of less than a year may be present as well. These episodes are marked by the ejection of considerable

amounts of dust, as determined from ultraviolet satellite observations of their spectra. Optically, they display high polarization. The mechanism responsible for this behavior is not understood presently. The R CrB stars, which are poorly understood, also span much of the temperature range of the stellar population, occupying spectral types Be to the carbon stars. These stars show very high luminosities, about $5 \times 10^4\ L_\odot$.

d. Semiregular variables. Finally, there are high-luminosity stars that show photometric and radial velocity changes on time scales of months to years that do not appear to be periodic, although there is a characteristic interval of time during which the brightness of the star is observed to change. The supergiants are among the best examples of this class of variable, perhaps the best known being the M supergiant Betelguese (α Orionis). In many ways, these stars resemble the RV Tau and Mira variables, but without the regularity of variability associated with these other stars.

The SRc, or supergiant semiregular variables, have amplitudes of up to 1 magnitude and periods from tens to thousands of days. Their luminosities range from 10^4 to $10^5\ L_\odot$, and temperatures are generally in the range of 2000 to 4000 K. The SRc's overlap with the R CrB stars. The OH/IR stars, red supergiants that display OH maser emission and large middle and far infrared luminosities, may be related to these stars. Other subclasses of the semiregular variables are the SRa and SRb stars, giants with similar periods but lower luminosities, and PV Tel stars, which are helium-rich, long-period supergiants of low amplitude (several tenths of a magnitude). The SRa and SRb stars overlap with the Mira variables.

6. Luminous Blue Variables

The luminous blue variables (LBVs) are a recently recognized class of massive supergiants. Also called S Dor variables, after their Large Magellanic Cloud prototype, these are the most luminous stars in a galaxy and easily identified in extragalactic systems. Their luminosities are often about $10^6\ L_\odot$, and inferred masses are in excess of $30\ M_\odot$ Their amplitudes range from less than 1 magnitude to over 5. Also called Hubble–Sandage variables, they display long (years) time scales for temperature and luminosity changes, often ejecting shells that form pseudophotospheres, causing their spectral types to change from O to as late as F. The galactic supergiant P Cyg, well known for its high rate of mass loss and irregular light variations, is a galactic member of this class. Others recently recognized as LBVs are AG Car and HR Car in the Galaxy, HD 269858f in the LMC, R40 in the SMC and AE And and AF And in M 31.

B. Nonradial Pulsators

1. β Cephei Stars

The name of this class is dependent on which side of the Atlantic one is on. The Europeans typically refer to these as β Canes Majoris stars, while the North Americans typically use β Cep as the prototype.

These stars are the prototypes of nonradial pulsation. They do not show simple mode structure, but often pulsate in several different modes at once. Their periods are extremely short, about 3 to 6 hr. Several Be stars (emission line B stars that are generally rapid rotators) are also in this class, the most notable being λ Eri. The best studied subclass, the short-period group, lies around spectral type B2-3 IV–V, displays periods from 0.02 to 0.04 days, and has amplitudes between 0.015 and 0.025 magnitudes. The short period and the high-mode number are indicative of atmospheric or shallow envelope pulsation, but the details of the driving mechanism for these stars are not currently understood.

The β Cep stars show an instability strip unrelated to that for the classical Cepheids. It is roughly parallel to the main sequence, and at luminosity class IV, characteristic of post-hydrogen-core exhaustion stars. The pulsation is opacity-driven at temperatures of about 10^5 K, a result that follows from the revision of stellar opacity that was required to explain the 5 Cep stars.

Apparently related to these stars is a class of massive pulsators, which generally lies between O8 and B6, covering the range in luminosity from main sequence to supergiant. These nonradial pulsators show periods between 0.1 and about 1 day and small amplitudes, usually less than 0.3 magnitudes. Maximum brightness corresponds to minimum radius, indicative of adiabatic pulsation but distinctly different from the normal behavior observed for the radial pulsators. Spectral lines in these stars show the effects of distortion of the photosphere, varying in shape during the pulsation period in a fashion that indicates distortion of the stellar surface.

2. White Dwarf Stars

Several classes of variables are found among degenerate stars. The hottest are called GW Vir stars, or PG 1159-035 objects, after their prototype. These stars are extremely hot helium white dwarfs, typically with temperatures of about 10^5 K. They show small amplitude pulsation, which is usually observed in many nonradial modes. Because of the dependence of the period on the temperature of the atmosphere, the star's period can change during the course of centuries as the dwarf cools. The pulsation periods are about 500 sec, indicative of nonradial modes; radial pulsation periods for degenerate stars are several orders of

magnitude smaller than this. Several of the central stars of planetary nebulae have been found to be members of this class, with long periods up to about 2000 sec, indicating extensive atmospheres and temperatures that may be as high as 150,000 K.

The cooler class, which lies in the instability strip and was one of the few predicted classes of variable stars, is the ZZ Ceti stars. These are DA-type white dwarf stars, with temperatures between 10^4 K and 2×10^4 K. Here, the driving is due mainly to the hydrogen convection zone. Periods are about 500 to 1000 sec, and masses are about 0.6 $M_\odot$. The DBVs, helium-rich analogs of the ZZ Cet stars, have temperatures between 2×10^4 K and 3×10^4 K but otherwise similar properties. In both cases, the amplitudes are between a few tens and hundreds of millimagnitudes. Among the cataclysmic systems, there are many white dwarf systems that show multiple periodicities, but it is not clear whether these should be ascribed to pulsation.

3. The Rapid Magnetic Pulsators

These are main-sequence A type stars with strong magnetic fields that inhabit the instability strip at the same place as the δ Sct stars. The difference between these and the normal variables of the main sequence is that they display time variable signatures of pulsation, which are due to the site of surface displacement. They provide a unique chance to study the interaction between pulsation and stellar magnetism.

The periods range from about 5 to 25 min, indicative of atmospheric pulsation, and they show very small amplitudes (typically less than $0.^m01$. The pulsation appears to be active only in the vicinity of the magnetic poles. Thus, as the star rotates, the polar region crosses the line of sight and the pulsation is observable. When the magnetic equator crosses the observer's sight line, the pulsation disappears. The mode of the pulsation is very high, typically $l = 20$–40. The presumed driving is essentially the same as for the δ Sct stars, that is, the hydrogen convection zone is unstable. According to Shibahashi and Cox, the modification introduced by the magnetic field is to suppress pulsation at the magnetic equator by the addition of a strong restoring force, while producing an overstability at the poles. About a dozen such systems are known, all confined to stars with magnetic fields of several hundred to several thousand gauss.

V. PRE-MAIN-SEQUENCE VARIABLES

The variability of stars in the stage before the onset of hydrogen core burning is not presently well understood. Intrinsic variations are observed in both the T Tauri stars, which are the young, late-type stellar population associated with, but not imbedded in, their parent molecular clouds, and the FU Orionis stars. The latter are notable for large amplitude outbursts, sometimes of more than 5 magnitudes, taking place over a period of months to years. These have been recently explained as arising from an accretion disk instability in the environment of the forming star. Neither of these classes of stars appears to be pulsationally unstable, although detailed modeling is difficult. All stars in evolutionary stages are characterized by emission line spectra and strong stellar winds. One subclass, the YY Ori stars, shows reverse stellar wind profiles, with absorption on the red side of the emission line suggestive of mass accretion.

The Herbig Ae/Be stars, more massive pre-main sequence counterparts of the τ Tau stars, have also been predicted to display an instability strip for 1 to 4 $M_\odot$ (Marconi and Palla, 1998). The stars for which such variations have been detected, HR 5999, V351 Ori, and HD 35929, have similar periods to δ Sct stars.

VI. CONCLUDING REMARKS

As a tool for the study of stellar interiors, pulsation theory and observations of stellar variability are only now maturing. Increased instrumental sensitivity is allowing the study of very small scale, nonperiodic oscillations in solar-type stars, the field of helio- and asteroseismology; and it is now possible to compute physically interesting nonlinear, nonadiabatic pulsation models for a wide variety of stellar interior conditions. The effects of magnetic fields and rotation are being included in theoretical computations. Improvements in detectors, especially the widespread use of CCDs, are quickening the pace of discovery and study of faint variables. Improved model atmospheres allow for the study of stellar abundances with increased precision. The advent of true flux calibration spectra, again using CCD detectors, allows for direct comparison with models and much improved radial velocity determinations at many different wavelengths.

The field of variable star research is also a fruitful one for both amateurs and professionals with access to small telescopes; it is a field in which every newly determined light curve makes a substantive contribution to our knowledge of the cause and behavior of pulsating stars.

SEE ALSO THE FOLLOWING ARTICLES

BINARY STARS • NEUTRON STARS • PULSARS • STAR CLUSTERS • STARS, MASSIVE • STELLAR SPECTROSCOPY • STELLAR STRUCTURE AND EVOLUTION

BIBLIOGRAPHY

Brown, T. M., and Gilliland, R. C. (1994). "Asteroseismology," *Ann. Rev. Astr. Ap.* **32,** 37.

Cox, J. P. (1980). "Theory of Stellar Pulsation," Princeton Univ. Press, Princeton, NJ.

Cox, A. N., Sparks, W., and Starrfield, S. G., eds. (1987). "Stellar Pulsation," Springer-Verlag, Berlin/New York.

Feast, M. (1999). "Cepheids as distance indicators," *PASP*, **III,** 775.

Gautschy, A., and Saio, H. (1996). "Stellar pulsation across the HR diagram: Part 2," *Annu. Rev. Astron. Astrophys*. **34,** 551.

Hansen, C. J. (1978). Secular stability, *Annu. Rev. Astron. Astrophys.* **16,** 15.

Heck, A., and Caputo, F., eds. (1999), "Post-Hipparcos Cosmic Candles," Kluwer, Dordrecht.

Hoffmeister, C., Richter, G., and Wenzel, W. (1985). "Variable Stars," Springer-Verlag, Berlin/New York.

Ibanoglu, Ca, ed. (2000). "Variable Stars as Essential Astrophysical Tools (NATO Science Series. Series C, Vol. 544)," Kluwer, Dordrecht.

Kholopov, P. N., ed. (1985). "General Catalogue of Variable Stars: 4th Edition" (3 Vols.), NAUKA, Moscow.

Kurtz, D. W. (1990). "Rapidly pulsating Ap stars," *Annu. Rev. Astr. Astrophys.* **28,** 607.

Ledoux, P., and Walraven, T. (1958). *In* "Handbuch der Physik" (E. Flügge, ed.), Vol. 51, p. 353, Springer-Verlag, Berlin/New York.

Marconi, M., and Palla, F. (1998). A_p *J (Letters).* **507,** L141.

Rosseland, S. (1949). "The Pulsation Theory of Variable Stars," Oxford Univ. Press, London/New York.

Unno, W., Osaki, Y., and Shibahashi, H. (1979). "Nonradial Oscillations of Stars," Tokyo Univ. Press, Tokyo.

Stellar Structure and Evolution

Peter Bodenheimer
Lick Observatory, University of California, Santa Cruz

GLOSSARY

Brown dwarf Object in the substellar mass range (0.01–0.075 solar masses) which, during its entire evolution, never attains interior temperatures high enough to account for 100% of its luminosity by nuclear conversion of hydrogen (^{1}H) to helium.

Degenerate gas Gas in which the elementary particles of a given type fill most of their available momentum quantum states as determined by the Pauli exclusion principle. Electron degeneracy occurs in the cores of highly evolved stars and in white dwarfs; neutron degeneracy occurs in neutron stars.

Effective temperature Surface temperature of a star calculated from its luminosity and radius under the assumption that it radiates as a black body.

Galactic cluster Gravitationally bound group of a few hundred or a few thousand stars found in the disk of the galaxy. In a given cluster all stars are assumed to have been formed at about the same time, but a wide range of ages is represented among the various clusters.

Globular cluster Compact, gravitationally bound group of 10^5–10^6 stars, generally found in the halo of a galaxy and formed early in the history of the galaxy.

Helioseismology Study of the internal structure of the sun through observation and analysis of the small oscillations at its surface.

Hertzsprung-Russell diagram A plot whose ordinate indicates the luminosity of a star and whose abscissa indicates the effective temperature of the star. A given star at a given time is represented by a point in this diagram. The evolution of a star is represented by a curve (or *track*) in this diagram.

Horizontal branch Sequence of stars on the Hertzsprung-Russell diagram of a globular cluster, above the main sequence. The stars all have approximately the same luminosity and are in the evolutionary phase where helium is burning in the core.

Luminosity Total rate of radiation of electromagnetic energy from a star, in all wavelengths and in all directions. Unit: energy per unit time.

Main sequence Sequence, or band, of stars in the Hertzsprung-Russell diagram, on which a large fraction of stars fall, running diagonally from high luminosity and high effective temperature to low luminosity and low effective temperature and associated with the evolutionary phase in which the stars burn hydrogen to helium in their cores.

Neutrino Subatomic neutral particle produced in stars chiefly in beta-decay reactions, but also by other processes, which travels at the speed of light and interacts only very weakly with matter.

Neutron star Highly compressed remnant of the evolution of a star of high mass. Its main constituent is free neutrons, the degenerate pressure of which supports the star against gravitational collapse. The mean density is comparable to that of nuclear matter, 10^{14} g cm^{-3}.

Nova Sudden, temporary, but recurrent brightening of a star by factors ranging from 10 to 10^6, occurring in a binary system where, in most cases, a white dwarf is accreting mass from a main-sequence companion. The outbursts are caused either by instability in the accretion disk surrounding the white dwarf (*dwarf nova*) or by nuclear reactions in the material recently accreted onto its surface (*classical nova*).

Nucleosynthesis Production of the elements through nuclear reactions in stars or in the early universe.

Protostar Object in transition between interstellar and stellar densities, during which it undergoes hydrodynamic collapse and is observable primarily in the infrared part of the spectrum.

Red giant Post-main-sequence star whose radius is much larger than its main-sequence value and whose effective temperature generally falls in the range 3000–5000 K.

Supernova Sudden increase in luminosity of a star, by a factor of up to 10^{10}, followed by a slower decline over a time of months to years. The event is caused by explosion of the star and dispersal of much of its matter.

White dwarf Compact star (mean density, 10^6 g cm^{-3}) representing the final stage of evolution of a star of low to moderate mass. It is supported against its gravity by the pressure of degenerate electrons.

Zero-age main sequence Line in the Hertzsprung-Russell diagram corresponding to the points where stars of different masses first arrive on the main sequence and representing the first moment, for a given star, when 100% of its luminosity is provided by the fusion of protons to helium.

STELLAR STRUCTURE is the study of the internal properties of a star, such as its temperature, density, and rate of energy production, and their variation from center to surface. The structure can be determined through a combination of theoretical calculations, based on known physical laws, and observational data. Stellar evolution refers to the change in these physical properties with time, again as determined from both theory and observation. The three main phases of stellar evolution are (1) the pre-main-sequence phase, during which gravitational contraction provides most of the star's energy; (2) the main-sequence phase, in which nuclear fusion of hydrogen to helium in the central region provides the energy; and (3) the post-main-sequence phase, in which hydrogen burning away from the center, as well as the burning of helium, carbon, or heavier elements, may provide the energy. The evolutionary properties are strongly dependent on the initial mass of the star and, to some extent, on its initial chemical composition. The end point of the evolution of a star can be a white dwarf, a neutron star, or a black hole. In accordance with this picture, a star can be defined as a gaseous object that, at some point in its evolution, obtains 100% of its energy from the fusion of protons (^{1}H) to helium nuclei. The boundary in mass between stars and substellar objects (also known as brown dwarfs) falls at about 0.075 solar masses ($M_\odot$), below which the 100% energy condition is never fulfilled.

I. INTRODUCTION

The study of the structure and evolution of the stars, which constitute the major fraction of the directly observable mass in the universe, is of critical importance for the understanding of the production of the chemical elements heavier than helium, of the evolution of the solar system and other planetary systems, of the structure and energetics of the interstellar medium, and of the evolution of galaxies as a whole. The structure of a star is determined by the interaction of a number of basic physical processes, including nuclear fusion; the theory of energy transport by radiation, convection, and conduction; atomic physics involving especially the interaction of radiation with matter; and the equation of state and thermodynamics of a gas. These principles, combined with basic equilibrium relations and assumed mass and chemical composition, allow the construction of mathematical models of stars that give, as a function of distance from the center, the temperature, density, pressure, and rate of change of chemical composition by nuclear reactions. A star is not static, however; it must evolve in time, driven by the loss of energy from its surface, primarily in the form of radiation. This energy is provided by two fundamental sources—nuclear energy and

gravitational energy—and in the process of providing this energy the star undergoes major changes in its structure. To follow this evolution mathematically requires the solution of a complicated set of equations, a solution that requires the use of high-speed computers to obtain sufficient detail. The goal of the calculations is to obtain a complete evolutionary history of a star, as a function of its initial mass and chemical composition, from its birth in an interstellar cloud to its final state as a compact remnant or possibly as an object completely disrupted by a supernova explosion.

The heart of the study of stellar structure and evolution is, however, the comparison of models and evolutionary tracks with the observations. There are numerous ways in which such comparisons can be made, for example, by use of the Hertzsprung-Russell (H–R) diagrams of star clusters, the mass–luminosity relation on the main sequence, the abundances of the elements at different phases of evolution, and the mass–radius relation for white dwarfs. There are many exotic stars that the theory is not yet able to fully explain, such as pulsars, novae, X-ray binaries, some kinds of supernovae, or stars showing rapid mass loss. These systems provide a challenge for the future theorist. However, the general outline of the phases of stellar evolution has by now fallen into place through a complex interplay between theoretical studies, observations of stars, and laboratory experiments, particularly those required to determine nuclear reaction rates. The period of development of ideas concerning the structure of stars extends more than 100 years into the past. Among the noteworthy historical developments were the clarification by Sir Arthur Eddington (1926) of the physics of radiation transfer; the development by S. Chandrasekhar (1931) of the theory of white dwarf stars and the derivation of their limiting mass; and the work of H. Bethe (1939) and others, which established the precise mechanisms by which the fusion of hydrogen to helium provides most of the energy of the stars. However, much of the detailed development of the subject has occurred since 1955, spurred by the availability of high-speed computers and by the extension of the observational database from the optical region of the spectrum into the radio, infrared, ultraviolet, X-ray, and gamma-ray regions. Numerous scientists have collaborated to advance our knowledge of the physics of stars in all phases of their evolution.

II. OBSERVATIONAL INFORMATION

The critical pieces of observational data include the luminosity, effective temperature, mass, radius, and chemical composition of a star. A further fundamental piece of information that is required to obtain much of this data is the distance to the star, a quantity that in general is difficult to measure because even the nearest star is 2.6×10^5 astronomical units (AU) away, where the AU, the mean distance of the earth from the sun, is 1.5×10^{13} cm. The AU, measured accurately by use of radar reflection off the surface of Venus, is used as the baseline for trigonometric determinations of stellar distances. The apparent shift (*parallax*) in the position of a star, against the background defined by more distant stars or galaxies, when viewed from different points in the earth's orbit, allows the distance to be determined. The parsec (pc) is defined as the distance of a star with an apparent positional shift of 1 sec of arc on a baseline of 1 AU and has the value of 3.08×10^{18} cm. The nearest stars (the Alpha Centauri triple system) are about 1.3 pc away. The most accurate available database of parallaxes was obtained from the Hipparcos satellite, which measured 120,000 stars with an accuracy down to 0.001 arcsec. Thus, the distance out to which accurate distance measurements are available is about 150 pc, still small compared with the distance to the center of our galaxy ($\approx$8000 pc). Thus, for most stars indirect determinations of distances must be used, based, for example, on period–luminosity relations for variable stars or on properties of the spectrum and apparent luminosity of a star compared to those of a star with a very similar spectrum and known distance.

A. Luminosity

The standard unit of luminosity is that of the sun, which is obtained by a direct measurement of the amount of energy $S_\odot$, received per unit area per unit time, over all wavelengths, outside the earth's atmosphere, at the mean distance of the earth from the sun. This quantity, known as the solar constant, is then converted into the solar luminosity by using the formula

$$L_\odot = 4\pi d_\odot^2 S_\odot = 3.86 \times 10^{33}\ \text{erg/sec},$$

where $d_\odot = 1$ AU. For other stars, in principle, the stellar flux S (energy per unit area per unit time) received at the earth is corrected for absorption in the earth's atmosphere and in interstellar space and is extended to include all wavelengths of radiation. If the star's distance d is known, its luminosity follows from $L = 4\pi d^2 S$.

B. Effective Temperature

A number of different methods are used to determine the effective (surface) temperature, most of which are based on the assumption that the stars radiate into space with a spectral energy distribution that approximates a black body. For a few stars, such as the sun, whose radius R can be measured directly, the value of T_{eff} is obtained from L and R by use of the black-body relation $L = 4\pi R^2 \sigma T_{\text{eff}}^4$,

where σ is the Stefan-Boltzmann constant. Otherwise the temperature may be estimated by four different methods.

1. The detailed spectral energy distribution is measured, and the temperature of the black-body distribution that best fits it is found.
2. The wavelength λ_{max} of maximum intensity in the spectrum is measured, and the temperature is found from Wien's displacement law for a black body: $\lambda_{max}T = 2.89 \times 10^7$, if λ_{max} is expressed in angstroms.
3. The "color" of the star is obtained by measurement of the stellar flux in two different wavelength bands. For example, the color "B − V" is obtained by comparing the star's flux in a wavelength band centered at 4400 Å and about 1000 Å broad (blue) with that in a band of similar width centered at 5500 Å (visual). In principle, the transmission properties of the B and V filters could be used in connection with black-body curves to determine the temperature. In practice, stars are not perfect black bodies, and the temperature is determined by comparison with a set of standard stars.
4. From the strength in the stellar spectrum of absorption lines of various chemical elements, the spectral type and thereby the temperature can be determined by comparison with a set of standard stars. This method does not depend on the black-body assumption, and the temperature is, in principles, determined from the degree of excitation and ionization of the atoms.

C. Radius

The radius can be measured directly for only a few stars. In the case of the sun, the angular size can be measured and the distance is known. In the case of Sirius and a few other stars, the angular diameter can be measured by use of interferometry or by high-resolution direct imaging with the Hubble Space Telescope. For certain eclipsing binary systems, if the orbital parameters are known and the light variation with time can be accurately measured, the radii of both components can be determined. In all other cases the radius must be estimated from measurements of L and T_{eff} and the formula $L = 4\pi R^2 \sigma T_{\text{eff}}^4$.

D. Mass

A stellar mass can be measured directly only if the star is a member of a binary system, in which case Kepler's third law can be applied. If M_1 and M_2 are the stellar masses, in units of the solar mass $M_\odot$, P is the orbital period of the system in years, and a is the semimajor axis of the relative orbit in astronomical units, then the law states that $(M_1 + M_2)P^2 = a^3$. In the case of the sun, the orbital periods and distances of the planets can be used for an accurate determination of solar masses. If both components of a binary are visible (visual binary) and if the angular separation as well as P can be determined, then the sum of the masses follows from Kepler's law as long as the distance is known. The individual masses can be found if the relative distances of the stars from the center of mass can be measured. If the binary system has such a close separation that the components cannot be visually resolved, it may be possible to resolve them spectroscopically. If the Doppler shifts of spectral lines as a function of time can be measured for both components, then the period can be determined as well as the mass ratio and minimum masses of the components. If in addition the system shows an eclipse, then the individual masses can be determined. There are relatively few systems with the required orbital characteristics to allow reasonably accurate mass determinations.

E. Abundances

The relative abundances of the elements in the solar system (at approximately the time of its formation) have been compiled by E. Anders and N. Grevesse, based in most cases on laboratory measurements of the oldest meteoritic material and in some cases from the strengths of absorption lines in the solar atmosphere. Abundances for selected elements are given in Table I. In stellar evolution theory, the fractional abundance of H by mass is known as X, that of He is known as Y, and that of all other elements are known as Z. In stars, the abundances are determined

TABLE I Solar System Abundances

Element	Log abundance by number of atoms (H = 10^{12})	Fractional abundance by mass
1 H	12.00	0.704
2 He	11.00	0.279
3 Li	3.31	9.89×10^{-9}
6 C	8.55	2.97×10^{-3}
7 N	7.97	9.12×10^{-4}
8 O	8.87	8.28×10^{-3}
10 Ne	8.07	1.66×10^{-3}
11 Na	6.33	3.43×10^{-5}
12 Mg	7.58	6.45×10^{-4}
13 Al	6.47	5.56×10^{-5}
14 Si	7.55	6.96×10^{-4}
16 S	7.21	3.63×10^{-4}
20 Ca	6.36	6.41×10^{-5}
24 Cr	5.67	1.70×10^{-5}
26 Fe	7.51	1.26×10^{-3}
28 Ni	6.25	7.29×10^{-5}

from the strengths of absorption lines in the spectrum combined with other parameters in the stellar atmosphere and compared with solar values. In practically all measured systems, the values of X and Y are deduced to be very similar to the solar values. However, Z can vary, and in the oldest stars and globular clusters, it can fall below the solar value by a factor of more than 100.

F. Hertzsprung-Russell Diagram

The Hertzsprung-Russell (H–R) diagram is a plot of luminosity versus surface temperature for a set of stars. Although the data can be plotted in various forms, the sample H–R diagram shown here (Fig. 1) gives data converted from observed quantities to L and T_{eff}. Most of the stars lie along the main sequence, which represents the locus of stars during the phase of hydrogen burning in their cores, with increasing T_{eff} corresponding to increasing mass. The stars well below the main sequence are in the white dwarf phase, having exhausted their nuclear fuel. The stars in the upper-right part of the diagram are red giants (e.g., Aldebaran); these are stars that have exhausted their central hydrogen and are now burning hydrogen in a shell region around the exhausted core. Some of them are also burning helium in a core or in a shell. The number of stars in a given region of the H–R diagram is roughly proportional

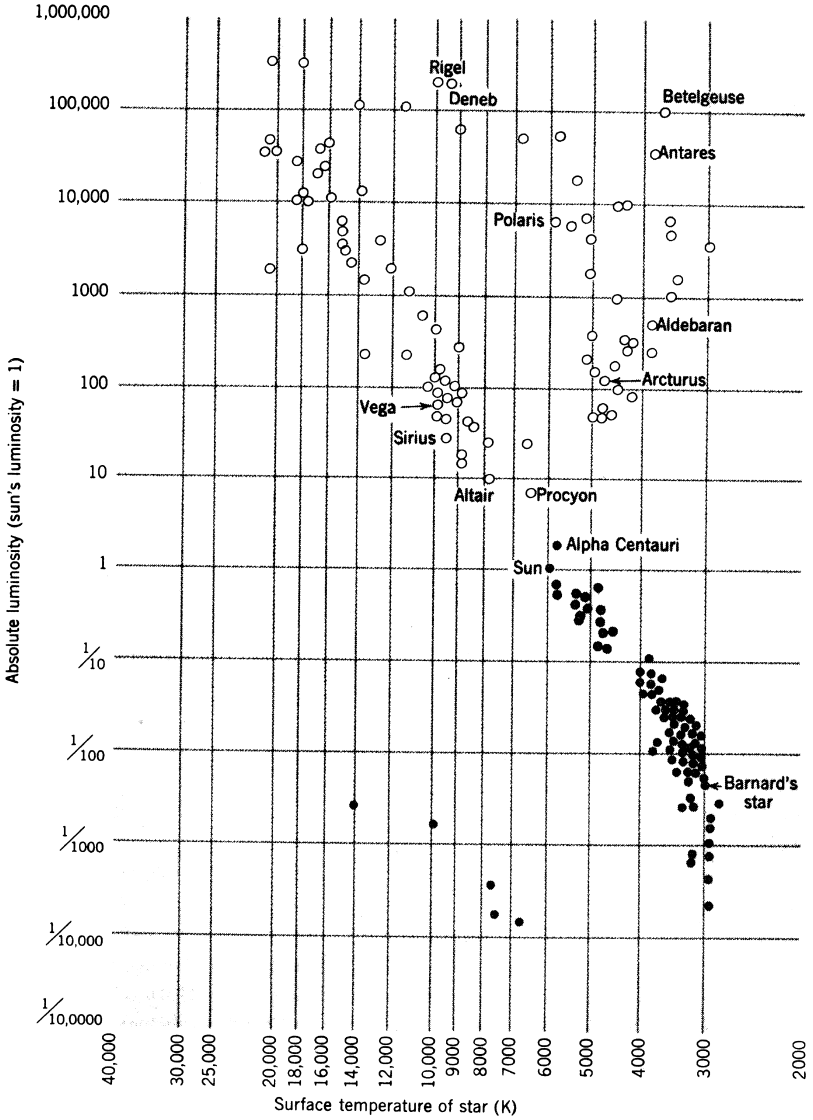


FIGURE 1 H–R diagram for the 100 brightest stars (open circles) and the 90 nearest stars (filled circles). [Reprinted with permission from Jastrow, R., and Thompson, M. H. (1984). "Astronomy: Fundamentals and Frontiers," 4th ed., Wiley, New York. © 1984, Robert Jastrow.]

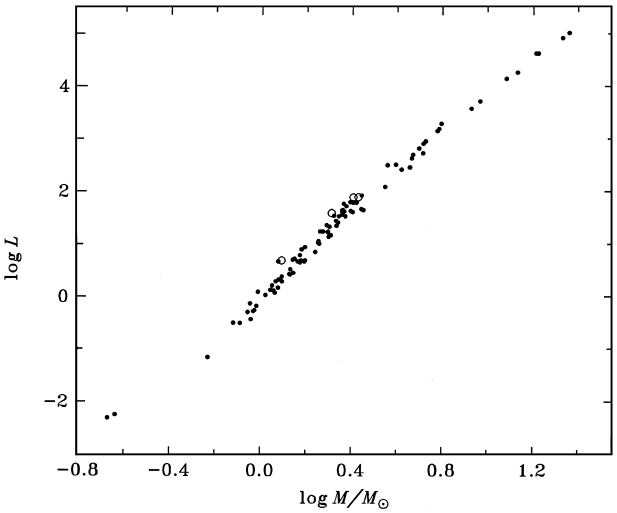


FIGURE 2 Observed masses of main-sequence stars in units of the solar mass as determined from binary orbits plotted against their luminosities in units of the solar luminosity (filled circles). The open circles are subgiants or giants. The observational errors in mass are less than 2%. [Reprinted with permission from Andersen, J. (2000). In "Unsolved Problems in Stellar Evolution" (M. Livio, ed.), Cambridge Univ. Press, Cambridge, UK. Reprinted with permission of Cambridge, University Press.]

to the evolutionary time spent in that region; thus, the main-sequence or core hydrogen-burning phase is that in which stars spend most of their lifetime.

G. Mass–Luminosity Relation

For stars known to be on the main sequence, the observational information on M and L can be combined to produce a reasonably smooth mass–luminosity relation (Fig. 2). The slope of this relation is not constant, however. Empirically, the luminosity increases approximately as M^3 for stars in the region of 10–20 $M_\odot$ and as $M^{4.5}$ for stars between 1–2 $M_\odot$. A few additional masses are available for stars less massive than the sun, but the accuracy is not as good as for those shown in Fig. 2. Theoretical models, with composition close to that of the sun at the zero-age main sequence, agree well with the observed points. There are two reasons for the scatter in the observed points: first, some of the observed stars are not strictly at zero age but have evolved slightly, with a corresponding increase in L; and second, there are small differences in metal abundance among the stars, which affect the luminosities. Because stars evolve and change their luminosity considerably while retaining the same mass, a mass–luminosity relation cannot be specified for most regions of the H–R diagram, only near the zero-age main sequence.

H. Stellar Ages

Although the age of the solar system (and therefore presumably the sun) can be determined accurately from radioactive dating of the oldest moon rocks and meteorites, the ages of other stars cannot be determined directly. Several indirect methods exist, which depend for the most part upon the theory of stellar evolution, and generally, there is some uncertainty in the derived ages.

1. The H–R diagrams of galactic or globular clusters can be compared with evolutionary calculations. The stars in a cluster are assumed to be all of the same age; to have the same composition; and, except for the very nearest clusters, to all lie at the same distance from the earth. The observed cluster diagram is compared with a theoretical line of constant age, known as an *isochrone*, obtained by calculating the evolution in the H–R diagram of a set of stars with different masses but the same composition and by connecting points on their evolutionary tracks that correspond to the same elapsed times since formation. This procedure is illustrated in Figs. 3 and 4, which show how the age is determined for two different clusters. Fig. 3 shows the Hyades, a nearby galactic cluster with a distance of 46.3 pc. The age is determined from the positions in the

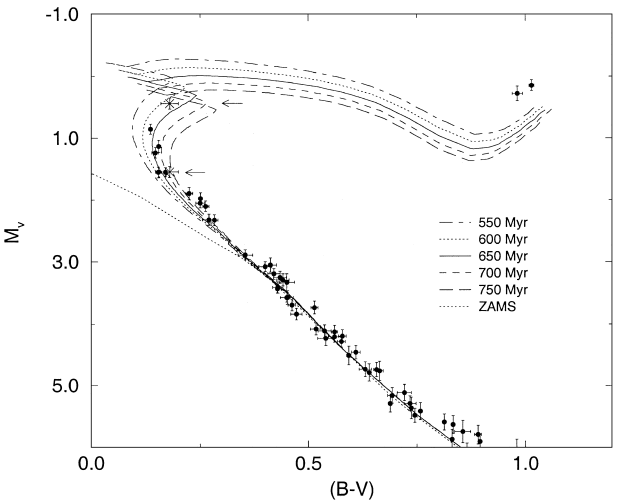


FIGURE 3 H–R diagram for the Hyades cluster. The luminosity is given in terms of the absolute visual magnitude M_V, and the observed color (B − V) is an indicator of surface temperature (increasing to the left). The filled circles with error bars correspond to the observations of the individual single stars, whose distances have been determined from Hipparcos parallaxes. The symbols with arrows correspond to the two components of a spectroscopic binary. The dotted curve corresponds to the theoretical zero-age main sequence for the composition of this cluster. Other curves are theoretical isochrones, that is, predictors of how the H–R diagram should look at the indicated ages. [Reprinted with permission from Perryman, M. A. C., *et al*. (1998). *Astron. Astrophys.* **331,** 81. © European Southern Observatory.]

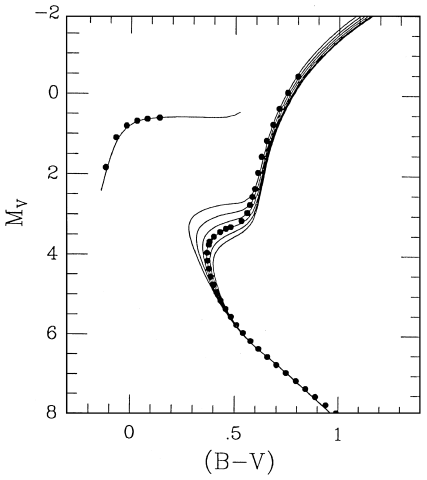


FIGURE 4 H–R diagram for the stars in the globular cluster M92. The axes have the same meaning as in Fig. 3. The filled circles are averaged observed stellar positions. The solid curves are theoretical isochrones for the ages (top to bottom) of 8, 10, 12, 14, 16, and 18 Gyr. The evolved stars in the upper-left part of the diagram are on the "horizontal branch," where they are burning He in their cores. [Adapted with permission from Pont, F., Mayor, M., Turon, C., and VandenBerg, D. A. (1998). *Astron. Astrophys.* **329,** 87. © European Southern Observatory.]

H–R diagram of the stars near the "turnoff" point, usually identified as the point of highest effective temperature on the main sequence. The time spent by a star on the main sequence decreases with increasing mass; thus, stars with masses above that corresponding to the turnoff mass have already left the main sequence and have evolved to the red giant region. In this case the age is determined to be 625 Myr, with an error of ± 50 Myr. Fig. 4 shows the globular cluster M92, one of the oldest objects in the galaxy. Its distance is about 9000 pc, and the abundance of metals, such as iron, is less than $1/100$ that of the sun. The distance to the cluster is determined by fitting the location of its main sequence in the H–R diagram to the locations of nearby stars of the same (reduced) metallicity whose distances are known from Hipparcos parallaxes. The age derived from isochrone fitting is 14 ± 1.2 Gyr. Note that the luminosity of the turnoff is much fainter than that for the Hyades cluster. In the cluster fitting method, the age of the stars is really being determined from the nuclear burning time scale.

2. Main-sequence stars like the sun have chromospheric activity, analogous to solar activity, such as flares, prominences, and chromospheric emission lines, which declines with age. An often-used indicator of chromospheric activity is the strength of two emission features in the Ca lines between 3900 and 4000 Å. The age-activity relation is calibrated by use of objects whose ages have been determined by other methods, such as the sun and the Hyades cluster.

3. Certain special types of stars are known to be young, that is, with ages of 1 to a few million years. One example is the T Tauri stars, which are identified by their high lithium abundances, the presence of emission lines of hydrogen, their irregular variability, their association with dark clouds (star-forming regions) in the galaxy, and their location in the H–R diagram above and to the right of the main sequence. Typical masses are 0.5–2 $M_\odot$, and typical values of $T_{\rm eff}$ are around 4000 K. A second example is the massive stars near the upper end of the main sequence, with $L/L_\odot \approx 10^5$. They are also known to be young because their nuclear burning time scale is only a few Myr.

4. The stars in the galaxy have been roughly divided, according to age, into two populations. The Population I stars are associated with the galactic disk, have relatively small space motions with respect to the sun's, and have metal (such as iron) abundances similar to the sun's. Although precise ages cannot be determined, this population is younger as a group than the Population II stars, which are distributed in the galactic halo and in globular clusters and have high space velocities and low metal abundances. These objects were formed early in the history of the galaxy (see Fig. 4), and their low metal abundance supports the point of view that the elements heavier than helium have been synthesized in the interiors of the stars. Subsequently, some of the synthesized material was ejected from the stars in supernova explosions and in winds from evolved stars, so that later generations of stars form from interstellar material that has been gradually enriched in the heavy elements.

III. PHYSICS OF STELLAR INTERIORS

A. Time Scales

The dynamical time scale, the contraction time scale, the cooling time scale, and the nuclear time scale are important at different stages of stellar evolution. If the star is not in hydrostatic equilibrium, it will evolve on the dynamical time scale given by $t_{\rm ff} = [3\pi/(32G\rho)]^{1/2}$; this is, in fact, the free-fall time from an initial density ρ. Examples of unstable stars that evolve on this time scale are protostars or the evolved cores of massive stars, which undergo photodissociation of the iron nuclei and consequently are forced into gravitational collapse. The resulting supernova outburst, of course, also takes place on a dynamical time scale. Another example is the light variation in Cepheid variables, which is caused by radial oscillations about an equilibrium state; the oscillation period is of the same order as $t_{\rm ff}$. Characteristic time scales are 10^5 years for the protostar collapse, 0.1 sec for the collapse of the iron core, and 10 days for the pulsation period of a Cepheid.

Stars in hydrostatic equilibrium but without a substantial nuclear energy source undergo a slow contraction with the release of gravitational energy. The associated time scale, known as the Kelvin-Helmholtz time scale, is given by the gravitational energy divided by the luminosity: $t_{\rm KH} \approx GM^2/(RL)$, where R is the final radius and L is the average luminosity. Stars contracting to the main sequence evolve on this time scale as do, for example, stars at the end of the main-sequence phase when they run out of hydrogen fuel in the core. For the present sun, the value of $t_{\rm KH}$ is about 3×10^7 years, which represents the time required for it to contract to its present size without any contributions from nuclear reactions. Because $t_{\rm KH}$, as well as the other time scales discussed later, depend on L, they are controlled by the time required for energy to be transported from the interior of the star to the surface.

Stars that derive all their energy from nuclear burning evolve on a nuclear time scale, which can be estimated from the total available nuclear energy divided by the luminosity. For the case of hydrogen burning, four protons, each with a mass of 1.008 atomic mass units (amu), combine to form a helium nucleus which has 4.0027 amu. The difference in mass of 0.0073 amu per proton is released as energy. The corresponding time scale is $t_{\rm nuc} = 0.007Mc^2/L$, where M is the amount of mass of H that is burned and c is the velocity of light. For the sun on the main sequence, $X \approx 0.71$, $L \approx L_\odot$, and about 14% of the hydrogen is actually burned, so the estimate gives $t_{\rm nuc} \approx 1. \times 10^{10}$ years. In subsequent evolutionary phases the sun continues to burn hydrogen but at a higher rate. For a star of 30 $M_\odot$, $L \approx 2 \times 10^5\, L_\odot$, about half of the hydrogen is burned, and $t_{\rm nuc} \approx 5 \times 10^6$ years. For helium burning, where three helium nuclei combine to produce carbon, the energy release per atomic mass unit is a factor 10 smaller than for hydrogen burning, and correspondingly, the time spent by a star in the core helium burning phase is a factor of 5–10 shorter than its main sequence lifetime, depending on the luminosity and the amount of hydrogen burning going on in a shell at the same time.

A final time scale associated with stellar evolution is the cooling time scale. For white dwarf stars that can no longer contract and also have no more nuclear fuel available, the radiated energy must be supplied by cooling of the hot interior. The electrons by this time are highly degenerate (see later discussion), so the available thermal energy is only that of the ions. The time scale can be estimated from

$$t_{\rm cool} \approx E_{\rm thermal}/L = 1.5R_{\rm g}TM/(\mu_{\rm A}L),$$

where T is the mean internal temperature, $R_{\rm g}$ is the gas constant, and $\mu_{\rm A}$ is the mean atomic weight (in amu) of the ions. For example, the cooling time of a white dwarf of 0.7 $M_\odot$ composed of carbon from the beginning of the white dwarf phase to a luminosity of $L = 10^{-3}\ L_\odot$ is 1.15×10^9 years. The cooling time scale also applies to substellar objects once they have contracted to their limiting radius.

B. Equation of State of a Gas

In the pre-main-sequence and main-sequence stages of the evolution of most stars, the ideal gas equation holds. The pressure of the gas is given by $P = NkT$, where k is the Boltzmann constant and N is the number of free particles per unit volume. If X, Y, and Z are the mass fractions of H, He, and heavy elements, respectively, and the gas is fully ionized, then

$$N = (2X + 0.75Y + 0.5Z)\rho/m_{\rm H},$$

where $m_{\rm H}$ is the mass of the hydrogen atom. It has been assumed that the number of particles contributed per nucleus is 2 for H, 3 for He, and $A/2$ (an approximation) for the heavy elements, where A is the atomic weight. In the outer layers of the star, N must be adjusted to take into account partial ionization. The internal energy for an ideal gas is $1.5kT$ per particle or $1.5kTN/\rho = 1.5P/\rho$ per unit mass. The equation of state can also be written $P = R_{\rm g}\rho T/\mu$, where μ, the mean atomic weight per free particle, is given by $\mu^{-1} = 2X + 0.75Y + 0.5Z$ for a fully ionized gas.

Under conditions of high temperature and low density, the pressure of the radiation must be added. According to quantum theory, a photon has energy $h\nu$, where ν is the frequency and h is Planck's constant, and momentum $h\nu/c$. The radiation pressure is the net rate of transfer of momentum per unit area, normal to an arbitrarily oriented surface. Under conditions in stellar interiors where the radiation is nearly isotropic and the system is near thermodynamic equilibrium, it can be shown that the radiation pressure is given by $P_{\rm R} = \frac{1}{3}aT^4$, where a, the radiation density constant, equals 7.56×10^{-15} erg cm^{-3} degree^{-4}.

At high densities an additional physical effect must be considered, the phenomenon of degeneracy. The effect is a consequence of the Pauli exclusion principle, which does not permit more than one particle to occupy one quantum state at the same time; it applies to elementary particles with half-integral spin, such as electrons or neutrons. In the case of electrons bound in an atom, the principle governs the distribution of electrons in various energy states. If the lowest energy states are filled, any electrons added subsequently must go into states of higher energy; only a discrete set of states is available. The same principle applies to free electrons. In a momentum interval $dp_x\, dp_y\, dp_z$ and in a volume element $dx\, dy\, dz$, the number of states is $(\frac{2}{h^3})\, dp_x\, dp_y\, dp_z\, dx\, dy\, dz$, where the factor of 2 arises from the two possible directions of the electron spin. Degeneracy occurs when most of these states,

up to some limiting momentum p, are occupied by particles. Complete degeneracy is defined as the situation in which all states are occupied up to the limiting momentum p_0 and all higher states are empty. Under this assumption, if the particles are electrons, their pressure can be derived to be $P_e = 1.0036 \times 10^{13}(\rho/\mu_e)^{5/3}$ dyne cm^{-2}, where μ_e, the mean atomic weight per free electron, is $2/(1+X)$ for a fully ionized gas. This formula is valid if p_o is low enough so that the electron velocities are not relativistic. In the limiting case that all electrons are moving with the velocity of light, the corresponding expression is $P_e = 1.2435 \times 10^{15}(\rho/\mu_e)^{4/3}$ dyne cm^{-2}. In either case most of the electrons are forced into such high momentum states that their pressure, defined again as the rate of transport of momentum per unit area, is much higher than the ideal gas pressure. The electrons in stellar interiors become degenerate when the density reaches 10^3–10^6 g cm^{-3}, depending on T. The protons or neutrons become degenerate only at much higher densities, about 10^{14} g cm^{-3}; these densities are characteristic of neutron stars, and it is the degenerate pressure of the neutrons that supports the star against gravity. The regions in density and temperature where ideal gas, radiation pressure, and electron degeneracy dominate in the equation of state are shown in Fig. 5.

C. Energy Sources

From the previous discussion of time scales, it is clear that only two fundamental energy sources are available to a star. Gravitational energy can be released either on a dynamical or a Kelvin-Helmholtz time scale. Nuclear energy in moderate-mass stars is produced by conversion of hydrogen to helium or by conversion of helium to carbon and oxygen. Only in the most massive stars do nuclear reactions proceed further to the synthesis of the heavier elements up to iron and nickel. During cooling phases stars draw on their thermal energy, which, however, has been produced in the past as a consequence of nuclear reactions or gravitational contraction. As the star evolves, the energy is taken up by radiation, heating of the star, expansion, neutrino emission, mass loss, ionization of atoms, and dissociation of nuclei. Here, we discuss in somewhat more detail the nuclear energy source.

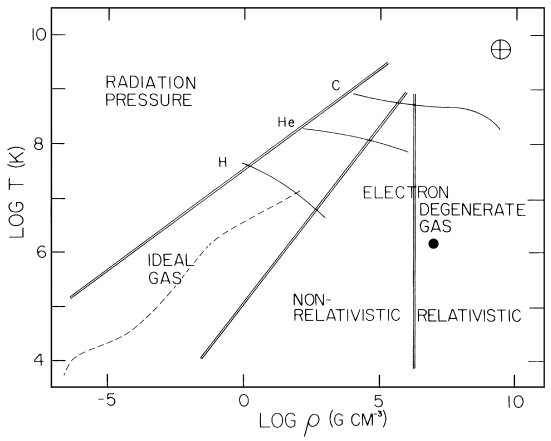


FIGURE 5 Density–temperature diagram. The double lines separate regions in which different physical effects dominate in the equation of state. Solid lines, denoted by H, He, and C, respectively, indicate the central conditions in stars that are undergoing burning of hydrogen, helium, and carbon in their cores. The dashed line represents the structure of the present sun, while the solid dot gives typical central conditions in a white dwarf of 0.8 $M_\odot$. The circled cross gives central conditions in a star of 25 $M_\odot$ just before the collapse of the iron core.

The nuclear processes involve reactions between charged particles. At the temperatures characteristic of nuclear burning regions in stars, the gas may safely be assumed to be fully ionized. The Coulomb repulsive force between particles of like charge presents a barrier for reactions between the ions. Particles must come within 10^{-13} cm of each other before the strong (but short-range) nuclear attractive force overcomes the Coulomb force. The energy that is required for particles to pass through the Coulomb barrier is $E_{\text{coulomb}} = Z_1 Z_2 e^2/r$, where r is the required separation of 10^{-13} cm, e is the electronic charge, and Z_1 and Z_2 are, respectively, the charges on the two particles involved, in units of the electronic charge. Evidently, favorable conditions for reactions will occur most readily for particles of small charge, but even for two protons $E_{\text{coulomb}} = 1000$ keV. The only source for the energy is the thermal energy of the particles, which, however, in the temperature range around 10^7 K is only $\frac{3}{2}kT \approx 1$ keV. It would seem that nuclear reactions under these conditions would not be possible.

However, three factors contribute to a small but nonzero probability of reaction. (1) The particles have a Maxwell velocity distribution, so that a small fraction of them have thermal energies much higher than the average. (2) According to the laws of quantum mechanics, there is a small probability that a particle with much less energy than required can actually "tunnel" through the Coulomb barrier. (3) A star has a very large number of particles, so that even if an individual particle has a very small probability of reacting, enough reactions do occur to supply the required energy. It turns out that in main-sequence stars particles with energies of about 15–30 keV satisfy the two requirements of existing in sufficient numbers and having enough energy to have a reasonable change of passing through the Coulomb barrier.

At temperatures of 1 to 3×10^7 K, which characterize the interiors of most main-sequence stars, only reactions between particles of low atomic number Z need be considered, usually those involving protons. However,

even if the Coulomb barrier is overcome, there is still only a small probability that a nuclear reaction will occur. This nuclear probability varies widely from one reaction to another, and it can be determined theoretically for only the simplest systems; in general, these probabilities must be measured for each reaction in laboratory experiments at low energy. An intense effort was made during the period 1960–1980 to make measurements of reasonable accuracy of all the nuclear reactions that are important in stars. Nevertheless, in some cases measurements cannot be made at the low energies appropriate for main-sequence stars, and the probabilities have to be extrapolated from measurements at higher energy. The group headed by W. A. Fowler at the California Institute of Technology led in the task of carrying out these difficult measurements.

Two reaction sequences have been identified that result in the conversion of four protons into one helium nucleus, the proton-proton (pp) chains and the CNO cycle. The reactions of the main branch of the pp chains (PP I) are as follows:

$$^1\mathrm{H} + {}^1\mathrm{H} \rightarrow {}^2\mathrm{D} + e^+ + \nu \tag{1}$$

$$^2\mathrm{D} + {}^1\mathrm{H} \rightarrow {}^3\mathrm{He} + \gamma \tag{2}$$

$$^3\mathrm{He} + {}^3\mathrm{He} \rightarrow {}^4\mathrm{He} + {}^1\mathrm{H} + {}^1\mathrm{H} \tag{3}$$

The symbol e^+ denotes a positron; ν is the neutrino; and ^2D is the deuterium nucleus, composed of one proton and one neutron. The positron immediately reacts with an electron, with the annihilation of both and the production of energy. The total energy production of the sequence is 26.7 MeV or 4.27×10^{-5} erg, corresponding to the mass difference between four protons and one ^{4}He nucleus, times c^2. Each sequence produces two neutrinos, which immediately escape from the star, as they interact only very weakly with matter, carrying with them a certain fraction of the energy. The remainder of the energy is deposited locally in the star, typically in the form of gamma rays.

Reaction (1) turns out to be very improbable, as the conversion of a proton into a neutron requires a beta decay. The reaction rate is so slow that it cannot be measured experimentally and must be calculated theoretically. At the center of the sun, a proton has a mean lifetime of almost 10^{10} years before it reacts with another proton; this reaction controls the rate of energy production, as subsequent reactions are more rapid. The neutrino in Reaction (1) carries away an average energy of 0.265 MeV. Once ^{2}D is formed, it immediately (within 1 sec) captures another proton to form ^{3}He. This reaction can be measured in the laboratory down to energies close to those where reactions can occur in stars. Any ^{2}D initially present in the star will be burned by this reaction at temperatures around 10^6 K. When Reactions (1) and (2) have occurred twice, the two resulting nuclei of ^{3}He combine to form ^{4}He and two protons. There is perhaps a 10% uncertainty in this reaction rate and in most others in the pp chains.

The second branch of the pp chains (PP II) occurs when the ^{3}He nucleus produced in Reaction (2) reacts with ^{4}He. The sequence of reactions is then

$$^3\mathrm{He} + {}^4\mathrm{He} \rightarrow {}^7\mathrm{Be} + \gamma \tag{4}$$

$$^7\mathrm{Be} + e^- \rightarrow {}^7\mathrm{Li} + \nu \tag{5}$$

$$^7\mathrm{Li} + {}^1\mathrm{H} \rightarrow {}^4\mathrm{He} + {}^4\mathrm{He} \tag{6}$$

Under conditions of the solar interior, Reaction (4) proceeds at approximately one-sixth of the rate of Reaction (3), and as the temperature increases it becomes relatively more important. Reaction (5) involves an electron capture on the ^{7}Be and the production of a neutrino with energy 0.86 MeV (90% of the time) or 0.38 MeV (10% of the time). The ^{7}Li nucleus immediately captures a proton and produces two nuclei of ^{4}He. This reaction also results in the destruction of any ^{7}Li that may be present at the time of the star's formation, since it becomes effective at temperatures of 2.8×10^6 K or above. Note that the second branch of the pp chains has an average neutrino loss of about 1.1 MeV, but that otherwise the net result is the same: the conversion of four protons to a ^{4}He nucleus.

The third branch of the pp chains (PP III) occurs if a proton, rather than an electron, is captured by ^{7}Be:

$$^7\mathrm{Be} + {}^1\mathrm{H} \rightarrow {}^8\mathrm{B} + \gamma \tag{7}$$

$$^8\mathrm{B} \rightarrow e^+ + \nu + {}^8\mathrm{Be} \rightarrow {}^4\mathrm{He} + {}^4\mathrm{He} \tag{8}$$

Less than 0.1% of pp-chain completions in the sun occur through Reactions (7) and (8); however, they are very important because the neutrino produced has an average energy of 6.7 MeV, high enough to be detectable in all current solar neutrino experiments.

The total rate of energy generation by the pp chains involves a complicated calculation of the rates of all the reactions given above, which are functions of the temperature, density, and concentration of the species involved. If T_6 is the temperature in units of 10^6 K, then under the assumption of equilibrium, that is, the abundances of the intermediate species ^{2}D, ^{3}He, ^{7}Be, and ^{7}Li reach steady state, the energy generation rate for the pp chains can be written

$$\epsilon_{\mathrm{pp}} = 2.38 \times 10^6 \psi g_{11} \rho X^2 T_6^{-2/3} \times \exp\left(-33.80 T_6^{-1/3}\right) \mathrm{erg\ g^{-1}\ sec^{-1}}, \tag{9}$$

where ρ is the density, X is the fraction of hydrogen by mass, and

$$g_{11} = 1 + 0.0123\, T_6^{1/3} + 0.0109\, T_6^{2/3} + 0.0009\, T_6.$$

The function ψ is a correction factor, between 1 and 2, that accounts for (1) the increase by a factor of 2 in the energy production rate, with respect to that for PP I, that occurs when the reactions go through PP II or PP III, and (2) the different neutrino energies in the three chains. For the center of the present sun, $\psi \approx 1.5$. For temperatures lower than and densities higher than those of the present solar center, the effect of electron shielding as well as departure from equilibrium may have to be taken into account.

Protons can also interact with the CNO nuclei, but because the Coulomb barrier is higher these reactions become more important than the pp chains only at temperatures above 2×10^7 K. The reactions are the following:

$$^{12}\mathrm{C} + {}^{1}\mathrm{H} \rightarrow {}^{13}\mathrm{N} + \gamma \quad (10)$$

$$^{13}\mathrm{N} \rightarrow {}^{13}\mathrm{C} + e^{+} + \nu \quad (11)$$

$$^{13}\mathrm{C} + {}^{1}\mathrm{H} \rightarrow {}^{14}\mathrm{N} + \gamma \quad (12)$$

$$^{14}\mathrm{N} + {}^{1}\mathrm{H} \rightarrow {}^{15}\mathrm{O} + \gamma \quad (13)$$

$$^{15}\mathrm{O} \rightarrow {}^{15}\mathrm{N} + e^{+} + \nu \quad (14)$$

$$^{15}\mathrm{N} + {}^{1}\mathrm{H} \rightarrow {}^{12}\mathrm{C} + {}^{4}\mathrm{He} \quad (15)$$

The net effect is the conversion of four protons into a helium nucleus, two positrons (which annihilate), and two neutrinos (which carry away energies of 0.71 and 1.0 MeV, respectively). The CNO nuclei simply act as catalysts; however, their relative abundances change as a result of the operation of the cycle. Note from Table I that ^{12}C is considerably more abundant than ^{14}N in the solar system. However, the rate of Reaction (10), which uses up ^{12}C, under stellar conditions is about 100 times faster than that of Reaction (13), which uses up ^{14}N. All other reactions are much faster than these two. The reaction chain tends to reach equilibrium, in which each CNO nuleus is produced as fast as it is destroyed. The equilibrium is obtained only when most of the ^{12}C and other participating nuclei are converted to ^{14}N. This change in relative abundances could be observed if the layers in which the reactions occur could later be mixed to the surface of the star.

A secondary branch of the CNO cycle affects the abundance of ^{16}O. Starting at ^{15}N, the reactions are

$$^{15}\mathrm{N} + {}^{1}\mathrm{H} \rightarrow {}^{16}\mathrm{O} + \gamma \quad (16)$$

$$^{16}\mathrm{O} + {}^{1}\mathrm{H} \rightarrow {}^{17}\mathrm{F} + \gamma \quad (17)$$

$$^{17}\mathrm{F} \rightarrow {}^{17}\mathrm{O} + e^{+} + \nu \quad (18)$$

$$^{17}\mathrm{O} + {}^{1}\mathrm{H} \rightarrow {}^{14}\mathrm{N} + {}^{4}\mathrm{He} \quad (19)$$

This branch accounts for only 0.1% of the completions, but the ^{16}O is generated less rapidly than it is converted into ^{14}N by Reactions (17)–(19), so that when the branch comes into equilibrium the abundance ratio O/N will be reduced. The overall effect of the CNO cycle, apart from its energy generation, is the conversion of 98% of the CNO isotopes into ^{14}N. Once the cycle reaches equilibrium, the energy generation can be calculated by the rate of the slowest reaction in the main chain (Reaction 13) multiplied by the energy released over the whole cycle (24.97 MeV, after neutrino losses):

$$\epsilon_{\mathrm{CNO}} = 8.67 \times 10^{27} g_{14,1} \rho X X_{\mathrm{CNO}} T_6^{-2/3} \times \exp\left(-152.28 T_6^{-1/3}\right) \mathrm{erg\ g^{-1}\ sec^{-1}}, \quad (20)$$

where X_{CNO} is the total mass fraction of all CNO isotopes and

$$g_{14,1} = 1 + 0.0027\ T_6^{1/3} - 0.00778\ T_6^{2/3} - 0.000149\ T_6.$$

At $T_6 \approx 25$, the energy generation goes approximately as T^{17}.

The synthesis of elements heavier than helium has proved in the past to be a considerable problem, primarily because there is no stable isotope of atomic mass 5 or 8. Thus, if two ^{4}He nuclei react to produce ^{8}Be, the nucleus will immediately decay back to He. If a proton reacts with ^{4}He, the result is ^{5}Li, which also is unstable. Helium burning must, in fact, proceed through a three-particle reaction, which can be represented as follows:

$$^{4}\mathrm{He} + {}^{4}\mathrm{He} \leftrightarrow {}^{8}\mathrm{Be} \quad (21)$$

$$^{8}\mathrm{Be} + {}^{4}\mathrm{He} \leftrightarrow {}^{12}\mathrm{C} + \gamma + \gamma \quad (22)$$

Although the ^{8}Be is unstable, its decay is not instantaneous. Under suitable conditions, e.g., $T = 10^8$ K and $\rho > 10^5$ g cm^{-3}, it was shown by E. Salpeter that a sufficient equilibrium abundance exists so that a third ^{4}He nucleus can react with it [Reaction (22)]. It was also predicted by F. Hoyle that in order for this triple-alpha process to proceed at a significant rate, Reaction (22) must be resonant; that is, there must be an excited nuclear state accessible in the range of stellar energies whose presence greatly enhances the probability that the reaction will occur. This resonance was later found in experiments performed at the Kellogg Radiation Laboratory at the California Institute of Technology. The amount of energy produced per reaction is 7.275 MeV, or 0.606 MeV per atomic mass unit, a factor of 10 less than in hydrogen burning (per atomic mass unit). An expression for the rate of energy production is

$$\epsilon_{3\alpha} = 5.09 \times 10^{11} \rho^2 Y^3 T_8^{-3} \times \exp\left(-44.027 T_8^{-1}\right) \mathrm{erg\ g^{-1}\ sec^{-1}}, \quad (23)$$

where $T_8 = T/10^8$ K, Y is the mass fraction of ^{4}He, and electron screening has not been included. This reaction has a very steep temperature sensitivity, with $\epsilon \propto T^{40}$ at $T_8 = 1$. A further helium burning reaction, which occurs at slightly higher temperatures than the triple-alpha process, is

$$^{12}\mathrm{C} + {}^{4}\mathrm{He} \rightarrow {}^{16}\mathrm{O} + \gamma, \tag{24}$$

with an energy production of 7.162 MeV per reaction and an uncertain reaction rate. Reactions (21), (22), and (24) probably produce most of the carbon and oxygen in the universe.

D. Energy Transport

The transport of energy outward from the interior of a star to its surface depends in general on the existence of a temperature gradient. Heat will be carried by various processes from hotter regions to cooler regions; the processes that need to be considered include (1) radiation transport, (2) convective transport, and (3) conductive transport. Neutrino transport must be considered only in exceptional circumstances when the density is above 10^{10} g cm^{-3}; at lower densities, neutrinos produced in stars simply escape directly without interacting with matter. In each case a relation must be found between the energy flux F_r, defined as the energy flow per unit area per unit time at a distance r from the center of the star, and the temperature gradient dT/dr.

Energy transport by radiation depends on the emission of photons in hot regions of the star and absorption of them in slightly cooler regions. The radiation field may be characterized by the *specific intensity* I_ν, which is defined so that $I_\nu \cos\theta \, d\nu \, d\omega \, dt \, dA$ is the energy carried by a beam of photons across an element of area dA in time dt in frequency interval ν to $\nu + d\nu$ into an element of solid angle $d\omega$, in a direction inclined by $\cos\theta$ to the normal to dA. The equation of transfer shows how the intensity of a beam is changed as it interacts with matter. The *mass emission coefficient* j_ν is defined so that $j_\nu \, \rho \, dV \, d\nu \, d\omega \, dt$ is the energy emitted by the volume element dV into $d\omega$ in time dt in the frequency range $d\nu$. The corresponding *mass absorption coefficient* κ_ν is defined so that the energy absorbed in the same intervals is $\kappa_\nu \, \rho \, I_\nu \, dV \, d\nu \, d\omega \, dt$. If we consider both the absorption and the emission of the radiation passing through a cylinder with length ds and cross section dA, the equation of transfer becomes

$$\frac{dI_\nu}{ds} = -\kappa_\nu \rho I_\nu + j_\nu \rho. \tag{25}$$

Suppose that the direction ds is inclined by an angle θ to the radial direction in the star so that the projected distance element $dr = ds \cos\theta$. Also, define the optical depth τ_ν by

$$d\tau_\nu = -\kappa_\nu \rho dr \tag{26}$$

and the equation of transfer becomes

$$\cos\theta \frac{dI_\nu}{d\tau_\nu} = I_\nu - \frac{j_\nu}{\kappa_\nu}. \tag{27}$$

In the stellar interior the mean free path of a photon before it is absorbed is only 1 cm or less. Thus, a typical photon is absorbed at practically the same temperature as it is emitted. These conditions are so close to strict thermodynamic equilibrium, where the specific intensity is given by the Planck function $B_\nu(T)$, that the ratio j_ν/κ_ν can be shown to be the same as it is in strict thermodynamic equilibrium, namely, $j_\nu/\kappa_\nu = B_\nu(T)$. The equation of transfer in stellar interiors can thus be expressed as

$$\cos\theta \frac{dI_\nu}{d\tau_\nu} = I_\nu - B_\nu(T). \tag{28}$$

This equation can be integrated over all frequencies and solved for the flux in terms of the temperature gradient for conditions appropriate to the stellar interior, that is, where the temperature change is negligible over the mean free path of a photon, to give

$$F_r = -(4ac)(3\kappa\rho)^{-1} T^3 (dT/dr), \tag{29}$$

where c is the velocity of light and a is the radiation density constant. Here, κ is the absorption coefficient averaged over frequency according to the so-called *Rosseland mean*:

$$\frac{1}{\kappa} = \frac{\int_0^\infty \frac{dB_\nu(T)/dT \, d\nu}{\kappa_{\nu,a}[1 - \exp(-h\nu/kT)] + \kappa_{\nu,s}}}{\int_0^\infty dB_\nu(T)/dT \, d\nu}. \tag{30}$$

Here, $\kappa_{\nu,a}$ refers to processes of true absorption, which have to be corrected for induced emission, and $\kappa_{\nu,s}$ refers to scattering processes. Equation (29) is known as the diffusion approximation for radiation transfer, an appropriate nomenclature because a photon is absorbed almost immediately after it is emitted, so that an enormous number of absorptions, re-emissions, and scatterings must occur before the energy of a photon as transmitted to the surface. The time required for energy to diffuse in this manner from the center of the sun to its surface is about 10^5 years. The quality of the radiation changes during this process. As the photons diffuse to lower temperatures, their energy distribution corresponds closely to the Planck distribution at the local T, because matter and radiation are well coupled. Thus, the gamma rays produced by nuclear reactions at the center are gradually transformed into optical photons by the time the energy reaches the surface.

As Eq. (29) indicates, the energy transport in a radiative star is controlled by the opacity κ. Numerous atomic processes contribute to this quantity, and in general, the structure of a star can be calculated only with the aid of detailed tables of the opacity, calculated as a function of ρ, T, and the chemical composition. Starting at the highest temperatures characteristic of the stellar interior and proceeding to lower temperatures, the main processes are (1) electron scattering, also known as Thomson scattering, in which a proton undergoes a change in direction but

no change in frequency during an encounter with a free electron; (2) free-free absorption, in which a photon is absorbed by a free electron in the vicinity of a nucleus, with the result that the photon is lost and the electron increases its kinetic energy; (3) bound-free absorption on metals, also known as photoionization, in which the photon is absorbed by an atom of a heavy element (e.g., iron) and one of the bound electrons is removed; (4) bound-bound absorption of a heavy element, in which the photon induces an upward transition of an electron from a lower quantum state to a higher quantum state in the atom; (5) bound-free absorption on H and He, which generally occurs near stellar surfaces where these elements are being ionized; (6) bound-free and free-free absorption by the negative hydrogen ion H−, which forms in stellar atmospheres in layers where H is just beginning to be ionized (example, the surface of the sun); (7) bound-bound absorptions by molecules, which can occur only in the atmospheres of the cooler stars ($T_{\rm eff} < 4000$ K, although even the sun shows a few molecular features in its spectrum); and (8) absorption by dust grains, which can occur in the early stages of protostellar evolution and possibly in the atmospheres of brown dwarfs, at temperatures below the evaporation temperature of grains (1400–1800 K).

The Thomson scattering from free electrons (in units of $\mathrm{cm}^2\ \mathrm{g}^{-1}$) is given by $\kappa = 0.2(1 + X)$. For stars this process is important generally in regions above the uppermost double line in Fig. 5, corresponding to the interiors of massive stars. The free-free absorptions and the bound-free absorptions on heavy elements have the approximate dependence $\kappa \propto \rho T^{-3.5}$. As one moves outward in a star from higher to lower T, the number of bound electrons per atom increases and, correspondingly, κ increases. A maximum in κ occurs in the range $10^4\ \mathrm{K} < T < 10^5$ K depending on density; this range corresponds to the zones where H and He, the most abundant elements, undergo ionization. Below this range in T, the above dependence is no longer valid, and κ drops rapidly with decreasing T, reaching a mimimum of about 10^{-2} at $T = 2000$ K. At lower T, where grains exist, the opacities are higher, on the order of $1\ \mathrm{cm}^2\ \mathrm{g}^{-1}$.

Under certain conditions, the temperature gradient given by Eq. (29) can be unstable, leading to the onset of convection. Suppose that a small element of material in a star is displaced upward from its equilibrium position. It may reasonably be assumed that the element remains in pressure equilibrium with its surroundings. Thus, if the material has the equation of state of an ideal gas, an upwardly displaced element with density less than that of the surroundings will have a temperature greater than that of the surroundings. If, as the element rises, its density decreases more rapidly than that of the surroundings, it will continue to feel a buoyancy force and will continue to rise. In this case the layer is unstable to convection, and heat is transported outward by the moving elements themselves. If, however, the density of the upward-moving element decreases less rapidly than that of the surroundings, the densities soon will be equalized, there will be no buoyancy force, and the layer will be stable.

The condition for occurrence of convection in a star can be expressed in another form. If the temperature gradient in a layer, calculated under the assumption of radiation transfer, is steeper than the adiabatic temperature gradient, then convection will occur:

$$|dT/dr| = 3\kappa\rho F_r(4acT^3)^{-1} > |dT/dr|_{\rm ad}. \qquad (31)$$

A layer with high opacity, or opacity rapidly increasing inward, is therefore likely to be unstable; this situation can occur near the surface layers of cool stars. A layer with a high rate of nuclear energy generation in a small volume is also likely to be unstable because of high F_r; this situation is likely to occur in the cores of stars more massive than the sun, where the temperature-sensitive CNO cycle operates.

Convection involves complicated, turbulent motions with continuous formation and dissolution of elements of all sizes. The existence of convection zones is important in the calculation of stellar structure because the overturning of material in almost all phases of evolution is much more rapid than the evolution time, so that the entire zone can be assumed to be mixed to a uniform chemical composition at all times. It is also important for heat transport, which is accomplished mainly by the largest elements. No satisfactory analytic theory of convection exists, although three-dimensional numerical hydrodynamical simulations, for example, of the outer layers of the sun, have been able to attain spatial resolution high enough so that they can reproduce many observed features. For long-term stellar evolution calculations, a very simplified "mixing length" theory is employed. It is assumed that a convective element forms, travels one mixing length vertically, and then dissolves and releases its excess energy to the surroundings. The mixing length is approximated by αH, where α is a parameter of order unity and H is the distance over which the pressure drops by a factor e. By use of this theory the average velocity of an element, the excess thermal energy, and the convective flux can be estimated. In most situations in a stellar interior the estimates show that it is adequate to assume that $dT/dr = (dT/dr)_{\rm ad}$. Although an excess in dT/dr over the adiabatic value is required for convection to exist, convection is efficient enough to carry the required energy flux even if this excess is negligibly small. There is a small uncertainty in stellar models because of possible overshooting, and resulting mixing, of convective elements beyond the boundary of a formal convection zone; otherwise, the uncertainty in the use of the mixing length theory is limited to the surface layers of

stars, where in fact the actual $|dT/dr|$ in a star can be significantly greater than $|dT/dr|_{\rm ad}$. The parameter α can be calibrated by comparison of calculated mixing lengths with the size of the granular elements in the solar photosphere that represent solar convective elements. Also, the radius of a theoretical model of the sun, which is sensitive to α, can be compared with and adjusted to the observed radius. In this manner it has been determined that the parameter α should be in the range 1–2. Calibration can be also made for stars in other phases of evolution by the use of numerical hydrodynamical models; however, it is often assumed that stars evolve with the same α that has been obtained for the sun.

Conduction of heat by the ions and electrons is, in general, inefficient in stellar interiors because the density is high enough that the mean free path of the particles is small compared with the photon mean free path. In the interiors of white dwarfs and the evolved cores of red giants, however, the electrons are highly degenerate, their mean free paths are very long, and conduction becomes a more efficient mechanism than radiative transfer. The flux is related to the temperature gradient by $F_r = -K\,dT/dr$, where K, the conductivity, is calculated by a complicated theory involving the velocity, collision cross section, mean free path, and energy carried by the electrons. Numerical tables are constructed for use in stellar structure calculations.

E. Basic Equilibrium Conditions

During most phases of evolution, a star is in hydrostatic equilibrium, which means that the force of gravity on a mass element is exactly balanced by the difference in pressure on its upper and lower surface. In spherical symmetry, an excellent approximation for most stars, this condition is written

$$\frac{\partial P}{\partial M_r} = -\frac{GM_r}{4\pi r^4}, \tag{32}$$

where M_r is the mass within radius r, ρ is the density, and P is the pressure. A second equilibrium condition is the mass conservation equation in a spherical shell, which simply states that

$$\frac{\partial r}{\partial M_r} = \frac{1}{4\pi r^2 \rho}. \tag{33}$$

The third condition is that of thermal equilibrium, which refers to the situation where the star is producing enough energy by nuclear processes to exactly balance the loss of energy by radiation at the surface. Then,

$$L = \int_0^M \epsilon\, dM_r, \tag{34}$$

where M is the total mass and ϵ is the nuclear energy generation rate per unit mass after subtraction of neutrino losses. This condition holds on the main sequence, but in many stages of stellar evolution it does not, because the star may be expanding, contracting, heating, or cooling. The more general equation of conservation of energy, effectively the first law of thermodynamics, must then be used:

$$\frac{\partial L_r}{\partial M_r} = \epsilon - \frac{\partial E}{\partial t} - P\frac{\partial V}{\partial t}, \tag{35}$$

where $V = 1/\rho$; E is the internal energy per unit mass; and $L_r = 4\pi r^2 F_r$, the total amount of energy per unit time crossing a spherical surface at radius r. If $\epsilon = 0$, the star contracts and obtains its energy from the third term on the right-hand side.

To obtain a detailed solution for the structure and evolution of a star, one must solve Eqs. (32), (33), and (35) along with a fourth differential equation which describes the energy transport. Equation (29) for radiation transport can be re-written, with M_r as the independent variable, in a form which can be used for all three types of transport:

$$\frac{\partial T}{\partial M_r} = -\frac{GM_rT}{4\pi r^4 P}\nabla, \tag{36}$$

where, if the energy transport is by radiation,

$$\nabla = \nabla_{\rm rad} = \frac{3}{16\pi Gac}\frac{\kappa L_r P}{M_r T^4} = \left(\frac{\partial \log T}{\partial \log P}\right)_{\rm rad}. \tag{37}$$

If transport is by conduction, an equivalent "conductive opacity" κ_{cond} can be obtained from the conductivity K. However, each point in the star must be tested to see if the condition for convection is satisfied, and if so, ∇ in Eq. (36) is replaced by the adiabatic gradient $\nabla_{\rm ad} = (\partial \log T/\partial \log P)_{\rm ad}$ in the interior or by a "true" ∇ obtained from mixing-length theory in the surface layers. The four differential equations are supplemented by expressions for the equation of state (see Section III.B), nuclear energy generation (see Section III.C), and opacity κ (see Section III.D) as functions of ρ, T, and chemical composition.

Four boundary conditions must be specified. At the center the conditions $r = 0$ and $L_r = 0$ at $M_r = 0$ are applied. At the surface ($M_r = M$) the photospheric boundary conditions can be used for approximate calculations:

$$L = 4\pi R^2 \sigma T_{\rm eff}^4 \quad \text{and} \quad \kappa P = \frac{2}{3}g, \tag{38}$$

where g is the surface gravity. For detailed comparison with observations, however, model atmospheres are used as surface boundary conditions and are joined to the interior calculation a small distance below the photosphere.

To start an evolutionary calculation, one specifies an initial model by giving its total mass M and the distribution of

chemical composition with mass fraction. A typical starting point is the pre-main-sequence phase where the composition is uniform. A sequence of models is calculated, separated by small intervals of time. If nuclear burning is important, the change of composition caused by reactions (e.g., conversion of H to He) is calculated at each layer. The system of equations is solved numerically, generally by a method developed by L. G. Henyey, in which first-order corrections to a trial solution are determined simultaneously for all variables at all grid points; the process is repeated until the corrections become small.

A final important equilibrium condition is the virial theorem, which, for a non-rotating, non-magnetic spherical star in hydrostatic equilibrium, can be written

$$2T_i + W = 0. \tag{39}$$

Here, W is the gravitational energy, $-qGM^2/R$, where R is the total radius and q is a constant of order unity that depends on the mass distribution, and T_i is the total internal energy $\int_0^M E dM_r$. This expression can be used to estimate stellar internal temperatures for the case of an ideal gas. Then $T_i = 1.5(R_g/\mu)TM$, where T is the mean temperature, and $T \approx GM\mu/(3RR_g) \approx 5 \times 10^6$ K for the case of the sun. This expression also shows that a contracting, ideal-gas star without nuclear sources heats up, with T increasing approximately as $1/R$.

IV. STELLAR EVOLUTION BEFORE THE MAIN SEQUENCE

Early stellar evolution can be divided into three stages : star formation, protostar collapse, and slow (Kelvin-Helmholtz) contraction. The characteristics of these three stages are summarized in Table II. As the table indicates, there is a vast difference in physical conditions between the time of star formation in an interstellar cloud and the time when a star arrives on the main sequence and begins to burn hydrogen. An increase in mean density by 22 orders of magnitude and in mean temperature by 6 orders of magnitude occurs. In the star formation and protostar collapse stages, it is not sufficient to assume that the object is spherical, and a considerable variety of physical processes must be considered. The problem involves solution of the equations of hydrodynamics, including rotation, magnetic fields, turbulence, molecular chemistry, gravitational collapse, and radiative transport of energy. Some two- and three- dimensional hydrodynamic simulations have been performed for the star formation phase and for the protostar collapse phase, with the aim of investigating the following, as yet unsolved, problems: (1) What initiates collapse? (2) What is the rate and efficiency of the conversion of interstellar matter into stars? (3) What determines the distribution of stars according to mass? (4) What is the mechanism for binary formation, and what determines whether a star will become a member of a double or multiple star system or a single star?

A. Star Formation

It is clear that star formation is limited to regions of unusual physical conditions compared with those of the interstellar medium on the average. The only long-range attractive force to form condensed objects is the gravitational force. However, a number of effects oppose gravity and prevent contraction, including gas pressure, turbulence, rotation, and the magnetic field. The chemical composition of the gas, the degree of ionization or dissociation, the presence or absence of grains (condensed particles of ice, silicates, carbon, or iron with characteristic size 5×10^{-5} cm), and the heating and cooling mechanisms in the interstellar medium all have an important influence on star formation. The regions where star formation is observed to occur is in the molecular clouds, where they can form in a relatively isolated mode, as in the Taurus region or, more commonly, in clusters (Fig. 6).

We consider first only the effects of thermal gas pressure and gravity in an idealized spherical cloud with uniform density ρ, uniform temperature T, and mass M. The self-gravitational energy is $W = -0.6GM^2/R$, where R is the radius. The internal energy is $T_i = 1.5R_gTM/\mu$, where μ is the mean molecular weight per particle. Collapse will be

TABLE II Major Stages of Early Stellar Evolution[a]

Phase	Size (cm)	Observations	Density (g cm^{-3})	Internal temperature (K)	Time (year)
Star formation	10^{20}–10^{17}	Radio	10^{-22}–10^{-19}	10	10^7
Protostellar collapse	10^{17}–10^{12}	Infrared	10^{-19}–10^{-3}	10–10^6	10^6
Slow contraction	10^{12}–10^{11}	Optical	10^{-3}–1.0	10^6–10^7	4×10^{7b}

[a] Reproduced with permission from Bodenheimer, P. (1983). "Protostar collapse." *Lect. Appl. Math.* **20,** 141. ©American Mathematical Society.

[b] For 1 solar mass.

FIGURE 6 The Trifid Nebula in Sagittarius (M20, NGC 6514), a region of recent star formation (Shane 120-in. reflector). (Lick Observatory photograph.)

possible roughly when $|W| > T_i$; that is, when the cloud is gravitationally bound, a criterion that has been verified by numerical calculations. Thus, for collapse to occur the radius must be less than $R_J = 0.4GM\mu/(R_g T)$, known as the Jeans length. By eliminating the radius in favor of the density, we obtain the Jeans mass, which is the minimum mass a cloud with density ρ and temperature T must have for collapse to occur:

$$M_J = \left(\frac{2.5R_g T}{\mu G}\right)^{3/2}\left(\frac{4}{3}\pi\rho\right)^{-1/2}$$

$$= 8.5 \times 10^{22}\left(\frac{T}{\mu}\right)^{3/2}\rho^{-1/2}\text{g}. \qquad (40)$$

This condition turns out to be quite restrictive. A typical interstellar cloud of neutral H has $T = 50\,\text{K}$, $\rho \approx 1.7 \times 10^{-23}$ g cm^{-3}, and $\mu \approx 1$, and the corresponding $M_J = 3600\ M_\odot$. The actual masses are far below this value, so the clouds are not gravitationally bound. On the other hand, typical conditions in an observed molecular cloud are $T = 10\,\text{K}$, $\rho \approx 1.7 \times 10^{-21}$ g cm^{-3}, and $\mu \approx 2$, with $M_J \approx 10\ M_\odot$. Fragmentation into stellar mass objects seems quite possible.

However, this analysis must be modified to take into account rotation and magnetic fields. The angular momentum problem can be stated in the following way. The angular momentum per unit mass of a typical molecular cloud with mass $10^4\ M_\odot$ and a radius of a few parsecs is 10^{24} cm^2 sec^{-1}. The rotational velocities of young stars that have just started their pre-main-sequence contraction indicate that their $J/M \approx 10^{17}$ cm^2 sec^{-1}. Evidently, some fragments of dark cloud material must lose 7 orders of magnitude in J/M before they reach the stellar state. The angular momentum reduction may occur at several different evolutionary stages and by different physical processes. During the star formation stage, it has been proposed that if the matter is closely coupled to the magnetic field, the twisting of the field lines can generate Alfvén waves that would transfer angular momentum from inside the cloud to the external medium. It has been estimated that significant reduction of angular momentum could take place in 10^6–10^7 years. Once the density increases to about 10^{-19} g cm^{-3}, the influence of the magnetic force on the bulk of the matter becomes relatively small, because the density of charged particles becomes very low. However, observations of rotational motion at this density indicate that J/M has already been reduced to $\approx 10^{21}$ cm^2 sec^{-1}. These dense regions, known as *molecular cloud cores*, have sizes of about 0.1 pc and temperatures of about 10 K; they are observed in the radio region of the spectrum, often through transitions of the ammonia molecule. They are the regions where the onset of protostellar collapse is likely to occur.

However, the magnetic force also opposes collapse, and a magnetic Jeans mass can be estimated by equating the magnetic energy of a cloud with its gravitational energy:

$$M_{J,m} \approx 10^3\ M_\odot\left(\frac{B}{30\ \mu G}\right)\left(\frac{R}{2\ \text{pc}}\right)^2, \qquad (41)$$

where B is the magnetic field. In typical molecular cloud regions, with $\rho \approx 10^{-21}$ g cm^{-3}, $R \approx 2$ pc, and observed $B \approx 30\,\mu$G, magnetic effects are seen to be much more significant than thermal effects in preventing collapse. However, in the cores, the observed field, although not accurately determined, is still $\approx 30\,\mu$G and the radius is only 0.1 pc, giving a value of $M_{J,m} \approx 2.5\ M_\odot$, while $M_J \approx 1\ M_\odot$. The cores actually have masses of a few solar masses, so, again, they are likely regions to collapse. The value of B does not increase in the cores as compared with the average cloud material because during the compression the predominantly neutral matter is able to drift relative to the field lines.

There are actually two scenarios which could explain how collapse is initiated. The first is based on the discussion above: molecular cloud material gradually diffuses with respect to magnetic field lines until a situation is reached, in a molecular cloud core, where the core mass is about the same as both the thermal and the magnetic Jeans masses. Rotation is not important at this stage because of magnetic braking, and collapse can commence. Of course, if angular momentum is conserved, rotational effects will soon become very important as collapse proceeds. The second picture, known as *induced* star formation, relies

on a more impulsive event, such as the collision between two clouds or the sweeping of a supernova shock across a cloud, to initiate collapse. The resulting compression can, under the right conditions, lead to more rapid diffusion of the magnetic field, more rapid decay of turbulence, and enhanced cooling of the cloud, thereby reducing the magnetic, turbulent, and thermal energies, respectively, and increasing the likelihood of collapse.

B. Protostar Collapse

Once collapse starts, the subsequent evolution can be divided into an isothermal phase, an adiabatic phase, and an accretion phase. A typical initial condition is a cloud core with radius 2×10^{17} cm, $T = 10$ K, $M = 2\ M_\odot$, and mean density $\approx 1 \times 10^{-19}$ g cm^{-3} (cloud cores are actually observed to be somewhat centrally condensed). During the isothermal phase, the cloud is optically thin to its own infrared radiation. It tends to heat by gravitational compression, but the cooling radiation generated by gas-grain collisions rapidly escapes the cloud, and the temperature remains at about 10 K. The pressure cannot increase rapidly enough to bring the cloud to hydrostatic equilibrium, and instability to collapse continues. Numerical simulations show that even if the cloud initially has uniform density, a density gradient is soon set up the propagation of a rarefaction wave inward from the outer boundary; soon the denser central regions are collapsing much faster than the outer regions, and the protostar becomes highly condensed toward the center with a density distribution close to $\rho \propto r^{-2}$. During this process the thermal Jeans mass (Eq. 40) rapidly decreases with the increase in density. Thus, in principle, fragmentation of the cloud into smaller masses can occur. Calculations that include rotation show that the cloud tends to flatten into a disklike structure, and once it has become sufficiently flat, the central region of the cloud can break up into orbiting subcondensations with orbital specific angular momentums of 10^{20}–10^{21} cm^{-2} sec^{-1}, comparable to those of the wider binary systems (Fig. 7). That fragmentation during the protostar collapse is a reasonable mechanism for binary formation is supported by the observational evidence that many, if not most, very young stars are already in binary systems at the early phases of the pre-main-sequence contraction, with binary frequency similar to that observed for stars on the main sequence. A few binary protostars have also been detected. In the fragmentation process, each of the fragments still retains some angular momentum of spin and is unstable to collapse on its own, but most of the spin angular momentum of the original cloud goes into the orbital motion of the binary; the process thus provides another mechanism for solving the angular momentum problem.

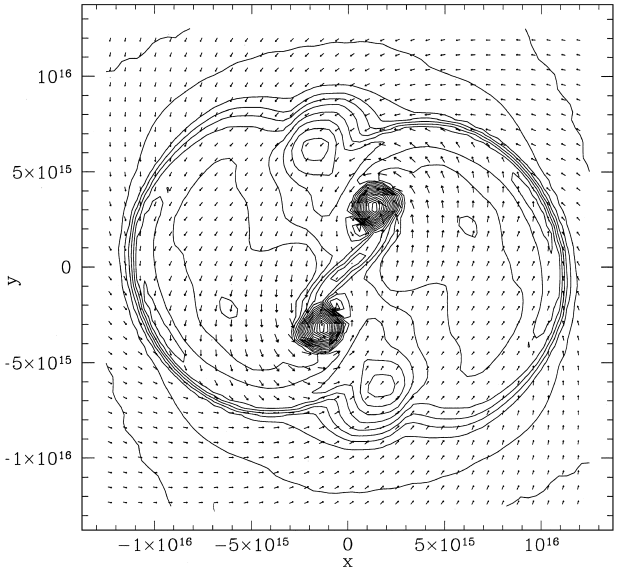


FIGURE 7 Numerical simulation of binary formation in a collapsing, rotating cloud. Contours of equal density are shown in the equatorial (x, y) plane of the inner part of the cloud. The maximum density in the fragments is 10^{-12} g cm^{-3}, and the minimum density at the outer edge is 10^{-18} g cm^{-3}. Velocity vectors have length proportional to speed, with a maximum value of 2.8×10^5 cm sec^{-1}. The forming binary has a separation of about 600 AU. The length scales are given in centimeters.

Once the density in the center of the collapsing cloud reaches about 10^{-13}g cm^{-3}, that region is no longer transparent to the emitted radiation, and much of the released gravitational energy is trapped and heats the cloud, marking the beginning of the adiabatic phase. By this time the Jeans mass (Eq. 40) has decreased to about 0.01 $M_\odot$. The heating results in an increase in the Jeans mass above this minimum value. The collapse time is still close to the free-fall time, and, as the density increases further, this time becomes short compared with the time required for radiation of diffuse out of the central regions. The collapse then becomes nearly adiabatic, and the pressure increases relative to gravity quite rapidly, so that a small region near the center approaches hydrostatic equilibrium. As this region continues to compress, the temperature rises to 2000 K. At that point the molecular hydrogen dissociates and thereby absorbs a considerable fraction of the released gravitational energy. As a result, the center of the cloud becomes unstable to further gravitational collapse, which continues until most of the molecules have dissociated at a density of 10^{-2} g cm^{-3} and a temperature of 3×10^4 K. The collapse is again decelerated and a core forms in hydrostatic equilibrium. Initially, this core contains only $\approx$1% of the protostellar mass; it serves as the nucleus of the forming star.

Even if the cloud had previously fragmented, the fragments will collapse through the adiabatic phase. Once

the final core has formed, the accretion phase begins. At first, infalling material has relatively little angular momentum, and it joins the core, falling through a shock front at its outer edge. However, material arriving later will have higher and higher angular momentum, so after some time it will not be able to join the core but will start forming a flattened disk surrounding the core, which can grow to a size of several hundred astronomical units. Thus, the protostar, at a typical stage of its development, consists of a slowly rotating central star, a surrounding disk, and an opaque envelope of collapsing material which accretes primarily onto the disk. Once the disk becomes comparable in mass to the central core, it can become gravitationally unstable. In some situations the instability could possibly result in the formation of subcondensations, with masses in the brown dwarf range, in orbit around the core. In other cases, the instability would produce spiral density waves, which would result in transport of angular momentum outward and mass inward, with the result that disk material collects onto the core and allows it to grow in mass. Once the disk mass decreases to less than 10% of the mass of the core, it becomes gravitationally stable. However, observations of young objects show that their disks, which are in the mass range of 1–10% that of the star, still continue to accrete onto the star. The mechanism by which they do so is not fully understood, and hydrodynamic as well as hydromagnetic instabilities are under investigation. Associated with the accretion of mass onto the star is the generation of jets and outflows, generally perpendicular to the plane of the disk, observed in most protostars, and originating through a complicated interaction between rotation, accretion, and magnetic effects near the interface between star and disk.

As material accretes onto the star, either directly or through a disk, it liberates energy at a rate given approximately by $L = GM_c\dot{M}/R_c$, where M_c and R_c are the core mass and radius, respectively, and $\dot{M}$ is the mass accretion rate. This energy is transmitted radiatively through the infalling envelope, where it is absorbed and re-radiated in the infrared. The typical observed protostar has a spectral energy distribution peaking at 60–100 μm, although the peak wavelength and the intensity of the radiation will depend on the viewing angle with respect to the rotation axis. Protostars viewed nearly pole-on will have bluer and more intense radiation than those viewed nearly in the plane of the disk. The time for the completion of protostellar evolution is essentially the free-fall time of the outer layers, which lies in the range 10^5–10^6 years, depending on the initial density. As the accretion rate slows down in the later phases and the infalling envelope becomes less and less opaque, the observable surface of the protostar declines in luminosity and increases in temperature. Once the infalling envelope becomes transparent, the observer can see through to the core, which by now can be identified in the H–R diagram as a star with a photospheric spectrum. The locus in the diagram connecting the points where stars of various masses first make their appearance is known as the *birth line*.

C. Pre-main-Sequence Contraction

Once internal temperatures are high enough (above 10^5 K) that hydrogen is substantially ionized, the star is able to reach an equilibrium state with the pressure of an ideal gas supporting it against gravity. Rotation, which was very important during the protostar stage, has now become negligible. The star radiates from the surface, and this energy is supplied by gravitational contraction, which can be regarded as a passage through a series of equilibrium states. The virial theorem shows that half of the released gravitational energy goes into radiation and the other half into heating of the interior, as long as the equation of state remains ideal. The calculation of the evolution can now be accomplished by solution of the standard structure equations (32), (33), (35), and (36).

The solutions are shown in the H–R diagram in Fig. 8. These tracks start at the birth line for each mass. For earlier times, a hydrostatic stellar-like core may exist, but it is hidden from view by the opaque infalling protostellar envelope. Note that the sun arrives on the H–R diagram

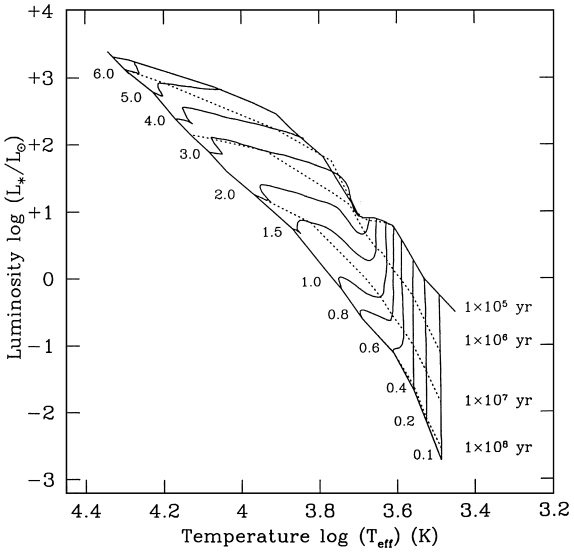


FIGURE 8 Evolutionary tracks for various masses during the pre-main-sequence contraction. Each track is labeled by the corresponding stellar mass (in solar masses, $M_\odot$). For each track, the evolution starts at the birth line (upper solid line) and ends at the zero-age main sequence (lower solid line). Loci of constant age are indicated by dotted lines. [Reproduced by permission from Palla, F., and Stahler, S. (1999). *Astrophys. J.* **525,** 772. © The American Astronomical Society.]

with about five times its present luminosity, and the higher mass stars actually do not appear until they have already reached the main sequence. The results of the calculations show in general that the stars first pass through a convective phase (vertical portions of the tracks) and later a radiative phase (relatively horizontal portions of the tracks). The relative importance of the two phases depends on the stellar mass. During the convective phase, energy transport in the interior is quite efficient, and the rate of energy loss is controlled by the thin radiative layer right at the stellar surface. The opacity is a very strongly increasing function of T in that layer, and that fact combined with the photospheric boundary conditions (38) can be shown to result in a nearly constant $T_{\rm eff}$ during contraction. As the surface area decreases, L drops and $T_{\rm eff}$ stays between 2000 and 4000 K, with lower masses having lower $T_{\rm eff}$. As the star contracts, the interior temperatures increase and in most of the star the opacity decreases as a function of T. The star gradually becomes stable against convection, starting at the center, because the radiative gradient [Eq. (37)] drops along with the opacity and soon falls below $\nabla_{\rm ad}$. When the radiative region includes about 75% of the mass, the rate of energy release is no longer controlled by the surface layer, but rather by the opacity of the entire radiative region. At this time, the tracks make a sharp bend to the left, and the luminosity increases gradually as the average interior opacity decreases.

Contraction times to the main sequence, starting at the birth line, for various masses are given in Table III. A summary of the evolution of the sun during the pre-main-sequence, main-sequence, and early post-main-sequence phases is given in Fig. 9. The higher mass stars have relatively high internal temperature and therefore relatively low internal opacities and are able to radiate rapidly. The contraction times are short, and because the luminosity is relatively constant during the radiative phase, during which they spend most of their time, the contraction time is well approximated by the

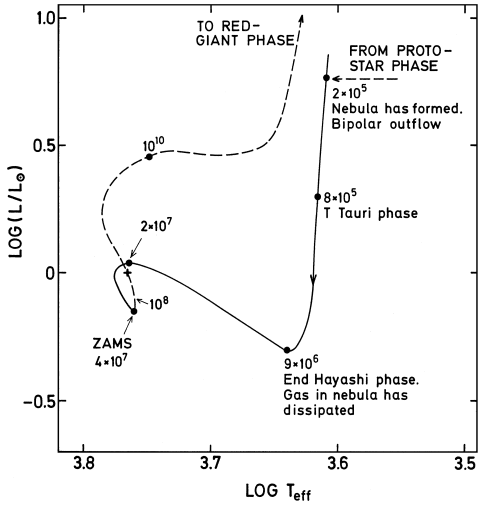


FIGURE 9 Sketch of the pre-main-sequence (solid line) and post-main-sequence (dashed line) evolution of a star of 1 $M_\odot$. Evolutionary times (years) are given for the filled circles. The present sun is indicated by a cross. The term "nebula" refers to the circumstellar disk. [Adapted with permission from Bodenheimer, P. (1989). In "The Formation and Evolution of Planetary Systems" (H. Weaver and L. Danly, eds.), p. 243, Cambridge Univ. Press, Cambridge, UK. Reprinted with permission of Cambridge University Press.]

Kelvin-Helmholtz time $t_{\rm KH} \approx GM^2/(RL)$. A star of 1 $M_\odot$ spends 10^7 years, on the vertical track, known as the Hayashi track after the Japanese astrophysicist who discovered it. He also pointed out that the region to the right of the Hayashi track is "forbidden" for a star of a given mass in hydrostatic equilibrium, at any phase of evolution. For the next 2×10^7 years, the star is primarily radiative, but it maintains a thin outer convective envelope all the way to the main sequence. The final 10^7 year of the contraction phase represents the transition to the main sequence, during which the nuclear reactions begin to be important at the center, the contraction slows down, and, as the energy source becomes more concentrated toward

TABLE III Evolutionary Times (Year)

Mass ($M_\odot$)	Pre-main-sequence contraction	Main-sequence lifetime	To onset of core He burning	Core He lifetime
30	—	4.8×10^6	1.0×10^5	5.4×10^5
15	—	1.0×10^7	3.0×10^5	1.6×10^6
9	—	2.1×10^7	9.1×10^5	3.8×10^6
5	2.3×10^5	6.5×10^7	4.8×10^6	1.6×10^7
3	2.0×10^6	2.2×10^8	2.9×10^7	6.6×10^7
2	8.5×10^6	8.4×10^8	3.1×10^8	—
1	3.2×10^7	9.2×10^9	2.5×10^9	1.1×10^8
0.5	1.0×10^8	1.5×10^{11}	—	—
0.3	1.8×10^8	7.1×10^{11}	—	—
0.1	3.7×10^8	5.9×10^{12}	—	—

the center, the luminosity declines slightly. For the lower mass stars, the evolution is entirely along the Hayashi track. Stars of 0.3 $M_\odot$ or less remain fully convective all the way to the main sequence. Because the luminosity varies continuously during the contraction, t_{KH} is determined by integration along the track. It takes a star of 0.1 $M_\odot$ about 4×10^8 years to reach the main sequence.

A number of comparisons can be made between the contraction tracks and the observations.

1. As predicted by Hayashi, no stars in equilibrium are observed to exist in the forbidden region of the H–R diagram.
2. The lithium abundances of stars that have just reached the main sequence can be compared with calculations of the depletion of lithium in the surface layers during the contraction. Lithium is easily destroyed by reactions with protons at temperatures above about $T = 2.8 \times 10^6$ K. The lower mass stars have outer convection zones extending down to this T during the contraction; the higher mass stars do not. Because the material in the convection zone is completely mixed, the low-mass stars would be expected on the average to have less surface (observable) Li when they arrive at the main sequence, in agreement with observations.
3. The youngest stars, known as the T Tauri stars, have high lithium abundances; are associated with dark clouds where star formation is taking place; display irregular variability in light, as well as mass loss that is much more rapid than that of most main-sequence stars; and have excess infrared emission over that expected from a normal photosphere, indicating the presence of a surrounding disk, and excess ultraviolet radiation, indicating accretion onto the star. All these characteristics are consistent with youth; their location in the H–R diagram falls along Hayashi tracks for stars in the mass range 0.5–2.0 $M_\odot$. The disk characteristics vanish after ages of a few million to 10^7 years, putting a constraint on the time available for the formation of gaseous giant planets.
4. Dynamical masses obtained from pre-main-sequence eclipsing binaries are an important tool for calibration of the pre-main-sequence theoretical tracks.
5. The H–R diagrams for young stellar clusters can be compared with lines of constant age (Fig. 8). The results show that there is a considerable scatter of observed points about a single such line, indicating that star formation in a given region probably occurs continuously over a period of about 10^7 years.

V. THE MAIN SEQUENCE

A. General Properties

The zero-age main sequence (ZAMS) is defined as the time when nuclear reactions first provide the entire luminosity radiated by a star. The chemical composition is assumed to be spatially homogeneous at this time, although actually in the centers of higher mass stars, some ^{12}C is converted to ^{14}N by Reactions (10)–(12) before that time. Later, the nuclear reactions result in the conversion of hydrogen to helium, with the change occurring most rapidly at the center. The characteristics of a range of stellar masses on the ZAMS are given in Table IV; the assumed composition is $X = 0.71$, $Y = 0.27$, $Z = 0.02$ (subscripts c refer to central values). The following general points can be made.

1. The equation of state in the interior is close to that of an ideal gas. There are deviations only at the high-mass end, where radiation pressure becomes important, and at the low-mass end, where electron degeneracy begins to be significant.
2. The radius increases roughly linearly with mass at the low-mass end and roughly as the square root of the mass above 1 $M_\odot$.
3. The theoretical mass–luminosity relation agrees well with observations (Fig. 2); the location of the theoretical ZAMS in the H–R diagram is also in good agreement with the observations.

TABLE IV Zero-Age Main Sequence

Mass ($M_\odot$)	log $L/L_\odot$	Log T_{eff} (K)	$R/R_\odot$	T_c (10^6 K)	ρ_c (g cm^{-3})	M_{core}/M	M_{env}/M	Energy source
30	5.15	4.64	6.6	36	3.0	0.60	0	CNO
15	4.32	4.51	4.7	34	6.2	0.39	0	CNO
9	3.65	4.41	3.5	31	10.5	0.30	0	CNO
5	2.80	4.29	2.3	27	17.5	0.23	0	CNO
3	2.00	4.14	1.7	24	40.4	0.18	0	CNO
2	1.30	4.01	1.4	21	68	0.12	0	CNO
1	−0.13	3.76	0.9	14	90	0	0.02	PP
0.5	−1.44	3.56	0.44	9	74	0	0.4	PP
0.3	−2.0	3.53	0.3	7.5	125	0	1.0	PP
0.1	−3.3	3.43	0.1	5	690	0	1.0	PP

4. The central temperature increases with mass, as expected from the virial theorem. The higher mass stars burn hydrogen on the CNO cycle, and because of the steep T dependence of these reactions and the corresponding strong degree of concentration of the energy sources to the center, they have convective cores. The fractional mass of these cores increases with total mass. Comparisons with cluster H–R diagrams suggests that a small amount of convective overshooting occurs at the edge of these cores. The lower mass stars run on the pp chains, and because of the more gradual T dependence, they do not have convective cores. However, because they have cooler surface layers and therefore steeply increasing opacity going inward from the surface, they have convection zones in their envelopes, which become deeper as the mass decreases. At 0.3 $M_\odot$, this convection zone extends all the way to the center. The lowest mass stars do not have high enough T_c to allow ^{3}He to react with itself [Reaction (3)]; therefore (at least on the ZAMS), the pp chain proceeds only through Reactions (1) and (2).

5. There is no substantial evidence for stars above 100 $M_\odot$. The upper limit for stellar masses arises either from the fact that radiation pressure from the core can prevent further accretion of envelope material during the protostellar phase when the core reaches some critical mass or from the fact that main-sequence stars above a given mass are pulsationally unstable which, in combination with strong radiation pressure, tends to result in rapid mass loss from their surfaces.

6. The lower end of the main sequence occurs at about 0.075 $M_\odot$. Below that mass nuclear reactions cannot provide sufficient energy. Physically, the limit arises from the fact that as low-mass objects contract, they approach the regime of electron degeneracy before T_c becomes high enough to start nuclear burning. As the electrons are forced into higher and higher energy states, much of the gravitational energy supply of the star is primarily used to provide the required energy to the electrons. The ions, whose temperatures determine nuclear reaction rates, are still and ideal gas, but they reach a maximum temperature and then begin to cool. Their thermal energy is required, along with the gravitational energy, to supply the radiated luminosity. Substellar objects below the limit simply contract to a limiting radius and then cool. Extensive observational searches for such objects, known as "brown dwarfs," have resulted in the discovery of many strong candidates.

B. Evolution on the Main Sequence

During the burning of H in their cores, stars move very slowly away from the ZAMS. The main-sequence lifetimes, up to exhaustion of H at the center, are given in Table III. High-mass stars go through this phase in a few million years, while stars below about 0.8 $M_\odot$ (with solar composition) have not had time since the formation of the galaxy to evolve away from the main sequence. The rate at which energy is lost at the surface, which determines the evolutionary time scale, is controlled by the radiative opacity through much of the interior. The stellar structure adjusts itself to maintain thermal equilibrium, which means that the energy production rate is exactly matched by the rate at which energy can be carried away by radiation. If for some reason the nuclear processes were producing energy too rapidly, some of this energy would be deposited as work in the inner layers, these regions would expand and cool, and the strongly temperature-dependent energy production rate would return to the equilibrium value.

The structure of the star on the main sequence changes during main-sequence evolution as a consequence of the change in composition. In the case of upper main-sequence stars the H is depleted uniformly over the entire mass of the convective core, while in the lower mass stars, which are radiative in the core, the H is depleted most rapidly in the center. The conversion of H to He results in a slight loss of pressure because of a decease in the number of free particles per unit volume; as a result, the inner regions contract very slowly to maintain hydrostatic equilibrium. The conversion also reduces the opacity somewhat, which tends to cause a slow increase in L. The outer layers see no appreciable change in opacity; a small amount of the energy received from the core is deposited there in the form of work, and these layers gradually expand. For example, in the case of the sun, L has increased from 0.7 to 1.0 $L_\odot$ since age zero, R has increased from 0.9 to 1.0 $R_\odot$, T_c has increased from 14×10^6 to 15.7×10^6 K, and ρ_c has increased from 90 to 152 g cm^{-3}. In the case of high-mass, stars when X goes to zero at the center, it does so over the entire mass of the convective core, and the star is suddenly left without fuel. A rapid overall contraction takes place, until the layers of unburned H outside the core reach temperatures high enough to burn. For solar-mass stars, the main-sequence evolution does not end so suddenly, because the H is depleted only one layer at a time, and the transition to the red giant phase takes place relatively slowly.

An important calibration for the entire theory of stellar evolution is the match between theory and observation of the sun. The procedure is to choose a composition and to evolve the model sun from age zero to the present solar age of 4.56×10^9 years . The value of Z is well constrained by observations of photospheric abundances; thus, if the solar L does not match the model L, the abundance of He in the model can be used as a parameter to be adjusted to provide agreement. Once L of the model is satisfactory, the radius can be adjusted to match the observed R by varying the mixing-length parameter α that is used to calculate

the structure of the outer part of the convection zone. A further check is provided by the results of helioseismology, a method of probing the interior structure of the sun by making precise measurements of the frequencies of its oscillations at the surface. The best-known periods are of the order of 5 min. The technique is able to measure the sound speed as a function of depth far into the sun, as well as the location of the bottom of the convection zone. Models that start with the composition consistent with that of Table I ($X = 0.70$, $Y = 0.28$, $Z = 0.02$) and $\alpha = 1.8$, and which include the slow settling of He downward during the evolution, provide excellent agreement with the results of helioseismology and the observed Z, L, and R. The present sun (see Table V) has $Y = 0.24$ in the outer layers; the convection zone includes the outer 29% of the radius with $T = 2.2 \times 10^6$ K at its lower edge.

C. The Solar Neutrino Problem

The results of terrestrial solar neutrino detection experiments can be summarized as follows:

1. The chlorine experiment, $^{37}\mathrm{Cl} + \nu \rightarrow {}^{37}\mathrm{Ar} + e^-$ with detection threshold 0.81 MeV, detects primarily neutrinos produced by ^{8}B decay [Reaction (8)]. A long series of experiments going back to 1970 shows a detection rate only one-third that predicted by the standard solar model.
2. The neutrino-electron scattering experiment carried out by the Superkamiokande project has a high detection threshhold (about 6 MeV) and can detect only neutrinos from Reaction (8). Results show about half the expected detection rats, but the experiment can determine from which direction the neutrinos arrive, and it is clear that they come from the sun.
3. Two independent gallium experiments, the SAGE project and the GALLEX project, detect neutrinos by $^{71}\mathrm{Ga} + \nu \rightarrow {}^{71}\mathrm{Ge} + e^-$ with detection threshhold 0.234 MeV. The experiments can detect neutrinos from all three branches of the pp chain; in particular, about half of the neutrinos produced by Reaction (1), which are by far the most numerous of the solar neutrinos, are detectable by this experiment. The actual detection rate in these experiments, which have been calibrated through the use of a terrestrial neutrino source of known production rate, is slightly more than half of the expected rate. The discrepancies in all case are far greater than the experimental or theoretical uncertainties.

The following possibilities have been considered as solutions to the solar neutrino problem.

1. The solar model is inappropriate. The production rate of the neutrinos produced by ^{8}B is very sensitive to the temperature near the center of the sun; those produced

TABLE V Standard Solar Model

$M(r)/M_\odot$	$r/R_\odot$	T (10^6K)	ρ (g cm^{-3})	P (erg cm^{-3})	$L(r)/L_\odot$	X
0.000	0.00	15.7	152	2.33×10^{17}	0.0	0.34
0.009	0.045	15.0	132	2.05×10^{17}	0.073	0.40
0.041	0.078	13.9	104	1.63×10^{17}	0.282	0.48
0.117	0.120	12.3	73.3	1.10×10^{17}	0.602	0.58
0.217	0.159	10.8	51.5	7.17×10^{16}	0.822	0.65
0.308	0.191	9.69	38.3	4.89×10^{16}	0.920	0.68
0.412	0.226	8.60	26.9	3.08×10^{16}	0.971	0.69
0.500	0.255	7.79	19.5	2.03×10^{16}	0.990	0.70
0.607	0.299	6.82	12.2	1.12×10^{16}	0.998	0.71
0.702	0.345	5.96	7.30	5.85×10^{15}	1.000	0.71
0.804	0.411	4.97	3.48	2.32×10^{15}	1.000	0.71
0.902	0.518	3.81	1.13	5.82×10^{14}	1.000	0.72
0.927	0.562	3.42	0.726	3.35×10^{14}	1.00	0.72
0.962	0.656	2.69	0.304	1.11×10^{14}	1.000	0.72
0.976[a]	0.714	2.18	0.185	5.53×10^{13}	1.000	0.74
0.990	0.804	1.33	0.088	1.59×10^{13}	1.000	0.74
0.998	0.900	0.60	0.026	2.10×10^{12}	1.000	0.74
0.9997	0.949	0.28	0.008	3.06×10^{11}	1.000	0.74
1.0000	1.000	0.0058	2.8×10^{-7}	1.20×10^5	1.000	0.74

[a] Base of convection zone.

by the p + p reaction are also sensitive, but less so. A decrease in this temperature in the models would bring the predictions more into line with the experiments. One of the main parameters in the solar model is the abundance of He. A decrease in its assumed value (and replacement by H) results in an increase in the opacity, so that a compensating decrease in Z is required so that the solar luminosity will be matched at the solar age. The value of T_c is reduced because in the presence of the increased value of X, Eq. (9) shows that the energy requirements of the sun can be met with a lower temperature and density. However, the value of Z is well enough known observationally that a change in Y sufficient to bring about agreement would result in a required value of Z that is outside the observational uncertainties. Furthermore, the value of Y is well constrained through helioseismology as a function of radius almost all the way to the center of the sun. The agreement between the observationally determined sound speed as a function of radius and the standard solar model is within 1%. In general, a change in the solar model sufficient to bring about consistency with the solar neutrino experiments will result in inconsistency with at least one other well-observed property of the sun.

2. The nuclear reaction rates, when extrapolated to stellar energies, are incorrect. The experimental rate of Reaction (7) is particularly difficult to determine, and it is directly proportional to the production rate of neutrinos from ^{8}B. However, the quoted uncertainty in all the rates, typically 10%, is not large enough to solve the problem.

3. Through so-called "neutrino oscillations," the electron neutrinos produced in the sun are converted to muon or tau neutrinos before they reach the earth; the detection experiments now running are insensitive to these other types. This possibility is the subject of intensive research by particle physicists. There is some direct experimental evidence that muon neutrinos undergo oscillations, but none so far for electron neutrinos.

4. The sun undergoes periodic episodes of mixing of the chemical composition in the interior. Some hydrogen outside the burning region would be mixed downward, providing extra fuel and increasing the energy generation rate given by Eq. (9). Some energy is deposited in the form of work in the burning layers, they expand and cool to re-establish the balance between L and the nuclear energy generation rate. The cooler sun then produces fewer detectable neutrinos. It is also possible that the sun is continuously mixed by a slow diffusive process; in either case, the mechanism for mixing is not understood, and there is no evidence for it in the helioseismological results. The solution to the solar neutrino experiment, may come about through different experimental techniques. For example, the SNO experiment, which uses heavy water (D_2O) rather than ordinary water as a detecting agent, in principle will be able to determine whether solar neutrinos have undergone oscillations.

VI. STELLAR EVOLUTION BEYOND THE MAIN SEQUENCE

Following the exhaustion of H at the center of a star, major structural adjustments occur; the physical processes that result in the transition to a red giant were first made clear by the calculations of F. Hoyle and M. Schwarzschild. The central regions, and in some cases the entire star, contract until the hydrogen-rich layers outside the exhausted core can be heated up to burning temperatures. Hydrogen burning becomes established in a shell source, which becomes thinner in terms of mass and hotter as the star evolves. The central region, inside the shell, continues to contract, while the outer layers expand considerably as the star evolves to the red giant region. The value of $T_{\rm eff}$ decreases to 3500–4500 K, and a convection zone develops in the outer layers and becomes deeper as the star expands. When T in the core reaches 10^8 K, helium burning begins, resulting in the production of ^{12}C and ^{16}O. Evolutionary time scales up to and during helium burning are given in Table III. Whether further nuclear burning stages occur that result in the production of still heavier elements depends on the mass of the star. For masses less than 9 $M_\odot$, interior temperatures never become high enough to burn the C or O. The star ejects its outer envelope, goes through a planetary nebula phase, and ends its evolution as a white dwarf. The more massive stars burn C and O in their cores, and nucleosynthesis proceeds to the production of an iron-nickel core. Collapse of the core follows, resulting in a supernova explosion, the ejection of a large fraction of the mass, and the production of a neutron star remnant for masses up to about 25 $M_\odot$ and possibly a black hole for higher masses. The details of the evolution are strongly dependent on mass. The following sections describe the post-main-sequence evolution of stars of Population I with masses of 1, 5, and 25 $M_\odot$ and of a Population II star of about 0.8 $M_\odot$, which represents a typical observable star in a globular cluster. Evolutionary tracks for the Population I stars are shown in Fig. 10.

A. Further Evolution of 1 $M_\odot$ Stars

The transition phase to the red giant region, during which the star evolves to the right in the H–R diagram, takes place on a time scale that is short compared with the main-sequence lifetime, but long enough so that stars undergoing the transition should be observable in old clusters. As the convective envelope deepens, the evolutionary track changes direction in the H–R diagram, and the luminosity increases considerably while $T_{\rm eff}$ decreases only slowly.

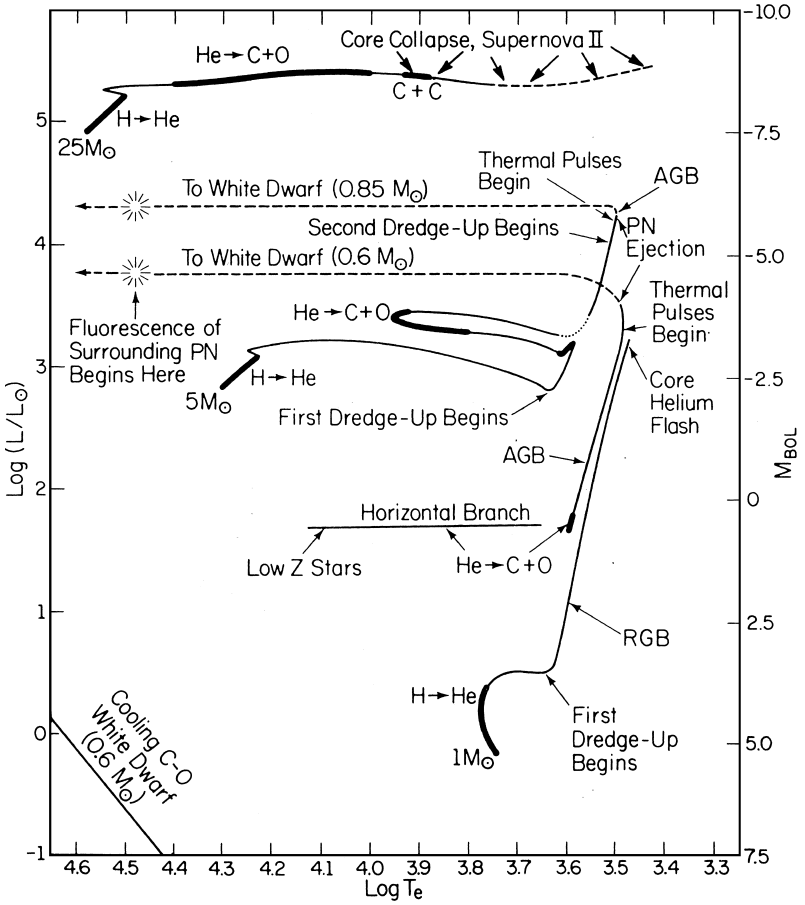


FIGURE 10 Post-main-sequence evolutionary tracks in the H–R diagram for stars of 1.0, 5.0, and 25 $M_\odot$ under the assumption that the metal abundance Z is comparable to that in the present sun. The heavy portions of each curve indicate where major nuclear burning stages occur in the core. The label RGB refers to the red giant branch, AGB refers to the asymptotic giant branch, and PN refers to the planetary nebula phase. Helium burning occurs on the horizontal branch only if Z is much less than the solar value. [Reproduced with permission from Iben, I., Jr. (1991). *Astrophys. J. Suppl.* **76,** 55. © The American Astronomical Society.]

The inner edge of the convection zone moves inward to a point just outside the hydrogen-burning shell; inside the shell is the dense, burned-out core consisting mainly of He and increasing in mass with time. Temperatures in the shell increase to the point where the CNO cycle takes over as the principal energy source. The outer convective envelope becomes deep enough so that it reaches layers in which C has been converted to N by Reactions (10)–(12), which proceed at a temperature slightly less than that required for the full cycle to go into operation. The modification of the C to N ratio at the surface of the star, which is expected because of convective mixing, has been verified in some cases by observations of abundances in red giant stars. The oxygen surface abundance is not affected. This process is known as the *first dredge-up*. As the sun evolves to high luminosity on the red giant branch, it undergoes a mass loss episode, which, depending on the uncertain mass loss rate, could result in the loss of the outer 25% of the mass.

Helium burning in the core finally begins when its mass is about 0.45 $M_\odot$. At this time, $L \approx 2 \times 10^3\ L_\odot$, $T_c = 10^8$K, and $\rho_c = 8 \times 10^5$ g cm^{-3}. At the center, the electron gas is quite degenerate; thus, the total gas pressure depends strongly on density but not on temperature. The helium-burning reaction rate is very sensitive to T; when the reaction starts, the local region heats somewhat and the reaction rate increases, resulting in further heating. Under normal circumstances, the increased T would result in increased pressure, causing the region to expand to the point where the rate of energy generation matched the rate at which the energy could be carried away. However, under degenerate conditions the pressure does not respond, only very slight expansion occurs, and the region simply heats, resulting in a runaway growth of the energy generation. This thermal instability is known as the *helium flash*, during which the luminosity can increase to $10^{11}\ L_\odot$ for a brief period at the flash site. This

enormous luminosity is absorbed primarily in the gradual expansion of the core, whose density decreases by a factor 40, and the luminosity at the surface does not increase. On the contrary, the expansion of the inner regions results in a decrease in the temperature at the hydrogen shell source, a reduction in the energy generation, a reduction of the surface luminosity, and a contraction of the outer layers to compensate. The temperature in the core rises to the point where it is no longer degenerate, cooling can occur, and the nuclear reaction rate once more becomes regulated. Relatively little He is actually burned during the flash, and the star settles down near the red giant branch, but with a reduced $L = 40\ L_\odot$ (see Fig. 10). Energy production comes from both core helium burning and shell hydrogen burning. Although a convection zone develops in the region of the helium flash, it does not link up with the outer convection zone, and so the products of helium burning are not mixed outward to the surface of the star at this time.

Helium burns in the core, converting it to C and O. When the helium is exhausted in the core, that region begins to contract until helium burning is established in a shell region. The hydrogen-burning shell is still active; thus, the star now has a double shell source surrounding the core, and the convective envelope still extends inward to a point just outside the hydrogen-burning shell. The star resumes its upward climb in luminosity along the asymptotic giant branch (AGB; see Fig. 10). The shell sources become thinner, hotter, and closer together. The core becomes degenerate, and the increase in its temperature stops because the energy released by contraction must go into lifting the electrons into higher and higher energy states and because of neutrino losses. The structure of the star is divided into three regions: (1) the degenerate C/O core, which grows in mass to about 0.6 $M_\odot$, in ρ_c to over 10^6 g cm^{-3}, and maintains a radius of about 10^9 cm; (2) the hydrogen- and helium-burning shells, which are separated by a very thin layer of He and which contain only a tiny fraction of the total mass; and (3) the extended convective envelope, which still has essentially its original abundances of H and He and a mean density of only about 10^{-7} g cm^{-3}, expands to a maximum size of about 200 $R_\odot$ (see Fig. 11). The helium-burning shell is subject to a thermal instability during which the energy generation increases locally to peak values of $10^6\ L_\odot$. These events, known as helium shell flashes, repeat with a period of about 10^5 years and result in surface luminosity changes of factors of 5–10.

As the star increases in luminosity, up to a maximum of about 5000 $L_\odot$, it again develops an increasingly strong stellar wind, as a result of which it undergoes further mass

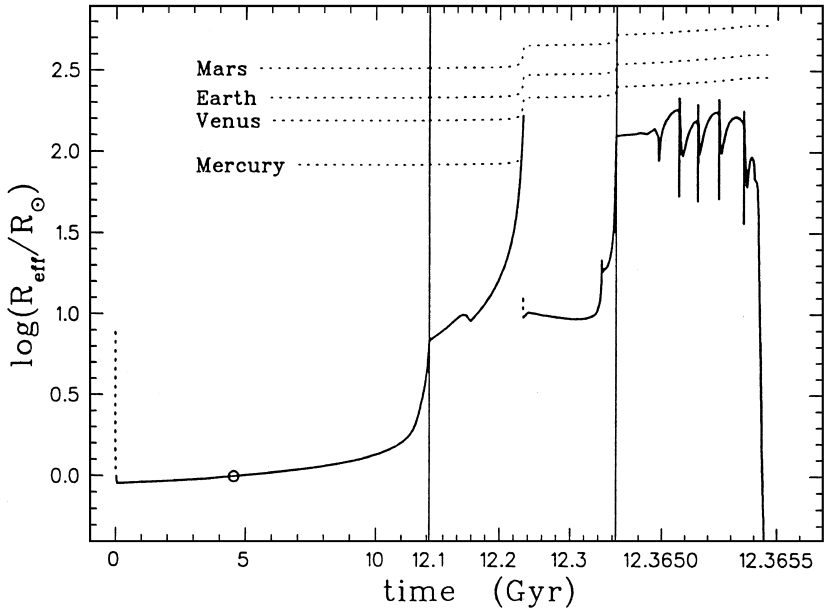


FIGURE 11 The outer radius (R_{eff}) of a model of the evolution of the sun, as a function of time, starting at the zero-age main sequence (solid line). The dashed line refers to the pre-main-sequence phase. The present sun is indicated by an open circle. Dotted lines indicate the orbital distances of the inner planets. These orbits move outward at times when the sun undergoes rapid mass loss. At the end of the simulation, the sun's mass has been reduced to 0.54 $M_\odot$. The oscillations in the right-hand part of the diagram are caused by helium shell flashes. [Adapted with permission from Sackmann, I.-J., Boothroyd, A. I., and Kraemer, K. E. (1993). *Astrophys. J.* **418,** 457. © The American Astronomical Society.]

loss. Although the mechanism behind this final mass ejection is not completely understood, it is probable that the star first becomes pulsationally unstable. In fact, many of the stars in this part of the H–R diagram are variable in light with periods of about a year. As the luminosity increases, the amplitude of the oscillations increases and eventually grows to the point where a small amount of mass at the outer edge is brought to escape velocity. The formation of grains in the outer, very cool layers of these pulsating stars may contribute to the mass loss; the radiation pressure on them can result in a stellar wind. In any case, practically the entire hydrogen-rich envelope is ejected, leaving behind the core, the shell sources, and a thin layer of hydrogen-rich material on top. The star now enters the planetary nebula phase.

The system now consists of a compact central star and an expanding diffuse envelope, known as the planetary nebula. The star evolves to the left in the H–R diagram at a nearly constant luminosity of about 5000 $L_\odot$, with T_{eff} increasing from the red giant value of about 4000 to over 10^5 K. Once T_{eff} exceeds 30,000 K, the ultraviolet radiation results in fluorescence in the nebula, causing it to glow in optical light (Fig. 12). The evolution time from the red giant region to maximum T_{eff} is about 10^5 years, during which time the radius decreases from its maximum value to about 0.1 $R_\odot$ (Fig. 11). The nuclear energy source from one or both shells is still active. The hydrogen-rich outer envelope decreases in mass as the fuel is burned, and eventually, when the envelope mass decreases to $\sim 10^{-4}$ $M_\odot$, it is no longer hot enough to burn. The outer layers contract until the radius approaches a limiting value, at which point only a very slight amount of additional gravitational contraction is possible because practically the entire star is highly electron degenerate, and the high electron pressure supports it against gravity. Essentially, the only energy source left is from the cooling of the hot interior. From this point the star evolves downward and to the right in the H–R diagram and soon enters the white dwarf region. Thus, the final state of the evolution of 1 $M_\odot$ is a white dwarf of about 0.6 $M_\odot$.

FIGURE 12 The giant planetary nebula NGC 7293 (Shane 120-in. reflector). (Lick Observatory photograph.)

B. Evolution of 5 $M_\odot$ Stars

The evolution of higher mass stars differs from that of the lower mass stars in several respects. On the main sequence the star of 5 $M_\odot$ burns hydrogen on the CNO cycle and has a convective core that includes about 23% of the mass at first, but which decreases in mass with time. The H is uniformly depleted in the core. With the exhaustion of H in the core, the star undergoes a brief overall contraction that leads to heating and ignition of the shell source; the convection zone has now disappeared. As the shell narrows, the central regions interior to it contract, while the outer regions expand. The star rapidly crosses the region between the main sequence and the red giant branch. Because of this rapid crossing, very few stars would be expected to be observed in this region, and for that reason it is known as the Hertzsprung gap. As T_{eff} decreases, a convection zone develops at the surface and advances inward; when it includes about half the mass, the star turns upward in the H–R diagram. When $\log L/L_\odot = 3.1$ and the hydrogen-exhausted core includes about 0.75 $M_\odot$, helium burning begins at the center. Because this region is not yet degenerate, the helium flash does not occur and the burning is initiated smoothly. A central convection zone develops because of the strong temperature dependence of the triple-alpha reaction. The structure of the star now consists of (1) the convective helium-burning region, (2) a helium-rich region that is not hot enough to burn He, (3) a narrow hydrogen-burning shell source, and (4) an extended outer envelope with essentially unmodified H and He abundances. However, the outer convection zone has extended downward into layers where some C has been converted into N through previous partial operation of the CNO cycle. The relative depletion of C and enhancement of N could now be observable at the stellar surface. This first dredge-up is analogous to that which occurs in 1 $M_\odot$ (see Fig. 13 at a time of 6×10^7 years).

As helium burning progresses, the central regions expand, resulting in a decrease in temperature at the hydrogen-burning shell and a slight drop in luminosity. The expansion of the outer layers is reversed, and the

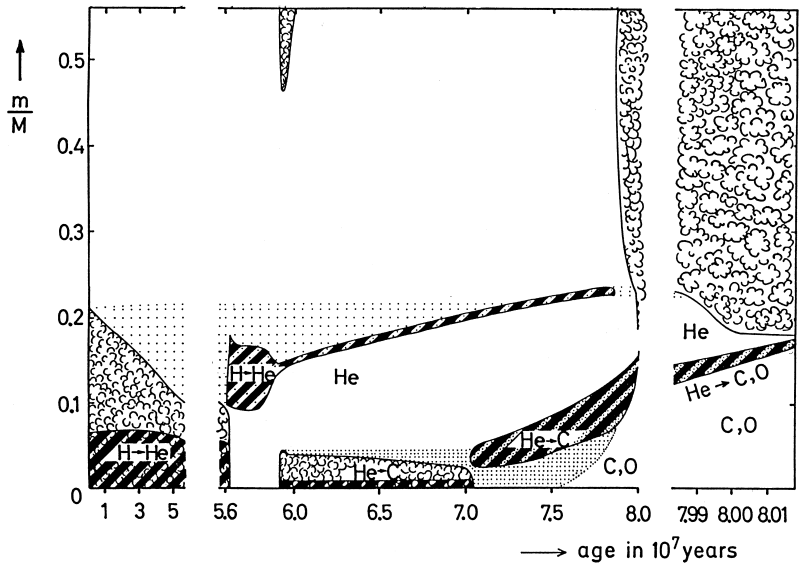


FIGURE 13 The evolution of the internal structure of a star of 5 $M_\odot$ with composition similar to that of the sun. The abscissa gives the age, starting from the zero-age main sequence. A vertical line through the diagram corresponds to a model at a given time. The figure shows only the inner half of the mass of the star. Cloudy regions indicate convection zones. Hatched regions are those with significant nuclear energy production. Regions which are non-uniform in composition are dotted. Breaks in the diagram correspond to changes in the time scale. [Adapted with permission from Kippenhahn, R., and Weigert, A. (1990). "Stellar Structure and Evolution," Springer-Verlag, Berlin. © Springer-Verlag.]

star evolves to higher $T_{\rm eff}$. The energy production is divided between the core and the shell, with the former becoming relatively more important as the star evolves. The total lifetime during helium burning is about 1.4×10^7 years, about 20% of the main-sequence lifetime. As Fig. 10 shows, most of this time is spent in a region where $T_{\rm eff} \approx 6000 - 8000$ K and $\log L/L_\odot \approx 3$, where the Cepheid variables are observed; they are interpreted as stars of 5–9 $M_\odot$ in the phase of core helium burning. The evolutionary calculations show that for solar metallicity only stars in this mass range make the excursion to the left in the H–R diagram during He burning. Lower mass stars stay close to the giant branch during this phase.

As He becomes depleted in the core, the region contracts and heats in order to maintain about the same level of energy production. The contraction results in a slight increase in T at the hydrogen-burning shell. The energy production increases there, and a small amount of this excess energy is deposited in the outer envelope in the form of mechanical work, causing expansion. The evolution changes direction in the H–R diagram and heads back toward the red giant region. The He is used up at the center, a helium-burning shell source is established, and the star resumes its interrupted climb up the giant branch. The surface convection zone develops again and reaches deep within the star to layers that have previously been enriched in He. In this *second dredge-up* the ratio of He to H is increased at the surface, and both C and O are decreased with respect to N. As the star rises to higher L, its structure includes the following regions: (1) a degenerate core consisting of a mixture of ^{12}C and ^{16}O, (2) a narrow helium-burning shell source, (3) a thin layer composed mostly of He, and (4) the extended outer convective envelope, somewhat enriched in He and with modified CNO ratios relative to the original composition. The H-burning shell goes out, to become re-established later on during He shell flashes. The structure of the interior of the star as a function of time up to about this point is illustrated in Fig. 13.

The star increases in L up to about 2×10^4 $L_\odot$, and the mass of the C/O core increases to 0.85 $M_\odot$. A series of thermal flashes, analogous to those found for 1 $M_\odot$, occurs in the He-burning shell; during the flash cycle, the main energy production alternates between an H-burning shell and an He-burning shell. The flashes occur in non-degenerate layers, unlike the He core flash in the lower mass stars. After several flashes, the *third dredge-up* occurs. Some of the C produced in the He-burning shell can be mixed to the surface of the star, eventually producing a C star. Ejection of the envelope now takes place by processes similiar to those described for 1 $M_\odot$. The total mass lost during the evolution of the star is estimated to be 4.15 $M_\odot$; the remaining C/O white dwarf has a mass of 0.85 $M_\odot$. The same result will occur for stars up to 9 $M_\odot$,

with the maximum core mass being about 1.1 $M_\odot$. The second dredge-up occurs only for stars in the mass range from about 4.5 to 9 $M_\odot$; the third dredge-up occurs only in stars from about 2 to 9 $M_\odot$.

C. Evolution of 25 $M_\odot$ Stars

The high-mass stars evolve similarly to the stars of 5 $M_\odot$ during main-sequence and immediate post-main-sequence evolution, except that the fraction of the mass contained in the convective core is larger, and the very highest mass stars may lose mass even on the main sequence. The 25-$M_\odot$ star evolves toward the red giant region, but helium burning at the center occurs when T_{eff} is still in the range 10^4 to 2×10^4 K, and carbon burning starts soon afterward. The star expands to become a red giant with $L/L_\odot = 2 \times 10^5$ and $T_{\text{eff}} = 4500$ K. A sequence of several new nuclear burning phases now occurs in the core, rapidly enough that little concurrent change occurs in the surface characteristics. Carbon burns at about 9×10^8 K, neon burns at 1.75×10^9 K, oxygen burns at 2.3×10^9 K, and silicon burns at 4×10^9 K. A central core of about 1.5 $M_\odot$, composed of iron and nickel, builds up, surrounded by layers that are silicon rich, oxygen rich, and helium rich, respectively. Outside these layers is the envelope, still with its original composition. The layers are separated by active shell sources. The temperature at the center reaches 7×10^9 K and the density is 3×10^9 g cm^{-3} (see Fig. 5). However, the sequence of nuclear reactions that has built the elements up to the iron peak group in the core can proceed no further. These elements have been produced with a net release of energy at every step, a total of 8×10^{18} ergs g^{-1} of hydrogen converted to iron. However, to build up to still heavier elements, a net input of energy is required. Furthermore, the Coulomb barrier for the production of these elements by reactions involving charged particles becomes very high. Instead, an entirely different process occurs in the core. The temperature, and along with it the average photon energy, becomes so high that the photons can react with the iron, breaking it up into helium nuclei and neutrons. This process requires a net input of energy, which must ultimately come from the thermal energy of the gas. The pressure therefore does not rise fast enough to compensate for the increasing force of gravity and the core begins a catastrophic gravitational collapse. On a time scale of less than 1 sec, the central density rises to 10^{14} g cm^{-3} and the temperature rises to 3×10^{10} K. As the density increases, the degenerate free electrons are captured by the nuclei, reducing the electron pressure and further contributing to collapse. At the same time, there is very rapid energy loss from neutrinos. The point is reached where most of the matter is in the form of free neutrons, and when the density becomes high enough, their degenerate pressure increases rapidly enough to stop the collapse. At that point a good fraction of the original iron core has collapsed to a size of 10^6 cm and has formed a neutron star, nearly in hydrostatic equilibrium, with a shock front on its outer edge through which material from the outer parts of the star is falling and becoming decelerated. The core collapse is thought to be the precursor to the event known as a supernova of type II.

The question of what happens after core collapse is one of the most interesting in astrophysics. Can at least part of the gravitational energy released during the collapse be transferred to the envelope and result in its expansion and ejection in the form of a supernova? Present indications are that it is possible and that the shock will propagate outward into the envelope. A large fraction of the gravitational energy is released in the form of neutrinos, produced during the neutronization of the core. Most of these neutrinos simply escape, but the deposition of a small fraction of their energy and momentum in the layers just outside the neutron star is crucial to the ejection process. Assuming that the shock does propagate outward, it passes through the various shells and results in further nuclear processing, including production of a wide variety of elements up to and including the iron peak. It also accelerates most of the material outside the original iron core outward to escape velocities. When the shock reaches the surface of the red giant star, the outermost material is accelerated to 10,000 km sec^{-1}, and the deeper layers reach comparable but somewhat smaller velocities. Luminosity, velocity, and T_{eff} as a function of time in numerical simulations of such an outburst agree well with observations. The enormous luminosity arises, in the earlier stages, from the rapid release of the thermal energy of the envelope. At later times, most of the observed radiation is generated by the radioactive decay of the ^{56}Ni that is produced mainly by explosive silicon burning in the supernova shock. Supernova observations are best fit with total explosion energies of about 10^{51} erg. The Crab Nebula (Fig. 14) is consistent with this energy and an expansion velocity of 10,000 km sec^{-1}. Another good test of the theory of stellar evolution is the calculation of the relative abundances of the elements between oxygen and iron in the ejected supernova envelope. Integration of the yields of these elements over the range of stellar masses that produce supernovae gives values that are consistent with solar abundance ratios.

D. Evolution of Low-Mass Stars with Low-Metal Abundance

The oldest stars in the galaxy are characterized by metal abundances that are 0.1–0.01 that of the sun and by a

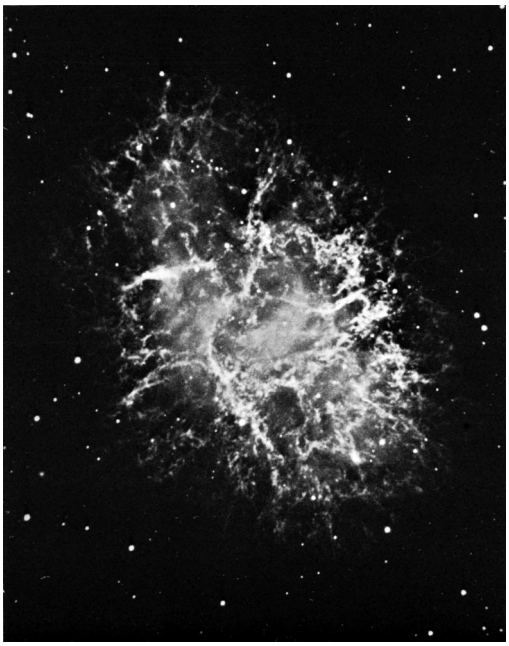


FIGURE 14 Crab Nebula in Taurus (M1, NGC 1952) in red light, the remnant of the supernova explosion in AD 1054 (Shane 120-in. reflector). (Lick Observatory photograph.)

distribution in the galaxy that is roughly spheroidal rather than disklike. These stars can be found in globular clusters or in the general field; in particular, the H–R diagrams of globular clusters (Fig. 15) give important information on the age of the galaxy and the helium abundance at the time the first stars were formed. There are observational difficulties in determining the properties of these stars, since most of them are very distant. For example, it has not been possible to determine observationally the mass–metallicity–luminosity relation for their main sequence. However, observations of the locations in the H–R diagram of a few nearby low-metal stars show that they fall below the main sequence, defined by stars of solar metal abundance. This information, combined with detailed analysis of the H–R diagrams of globular clusters, indicates that the helium abundances in the old stars are slightly less than that of the sun: $Y = 0.24 \pm 0.02$.

The age estimates of globular clusters, based on detailed comparisons of observed H–R diagrams with theoretical evolutionary tracks (see Fig. 4), range from 10 to 15×10^9 years. Recent work, based on improved observations, Hipparcos parallaxes for nearby low-metal stars which are used to calibrate the distances to the globular clusters, and improved theoretical tracks, favors ages in the range 13 to 14×10^9 years for the clusters of lowest metallicity. The high-mass stars in these clusters have long ago evolved to their final states, mostly white dwarfs. The mass of the stars that are now evolving off the main sequence and becoming relatively luminous red giants is about 0.8 $M_\odot$. On the main sequence, however, these stars have approximately the solar luminosity because of their lower metal content, and hence lower opacity, when compared with stars of normal metal abundance. The evolution during the main-sequence phase and the first ascent of the giant branch is very similar to that of 1-$M_\odot$ stars described previously. Hydrogen burning occurs on the pp chains; there is no convection in the core; and when hydrogen is exhausted in the center, the energy source shifts to a shell and the star gradually makes the transition to the red giant region. The core contracts and becomes degenerate, and the envelope expands and becomes convective. The first dredge-up occurs, and a comparison can be made between theoretical and observed element abundances at this stage. Indications are that there is a disagreement in the sense that observed abundances change at earlier phases of evolution than the theory suggests. The implication is that some additional mechanism besides convection is causing mixing into the deep interior. A helium flash in the core occurs when its mass is 0.45 $M_\odot$ and the total luminosity is above 10^3 $L_\odot$.

When the helium flash is completed and the core is no longer degenerate, the star settles down to a stable state with He burning in the core and H burning in a shell. However, the location in the H–R diagram during this phase differs from that of the low-mass stars with solar metals. The star evolves rapidly to a position on the horizontal branch (HB), with T_{eff} considerably higher than that on the giant branch and with L at about 40 $L_\odot$, considerably less than that at the He flash (see Figs. 10 and 15). Calculations show that as the assumed value of Z is decreased for a given mass, the position on the model on the HB shifts to the left. This behavior is in agreement with observations of globular clusters, which show that the leftward extent of the HB is well correlated with observed Z, in the sense that smaller Z corresponds to an HB extending to higher T_{eff}. In theoretical models, as Z is increased to 10^{-2} the HB disappears altogether, and the He-burning phase occurs very close to the red giant branch. This calculation is also in agreement with the observed fact that old clusters with solar Z do not have an HB. The HB itself is not an evolutionary track. The spread of stars along the HB in a cluster of given Z represents a spread in stellar mass, ranging in a typical case from about 0.6 $M_\odot$ at the left-hand edge to about 0.8 $M_\odot$ at the right. The mean HB mass is less than the mass of the stars that are just leaving the main sequence in a given cluster. The implication is that stars lose mass on the red giant branch during the period of their evolution just prior to the He flash and that the total mass lost varies from star to star. The mass loss probably occurs by a mechanism similar to that which drives the solar wind, but much stronger.

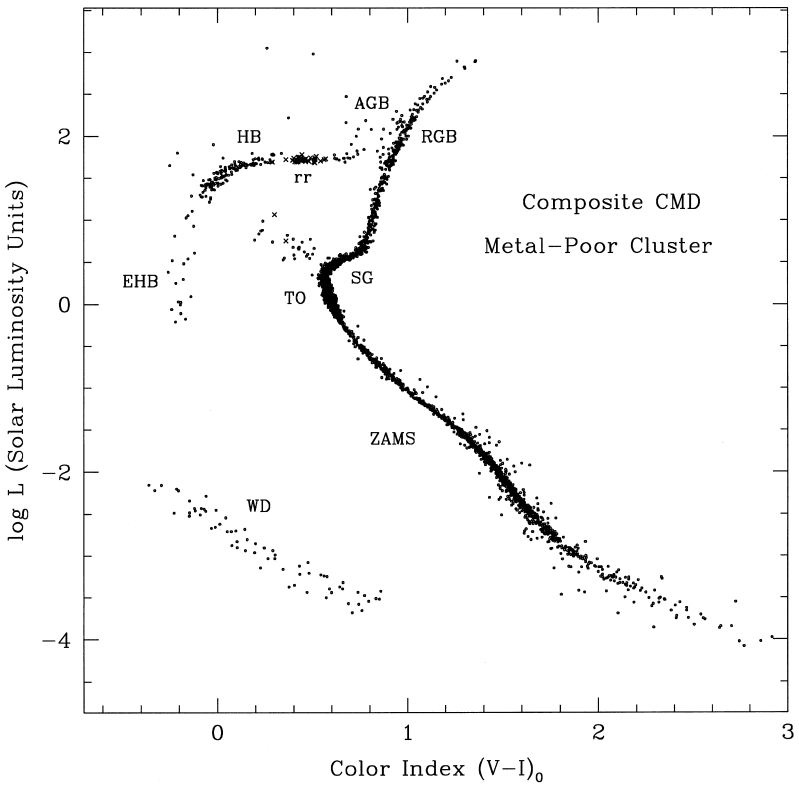


FIGURE 15 Color-magnitude diagram of a low-metal globular cluster composed of data from about five different actual clusters. The various features are ZAMS, zero-age main sequence; WD, white dwarfs; TO, turnoff; SG, subgiant branch; RGB, red giant branch; HB, horizontal branch; rr, RR Lyrae variables; EHB, extreme horizontal branch; and AGB, asymptotic giant branch. The location of the various features can be used to compare theory with observations. [Figure courtesy of William E. Harris.]

The phase of core helium burning on the HB lasts about 10^8 years, after which the star develops a double shell structure and evolves back toward the red giant branch, which during the following phase of evolution is referred to as the asymptotic giant branch. The evolution from this point on is very similar to that of the star of 1 $M_\odot$ with $Z = 0.02$; ascent of the asymptotic giant branch, development of a carbon/oxygen core that becomes increasingly degenerate, development of a pulsational instability in and a wind from the convective envelope, ejection of the envelope, evolution through the planetary nebula phase, and, finally, the transition to the white dwarf phase.

VII. FINAL STATES OF STARS

A. Brown Dwarfs

Although objects below 0.075 $M_\odot$ are not, strictly speaking, stars, their general observational parameters are now becoming available, and the physics that must be included in theoretical models is very similar to that for the lowest mass stars. For example, the equation of state must include non-ideal Coulomb interactions as well as partial electron degeneracy, and the surface layer calculation must include the effects of molecules. Substellar objects never attain internal temperatures high enough that nuclear burning can supply their entire radiated luminosity. They can, however, burn deuterium by Reaction (2), and this energy source can be significant for a short period of time. Below about 0.06 $M_\odot$, internal temperatures never become high enough to burn Li by Reaction (6); thus, most brown dwarfs should show Li lines in their spectra. However, above that mass, the Li does burn near the center, and because the objects are fully convective, the depletion of Li becomes evident at the surface. This *lithium test* has been successfully used observationally to identify brown dwarfs.

The brown dwarfs contract to the point where the electrons become degenerate. For 0.07 $M_\odot$, the value of $T_{\rm eff}$ during contraction is about 2900 K; for 0.02 $M_\odot$, it is about 2500 K. Following that time the contraction becomes very slow and the objects approach an asymptotic value of the radius, depending on mass and composition. At 0.07 $M_\odot$, the time to contract to maximum central

temperature, which corresponds to significant electron degeneracy, is about 3×10^8 years; for lower masses, it is less. Beyond this point the object cools, with decreasing T_{eff} and L. A typical brown dwarf after an evolution time of 3×10^9 years will have $T_{\text{eff}} = 1000$–1500 K and $L/L_\odot = 10^{-5}$ to 10^{-6} and is therefore very difficult to observe. Nevertheless, a number of them have been identified, for example, the companion to Gl 229, which falls exactly into this range and is thought to have a mass of about 0.03–0.05 $M_\odot$. The brown dwarfs are distinguished from white dwarfs first by their very cool surface temperatures and second by the fact that their internal composition is practically unchanged from the time of formation; that is, it is about 71% hydrogen by mass. As a result, the radii are larger than those of white dwarfs, with values near the end of contraction of about 0.1 $R_\odot$ over the mass range 0.01–0.08 $M_\odot$. The dividing line between brown dwarfs and planets is arbitrary. One possibility is the mass below which deuterium burning is no longer possible, about 0.012 $M_\odot$. Or they could be distinguished by their formation mechanism: planets form by accretion in protostellar disks, while brown dwarfs form in the same way as stars, by fragmentation in collapsing interstellar clouds.

B. White Dwarfs

Observational parameters of the most commonly observed white dwarfs are $T_{\text{eff}} = 10{,}000$–$20{,}000$ K and $\log L/L_\odot = -2$ to -3. The objects, therefore, lie well below the main sequence, and the deduced radii fall in the range 10^{-2} $R_\odot$, with mean densities above 10^6 g cm^{-3}. At these high densities, most of the mass of the object falls in the region of complete electron degeneracy; only a thin surface shell is non-degenerate. Only a few white dwarfs occur in binary systems with sufficiently accurate observations so that fundamental mass determinations can be made. Examples are Sirius B, Procyon B, and 40 Eridani B, with masses of 1.05, 0.63, and 0.43 $M_\odot$, respectively.

The structure of a white dwarf can be determined in a straightforward way under the assumptions of hydrostatic equilibrium and uniform composition, with the pressure supplied entirely by the degenerate electrons. The ion pressure and the effect of the thin surface layer are small corrections. Because the pressure is simply a function of density—for example, $P \propto \rho^{5/3}$ in the nonrelativistic limit—it can be eliminated from Eq. (32). The structure can then be calculated from the solution of Eqs. (32) and (33). The results show that for a given composition there is a uniquely defined relation between mass, radius, and central density. In other words, if the central density is considered to be a parameter, for each value of it there corresponds a unique value of the mass and of the radius. This theoretical mass–radius relation, appropriate for a typical C/O white dwarf or for a helium white dwarf, is shown in Fig. 16, where it is compared with the observations of a few white dwarfs. If a more accurate equation of state is included, the line is shifted slightly, depending on composition. Note that the radius decreases with increasing mass and vanishes altogether at 1.46 $M_\odot$.

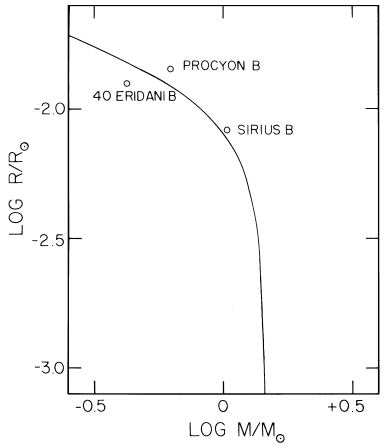


FIGURE 16 Theoretical mass–radius relation derived by S. Chandrasekhar for fully degenerate white dwarfs (solid line) compared with a few observed points (open circles). The masses of Sirius B and 40 Eridani B are known to within 5%; all radii and the mass of Procyon B are less accurately known.

This upper limit to the mass of a white dwarf was first derived by S. Chandrasekhar and corresponds to infinite central density in the formal solution of the equations. In fact, however, the physical process that limits the mass of a white dwarf is capture of the highly degenerate electrons by the nuclei, predominantly carbon and oxygen. Suppose that a white dwarf increases in mass upward toward the limit. The central density goes up, the degree of electron degeneracy goes up, and a critical density is reached, about 10^{10} g cm^{-3} for carbon, where electron capture starts. As free electrons are removed from the gas, the pressure is reduced and the star can no longer maintain hydrostatic equilibrium. Collapse starts, more electron capture takes place, and (assuming no nuclear burning takes place) neutron-rich nuclei are formed. Collapse stops when the neutron degeneracy results in a very high pressure, in other words, when a neutron star has been formed. The white dwarf mass limit is reduced by about 10%, depending upon composition, from the value of 1.46 $M_\odot$ given earlier.

The theoretical mass–radius relation agrees, within the uncertainties, with the observations. The location of the theoretical white dwarfs in the H–R diagram also agrees with observations. From the L and T_{eff} of observed white dwarfs, one can obtain radii and from them the corresponding masses, using the M–R relation. The masses of white dwarfs, determined in this manner, range from 0.42 to 0.7 $M_\odot$. These objects have presumably evolved from

stars of somewhat higher original mass, as indicated in Fig. 10. Their composition is therefore that of the evolved cores of these stars, mainly carbon and oxygen with only a thin layer of hydrogen or helium on the outside. The internal temperatures are not high enough to burn nuclear fuel and contraction is no longer possible, so the only energy source available is the thermal energy of the ions in the interior. The star's L, $T_{\rm eff}$, and internal temperature all decrease as the star evolves at a constant radius. The energy transport in the interior is by conduction by the free electrons, which are moving very rapidly and have long collisional mean free paths. The conduction is efficient with the result that the interior is nearly isothermal; the energy loss rate is controlled by the radiative opacity in the thin outer nondegenerate layers. These loss rates are sufficiently low that the cooling times are generally several billion years.

C. Neutron Stars

These objects, whose existence as condensed remnants left behind by supernova explosions was predicted by F. Zwicky and W. Baade in 1934, were first discovered by Bell and Hewish in 1967 from their pulsed radio radiation. The typical mass is 1.4 $M_\odot$, corresponding to the mass of the iron core that collapses for stars in the 10- to 20-$M_\odot$ range, the typical radius is 10 km, and the mean density is above 10^{14} g cm^{-3}, close to the density of the atomic nucleus. At this high density the main constituent is free neutrons, with a small concentration of protons, electrons, and other elementary particles. The surface temperature is very difficult to determine directly from observations. Neutron stars form as collapsing cores of massive stars and reach temperatures of 10^9–10^{10} K during their formation. However, they then cool very rapidly by neutrino emission; within one month $T_{\rm eff}$ is less than 10^8 K, and within 10^5 years it is less than 10^6 K. Their luminosity soon becomes so low ($\sim 10^{-5}\ L_\odot$) that they would not be detected unless they were very nearby. The fact that neutron stars are observable arises from the remarkable properties that (1) they are very strongly magnetic, (2) they rotate rapidly, and (3) they emit strong radio radiation in a narrow cone about their magnetic axes. The rotation axis and the magnetic axis are not coincident, and so if the star is oriented so that the radio beam sweeps through the observer's instrument, he or she will observe radio pulses separated by equal time intervals. The observed pulsars are thus interpreted as rapidly spinning neutron stars with rotation periods in the range 4.3 to 1.6 msec, with an average of 0.79 sec. The pulsar in the Crab nebula (Fig. 14) has a period of 1/30 sec. The origin of the beamed radiation, which is also observed in the optical and X-ray regions of the spectrum, is not known. The electromagnetic energy is emitted at the expense of the rotational energy, and the rotation periods are gradually becoming longer.

D. Black Holes

Neutron stars also have a limiting mass, but it is not known as well as that for white dwarfs because the physics of the interior, in particular the equation of state, is not fully understood. A reasonable estimate is 2–3 $M_\odot$. If the collapsing core of a massive star exceeds this limit, the collapse will not be stopped at neutron star densities because the pressure gradient can never be steep enough to balance gravity. The collapse continues indefinitely, to infinite density, and a black hole is formed. Stars with original masses in the range 10–20 $M_\odot$ form cores of around 1.4 $M_\odot$; they will become supernovae and leave a neutron star as a compact remnant. Stars above 25–30 $M_\odot$ have larger cores at the time of collapse, and the amount of mass that collapses probably exceeds the neutron star mass limit; the result is a black hole.

E. Supernovae

Only very small fraction of stars ever become supernovae. Two main types are recognized. The supernovae associated with the collapse of the iron core of massive stars are called type II. They tend to be observed in the spiral arms of galaxies, where recent star formation has taken place. The luminosity at maximum light is about $10^{10}\ L_\odot$, and the decline in luminosity with time takes various forms with a typical decline time of a few months. The observed temperature drops from about 15,000 to 5000 K during the first 30–60 days, and the expansion velocity decreases from 10,000 to 4000 km sec^{-1} over the same time period. The abundances of the elements are similar to those in the sun, with hydrogen clearly present. The type I supernovae, on the other hand, tend to be observed in regions associated with an older stellar population, for example, in elliptical galaxies. The luminosity at maximum is about the same as that for type II, and the curves of the decline in light with time are all remarkably similar, with a rapid early decline for 30 days and then a more gradual exponential decay in which the luminosity decreases by a factor 10 in typically 150 days. Temperatures are about the same as in type II, the expansion velocities are somewhat higher, and the element abundances are quite different: there is little evidence for hydrogen, or helium and the spectrum is dominated by lines of heavy elements. The origin of the type I supernova is controversial, but one promising model, which applies to the so-called type Ia, is that it represents the end point of the evolution of a close binary system. Just before the explosion the system consists of a white dwarf near the limiting mass and a companion of relatively low mass,

which lasts for several Gyr on the main sequence. When the companion evolves, it expands and at some point transfers mass to the white dwarf, which therefore is accreting material that is hydrogen rich. The hydrogen burns near the surface, producing a layer of helium that increases in mass. By the time the helium is hot enough to burn, it does so under degenerate conditions. The temperature increases rapidly to the point where the carbon ignites explosively. Much of the material is processed to heavier elements such as iron, nickel, cobalt, and silicon. It still is not entirely certain whether the whole star blows up or whether a neutron star or white dwarf remnant is left. The light emitted is produced primarily from the radioactive decay of ^{56}Ni. In any case, the type Ia supernova is essentially a nuclear explosion, while the type II supernova is driven ultimately by the rapid release of gravitational energy during core collapse.

VIII. SUMMARY: IMPORTANT UNRESOLVED PROBLEMS

The study of stellar structure and evolution has resulted in major achievements in the form of quantitative and qualitative agreement with observed features, in the pre-main-sequence, main-sequence, and post-main-sequence phases. A number of aspects of the physics need to be improved, and improvement in many cases means extending the theory from the relatively simple one-dimensional calculations into two- and three-dimensional hydrodynamic simulations. Considerable uncertainty is caused by a lack of a detailed theory of convection that can be applied in stellar interiors. On the main sequence this uncertainty could result in relatively minor changes in the surface layers of stars, mainly in the mass range 0.8–1.1 $M_{\odot}$. Three-dimensional hydrodynamic simulations of convection near the surface of sunlike stars provide very good agreement with the observed properties of solar granulation and can provide a calibration for the value of the mixing length to be used in long-term calculations over billions of years. However, the question of overshooting at the edges of interior convection zones still needs to be solved. In pre-main-sequence and post-main-sequence stars, which have deep extended convection zones, a better theory could result in some major changes. There are several examples of situations where observations suggest mixing of the products of nucleosynthesis to the surface of stars, but the theoretical models of convection zones do not provide sufficiently deep mixing. Other possible mixing mechanisms, including overshoot, need to be considered. Improvement in laboratory measurements of nuclear reaction rates, particularly for the pp chains, could reduce the theoretical error bar in the rate of solar neutrino production. The continued discrepancy in the solar neutrino detection rates, along with the excellent agreement between the solar model and the helioseismological results, suggests that neutrino oscillations may actually be the solution to the puzzle. New experiments under way, such as SNO, should provide further clues on this question.

The assumption of mass conservation is built into much of the theory discussed in this article. However, it is known from observation that various types of stars are losing mass, including the T Tauri stars in the early pre-main-sequence contraction phase, the O stars on the upper main sequence, and many red giant stars. The physics of the ejection process is not well explained, especially in the case of rapid mass loss leading to the planetary nebula phase. Also, in the early phases of evolution, the mass loss, often in the form of bipolar ouflows and jets, is associated with simultaneous accretion of mass onto the star through a disk. The physics of the mass and angular momentum transfer at this stage needs to be clarified.

Binary star evolution has not been treated here, but numerous fascinating issues remain for further study. If the two components interact, the problem again involves three-dimensional hydrodynamics. The first difficulty is the formation process. Most stars are found in double or multiple systems, but their origin, presumably a result of fragmentation of collapsing, rotating interstellar clouds, is not well understood, particularly for the close binaries. A further problem is the explanation of short-period systems consisting of a white dwarf and a nearby low-mass main-sequence companion. These systems are associated with nova outbursts and possibly supernovae of type I. The problem is that the orbital angular momentum of such a binary is much smaller than that of the original main-sequence binary from which the system must have evolved. A goal of current research is to uncover the process by which the loss of angular momentum occurs. A likely candidate is the "common envelope" phase. An expanding red giant envelope interacts with its main-sequence companion, so that the latter spirals into the giant, losing angular momentum and energy because of frictional drag and gravitational torques. The energy lost from the orbital motion is deposited in the envelope, resulting in its ejection, leaving the main-sequence star in a short-period orbit about the white dwarf core of the giant. Clarification of this process requires detailed three-dimensional hydrodynamic simulations.

The effects of rotation and magnetic fields have also not been discussed in this article. However, their interaction in the context of stellar evolution is certain to produce some very interesting modifications to existing theory. During the phase of star formation these processes are central, and consideration of them is crucial for the clarification of the processes of binary formation and generation of jets.

During main-sequence evolution, rotational and magnetic energies, although present, are almost certainly small compared with gravitational energy, and their overall effect on the structure is therefore insignificant. However, the circulation currents induced by rotation could be important with regard to mixing and the exchange of matter between the deep interior and the surface layers. During post-main-sequence phases, the cores of stars contract to very high densities, and if angular momentum is conserved in them from the main-sequence phase, they would be expected to rotate very rapidly by the time the core becomes degenerate. However, white dwarfs are rotating slowly, suggesting that at some stage almost all of the angular momentum is transferred, by an as yet unexplained process, out of the cores of evolving stars. The same problem applies to the rotation of neutron stars.

A number of other problems present themselves. The complicated physical processes involved in the generation of supernovae explosions and the formation of neutron stars require further study, in some cases involving three-dimensional hydrodynamics. Here, rotational and magnetic effects again become important, as well as neutrino transfer, development of jets, and the possible generation of gamma-ray bursts. The observational and theoretical study of stellar oscillations can provide substantial information on the structure and evolution of stars. For example, the interior of the sun has been probed through the analysis of its short-period oscillations, and the techniques can be extended to other stars. For many types of stars (including the sun!), the physical mechanism producing the oscillations is not understood. The determination of the ages of globular clusters, presumably containing the oldest stars in the galaxy, through the use of stellar evolutionary tracks, is a particularly interesting problem as it helps to constrain cosmological models. It is clear that the study of stellar evolution involves a wide range of interacting physical processes and that the findings are of importance in numerous other areas of astronomy and physics.

SEE ALSO THE FOLLOWING ARTICLES

BINARY STARS • DARK MATTER IN THE UNIVERSE • GALACTIC STRUCTURE AND EVOLUTION • NEUTRINO ASTRONOMY • NEUTRON STARS • STAR CLUSTERS • STARS, MASSIVE • STARS, VARIABLE • STELLAR SPECTROSCOPY • SUPERNOVAE

BIBLIOGRAPHY

Bahcall, J. N. (1989). "Neutrino Astrophysics," Cambridge Univ. Press, Cambridge, UK.

Bowers, R., and Deeming, T. (1984). "Astrophysics I. Stars," Jones & Bartlett, Boston.

Clayton, D. D. (1983). "Principles of Stellar Evolution and Nucleosynthesis," Univ. of Chicago Press, Chicago.

Goldberg, H. S., and Scandron, M. (1981). "Physics of Stellar Evolution and Cosmology," Gordon & Breach, New York.

Kippenhahn, R. (1983). "100 Billion Suns," Basic Books, New York.

Kippenhahn, R., and Weigert, A. (1990). "Stellar Structure and Evolution," Springer-Verlag, Berlin.

Phillips, A. C. (1994). "Physics of Stars," Wiley, Chichester.

Shklovskii, I. S. (1978). "The Stars: Their Birth, Life, and Death," Freeman, San Francisco.

Tayler, R. J. (1994). "The Stars: Their Structure and Evolution," Cambridge Univ. Press, Cambridge, UK.

Supernovae

David Branch
University of Oklahoma

GLOSSARY

Absolute magnitude The apparent magnitude that a star would have if it were at a distance of 10 pc.

Apparent magnitude A logarithmic measure of the brightness of a star as measured through, e.g., a blue or visual filter. A difference of five magnitudes corresponds to a factor of 100 in observed flux; one magnitude corresponds to a factor of 2.512. Increasing magnitude corresponds to decreasing brightness.

Light curve Apparent magnitude, absolute magnitude, or luminosity, plotted against time.

Luminosity The rate at which a star radiates energy into space. The luminosity of a supernova usually refers only to the thermal radiation in the optical, ultraviolet, and infrared bands, and does not include nonthermal γ rays, X-rays, and radio waves.

Neutron star A very compact star that has a radius of the order of 10 km and a density that exceeds that of nuclear matter. The maximum mass is thought to be less than 3 solar masses. A neutron star is one possible final state of the core of a massive star (the other being a black hole).

Parsec The distance at which a star would have an annual trigonometric parallax of 1 sec of arc. One parsec (pc) equals 206,625 astronomical units, or 3.1×10^{18} cm. Distances to the nearest stars are conveniently expressed as pc; distances across our Galaxy, as kiloparsecs (kpc); distances to external galaxies, as megaparsecs (Mpc).

Photosphere The layer in a star from which photons escape to form a continuous spectrum; equivalently, the layer at which the star becomes opaque to an external observer's line of sight.

Stellar wind The steady flow of matter from a (nonexploding) star into surrounding space.

White dwarf A compact star whose internal pressure is provided by degenerate electrons. A typical white dwarf has a radius comparable to that of Earth but a density of the order of 10^7 g/cm^3. The maximum mass, the Chandrasekhar mass, is 1.4 solar masses. A white

dwarf is the final state of a star that forms with less than 8 solar masses.

A SUPERNOVA is a catastrophic explosion of a star. The luminosity of a bright supernova can be 10^{10} times that of the Sun. The explosion throws matter into space at a few percent of the speed of light, with a kinetic energy of 10^{51} erg—the energy equivalent of 10^{28} Mtons of TNT. The ejected matter is enriched in heavy elements, some that were synthesized by nuclear fusion reactions during the slow preexplosion evolution of the star and some that were synthesized during the explosion itself. Thus, supernovae drive the nuclear evolution of the universe.

Some supernovae are the complete thermonuclear disruptions of white dwarf stars; others are initiated by the gravitational collapse of the cores of massive stars. The latter eject only their outer envelopes and leave behind compact stellar remnants—neutron stars and black holes. The ejected matter of a supernova sweeps up and heats interstellar gas to form an extended *supernova remnant*.

Supernovae are valuable indicators of extragalactic distances. They are used to establish the extragalactic distance scale (the Hubble constant) and to probe the history of the cosmic expansion (deceleration and acceleration).

I. BASIC OBSERVATIONS

A. Discovery

At least five of the temporary "new" stars that suddenly appeared in the sky during the last millennium—in 1006, 1054, 1181, 1572, and 1604—are now known to have been supernova explosions in our Galaxy. Some of these were bright enough to be seen by eye even in daylight. Radio, X-ray, and optical emission from the extended remnants of these historical Galactic supernovae is now detected. Hundreds of other extended remnants of supernovae that occurred in the Galaxy within the last 100,000 years also are recognized.

No Galactic supernova has been discovered since 1604, just a few years before the invention of the telescope. The supernova remnant known as Cassiopeia A, the brightest radio source in the sky beyond the solar system, was produced by a supernova that occurred near 1680, but the event was subluminous and either was not noticed or was dismissed as an ordinary variable star. Most of the supernovae that occur in our Galaxy are not detected by observers on Earth, owing to extinction of their light by dust grains that are concentrated within the disk of the Galaxy.

On February 24, 1987, an extraordinary astronomical event took place when a supernova appeared in the Large Magellanic Cloud (LMC), a small irregular satellite galaxy of our Galaxy. At a distance of only 50 kpc (160,000 light years), the LMC is the nearest external galaxy. At its brightest, SN 1987A reached the third apparent visual magnitude and was easily visible to the naked eye (from sufficiently southern latitudes). SN 1987A was the brightest supernova since 1604, and it became by far the most well-observed supernova. Its aftermath will be observed long into the future, perhaps as long as there are astronomers on Earth.

Until another outburst is seen in our Galaxy, or in one of our satellite galaxies, studies of the explosive phases of supernovae must be based on more distant events. The study of extragalactic supernovae began in August 1885, when a star of the sixth apparent visual magnitude appeared near the nucleus of the Andromeda Nebula, now known to be the nearest large external galaxy, at a distance of less than 0.8 Mpc. It was not until 1934 that the titanic scale of such events was generally recognized, and the term *supernova* began to be used. Systematic photographic searches for supernovae began in 1936. The following 35 years were primarily exploratory, as some of the basic characteristics of the various supernova types were established, and statistical information was gathered. A convention for designating each event eventually was adopted: e.g., SN 1987A was the first supernova to be discovered in 1987, and SN 1979C was the third of 1979. By the end of 1989, 687 extragalactic supernovae had been seen. During the late 1990s the discovery rate increased dramatically, owing mainly to searches with modern detectors (CCDs) for very distant supernovae, which can be discovered in batches. SN 1999Z was followed by SN 1999aa to SN 1999az, then SN 1999ba to SN 1999bz, and so on, to SN 1999gu. By the end of 1999, 1447 supernovae had been found (Fig. 1).

B. Types

The classification of supernovae is based mainly on the appearance of their optical spectra (Figs. 2 and 3). Type I supernovae (so named because the first well-observed events of the 1930s were of this kind) show no conspicuous spectral features produced by hydrogen; Type II supernovae do show obvious hydrogen lines. Type I supernovae are subdivided according to other spectroscopic characteristics: those of Type Ia have a very distinctive spectral evolution that includes the presence of a strong red absorption line produced by singly ionized silicon; the spectra of Type Ib supernovae include strong lines of neutral helium; neither ionized silicon nor neutral helium is strong in the spectra of Type Ic. Type II supernovae that have especially *narrow* lines are called Type IIn. Type II events also are divided into Type II-P and Type II-L according to the shapes of their light curves.

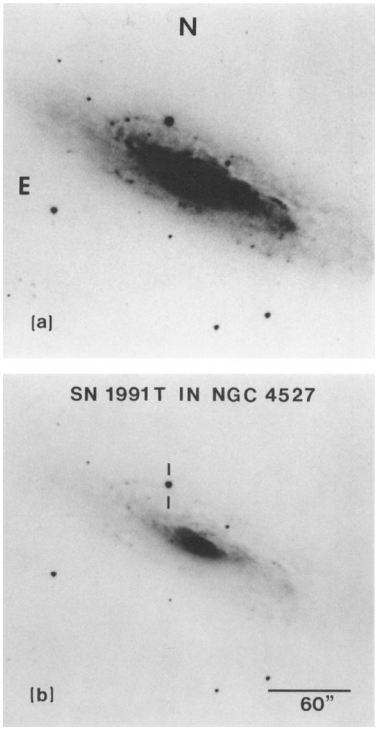


FIGURE 1 CCD images of SN 1991T in its parent galaxy NGC 4527, displayed at two contrast levels. [Reproduced by permission from Filippenko, A. V., *et al.* (1991). *Astrophys. J.* **384,** L15.]

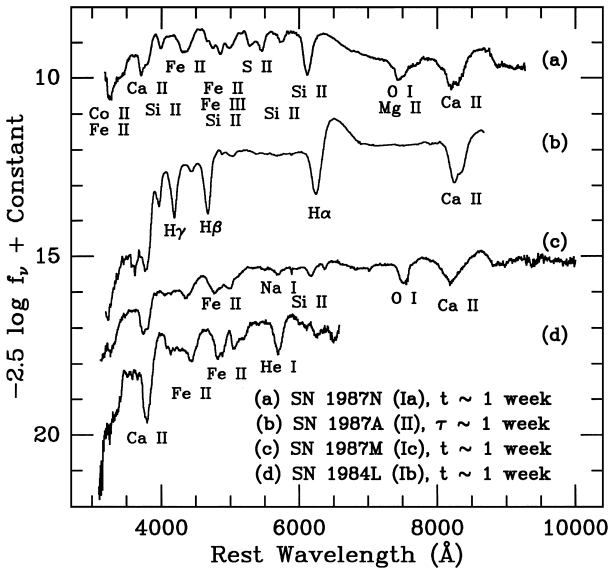


FIGURE 2 Photospheric-phase optical spectra of supernovae of various types, obtained 1 week after the time of maximum brightness (1 week after the time of explosion in the case of SN 1987A). [Reproduced by permission from Branch, D., Nomoto, K., and Filippenko, A. V. (1991). *Comments Astrophys.* **XV,** 221.]

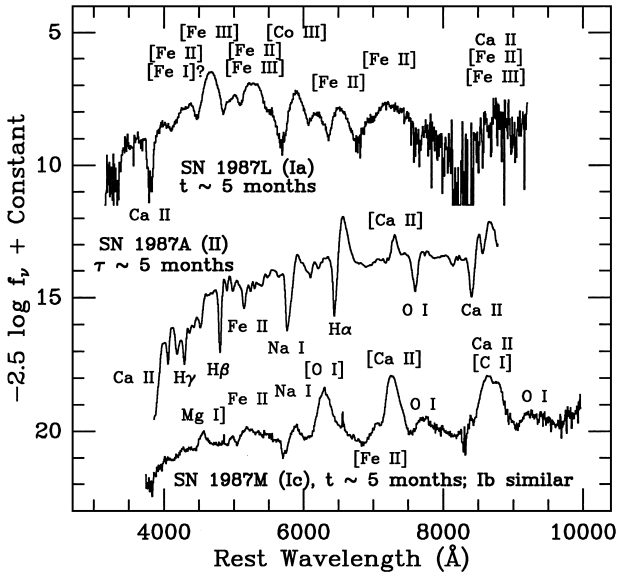


FIGURE 3 Nebular-phase optical spectra of supernovae, obtained 5 months after the time of maximum brightness (5 months after the time of explosion in the case of SN 1987A). [Reproduced by permission from Branch, D., Nomoto, K., and Filippenko, A. V. (1991). *Comments Astrophys.* **XV,** 221.]

C. Light Curves

A typical supernova reaches its maximum brightness about 20 days after explosion. At its brightest, a normal Type Ia supernova (SN Ia) reaches an absolute visual magnitude of -19.5 and has a luminosity exceeding 10^{43} erg/sec, billions of times that of the Sun. SNe Ia (plural) are highly homogeneous with respect to peak absolute magnitude as well as other observable properties. Supernovae of the other types show more observational diversity, and almost all of them are less luminous than SNe Ia. The time-integrated luminosity of a bright supernova exceeds 10^{49} erg, while the typical kinetic energy is 10^{51} erg; thus the radiative efficiency of a supernova is low.

Characteristic light-curve shapes of the various supernova types are compared in Fig. 4. The light curve of a SN Ia consists of an initial rise and fall (the early *peak*) that lasts about 40 days, followed by a slowly fading *tail*. The tail is nearly linear when magnitude is plotted against time, but since magnitude is a logarithmic measure of brightness, the tail actually corresponds to an exponential decay of brightness, The rate of decline of the SN Ia tail corresponds to a half-life of about 50 days.

The light curves of many SNe II interrupt the initial decline from their peaks to enter a "*plateau*" phase of nearly constant brightness for several months; these are designated Type II-P. Others show a nearly *linear* decline (in magnitudes) from peak, with little or no plateau; these are

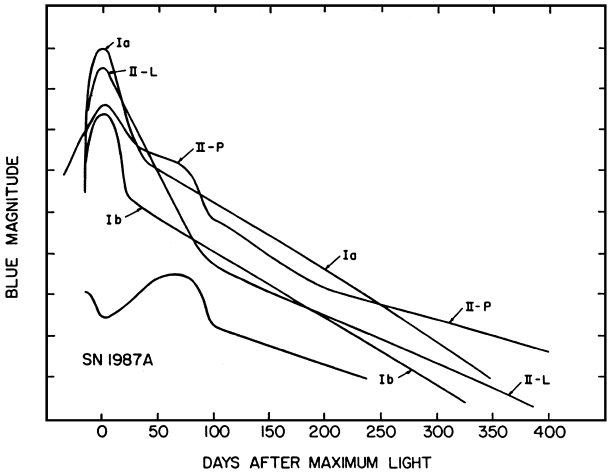


FIGURE 4 Schematic light curves of supernovae of various types. [Reproduced by permission from Wheeler, J. C. (1990). *In* "Supernovae" (J. C. Wheeler, T. Piran, and S. Weinberg, eds.), p. 1, World Scientific, Singapore.]

called Type II-L. Some SNe II, especially the SNe IIn, have slowly fading light curves that fit neither the SN II-P nor the SN II-L categories. Most SNe II eventually enter a linear tail phase, many with a rate of decline that corresponds to a half-life near 77 days.

SN 1987A was subluminous, and its light curve was unusual. It was observationally conspicuous only because it occurred in a very nearby galaxy; most such subluminous events in distant galaxies go undiscovered. This illustrates the importance of a very strong observational selection effect: in our observational sample of supernovae, luminous events are highly overrepresented relative to subluminous ones (Fig. 5).

D. Sites, Rates, and Stellar Populations

Galaxies are classified as spirals and ellipticals. SNe II, SNe Ib, and SNe Ic are found in spirals, usually in the spiral arms. SNe Ia are found in both kinds of galaxies, and those in spirals show little, if any, tendency to concentrate to the arms.

To estimate the relative production rates of the different types of supernovae in the various kinds of galaxies, several observational selection effects must be taken into account. The most important is the bias in favor of the discovery of luminous events. For example, because of their high peak luminosities, SNe Ia are the most numerous type in the observational sample, but they actually occur less frequently than SNe II. Another important effect is that it is difficult to discover supernovae in spiral galaxies whose disks are oriented such that we observe them from the side. When these and additional selection effects are allowed for, it is found that in spirals, SNe II are most frequent, followed by SNe Ic, SNe Ib, and SNe Ia, which occur at comparable rates. In spirals, the rates of all supernova types are correlated with galaxy color: the bluer the galaxy—and, by inference, the higher the star formation rate within the last billion years—the higher the supernova rate. Elliptical galaxies produce only SNe Ia, at a lower rate than spirals.

Absolute supernova rates are more uncertain than relative rates because the relevant selection effects are still more difficult to evaluate. In large spiral galaxies, the mean interval between supernovae is about 25 years. A similar rate is derived for our Galaxy itself, from the number of historical supernovae that have been seen, but this involves a large correction for events that have been missed owing to our location in the dusty disk of the Milky Way. The fact that the remnants of all of the historical Galactic supernovae are within a few kiloparsecs of the Sun, while the radius of the galaxy is 15 kpc, confirms that only (some of) the nearest Galactic supernovae of the past millennium have been noticed by observers on Earth.

The observation that SNe II, SNe Ib, and SNe Ic appear in the arms of spirals indicates that they are produced by massive stars. Only stars that are formed with a mass that is greater than about 8 solar masses have nuclear lifetimes short enough—less than about 30 million years—that they explode before drifting out of the arms in which they were born.

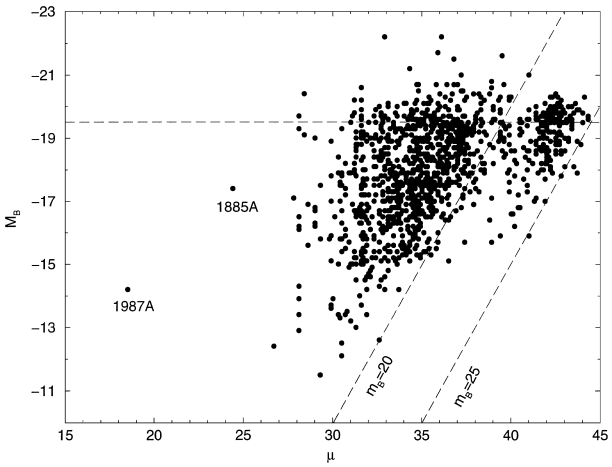


FIGURE 5 Absolute blue magnitudes of supernovae are plotted against the distance modulus of their parent galaxies. The distance modulus is five times the logarithm of the distance, expressed in units of 10 pc. The slanted dashed lines correspond to apparent blue magnitudes of 20 and 25. The horizontal dashed line is at the absolute magnitude of a normal SN Ia, $M_B = -19.5$. For each supernova, the absolute magnitude is derived from the brightest apparent magnitude that was observed, which was not in all cases at the time of maximum brightness. [Courtesy of D. Richardson, University of Oklahoma.]

The observation that SNe Ia occur in elliptical galaxies, but also in spirals in proportion to the recent rate of star formation, appears at first to be paradoxical. Elliptical galaxies stopped forming stars billions of years ago and now contain few stars that are much more massive than the Sun. On the other hand, the correlation between the SN Ia rate and the recent star-formation rate in spirals implies that most SNe Ia in spirals are produced by stars that are fairly short-lived and, therefore, born moderately massive. The resolution of the paradox is that a SN Ia is produced by a white dwarf that accretes matter from a binary companion star until it is provoked to explode. This can account for the low but nonzero SN Ia rate in ellipticals as well as for the correlation of the SN Ia rate with galaxy color in spirals (Section III.C).

II. THE INTERPRETATION OF SPECTRA

A. The Continuous Spectrum

The shape of a star's thermal continuous spectrum is determined primarily by the temperature at the photosphere. The absolute brightness depends also on the radius of the photosphere. In a supernova, the temperature and radius of the photosphere change with time. When a star explodes, its matter is heated and thrown into rapid expansion. The luminosity abruptly increases in response to the high temperature, but most of this energy is radiated as an X-ray and ultraviolet "flash" rather than as optical light. As the supernova expands, it cools. For about 3 weeks the optical light curve rises as the radius of the photosphere increases, and an increasing fraction of the radiation from the cooling photosphere goes into the optical band. Maximum optical brightness occurs when the temperature is near 10,000 K. At this time the radius of the photosphere is 10^{15} cm, or 70 AU, almost twice the radius of Pluto's orbit. The matter density at the photosphere is low, near 10^{-16} g/cm^3.

After maximum light, further cooling causes the light curve to decline. Owing to the expansion and consequent geometrical dilution, the photosphere recedes with respect to the matter, i.e., matter flows through the photosphere, and deeper and deeper layers of the ejecta are gradually exposed. Eventually, at a time that depends mainly on the amount of matter ejected but that ordinarily is a matter of months, the ejected matter becomes optically thin, the continuous spectrum fades away, and the "photospheric phase" comes to an end. The supernova then enters the transparent "nebular" phase.

B. The Spectral Lines

Supernova spectral lines carry vital information on the temperature, velocity, and composition of the ejected matter. However, because of the fast ejection velocities, typically 10,000 km/sec—3% of the speed of light—the spectral features are highly Doppler-broadened and overlapping, and extracting information from the spectra is difficult.

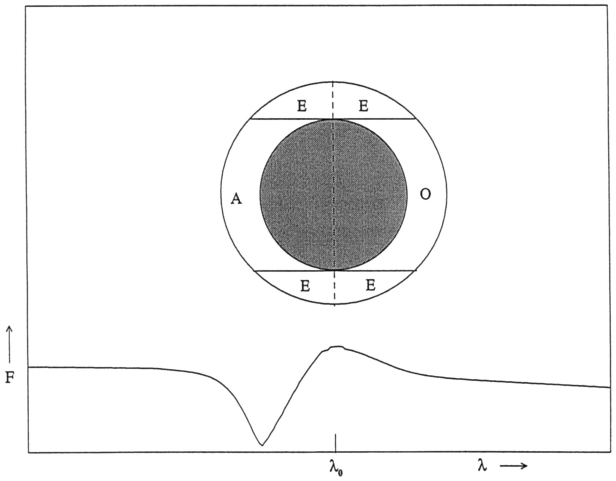


FIGURE 6 Top: A schematic drawing of the photosphere (shaded) and the surrounding atmosphere of a supernova. An observer to the left would see an emission line formed in region E and an absorption line formed in region A; region O would be occulted. Bottom: The shape (in flux versus wavelength) of a characteristic supernova spectral line is shown. [Courtesy of K. Hatano, University of Oklahoma.]

During the photospheric phase, broad absorption and emission lines form outside the photosphere and are superimposed on the continuous spectrum. A schematic model that is useful for understanding the formation of spectral lines during the photospheric phase is shown at the top in Fig. 6. The central region is the photosphere, and outside the photosphere is the line-forming region. The entire supernova is in differential expansion, with velocity proportional to distance from the center. This simple velocity law is a natural consequence of matter being ejected with a range of velocities and then coasting without further acceleration; after a time, each matter element has attained a distance from the center in proportion to its velocity.

As a photon travels through the expanding atmosphere, it redshifts in wavelength with respect to the matter that it is passing through (as a photon redshifts in the expanding universe). Thus a photon can redshift into resonance with an electronic transition in an atom or ion and be absorbed. In the very useful approximation of pure scattering, the absorbed photon is immediately reemitted, but in a random direction. From the point of view of an external observer, a photon scattered by a moving atom will be Doppler-shifted with respect to the rest wavelength of the transition. A photon scattered into the observer's line of sight from the right of the vertical line in Fig. 6 comes from

an atom that has a component of motion away from the observer and is seen to be redshifted; a photon scattered from the left of the vertical line is seen to be blueshifted. Therefore the region labeled E produces a broad, symmetrical emission line superimposed on the continous spectrum and centered on the rest wavelength of the transition. Region O is occulted by the photosphere. In region A, photons originally directed toward the observer can be scattered out of the line of sight. Because it has a component of motion toward the observer, matter in region A produces a broad, asymmetric, blueshifted, absorption component.

The shape of the full line profile is as shown at the bottom in Fig. 6. This kind of profile, having an emission component near the rest wavelength of the transition and a blueshifted absorption component, is characteristic of expanding atmospheres and is referred to as a "P Cygni profile" (after a bright star whose atmosphere is in rapid but nonexplosive expansion). The precise shape of a line profile in a supernova spectrum depends primarily on how the matter density decreases with radius and on the velocity at the photosphere. The higher the velocity, the broader the profile and the more the absorption component is blueshifted.

In supernovae the velocities are so high that line profiles usually overlap. Physically, this corresponds to multiple scattering in the atmosphere: after a photon is scattered by one transition, it continues to redshift and may come into resonance with another transition and be scattered again and again before it escapes from the atmosphere.

During the nebular phase that follows the photospheric phase, there is no continuum to serve as a source of photons to be scattered. However, an ion can be excited by a collision with a free electron and respond by emitting a photon. Thus the spectrum during the nebular phase consists of broad, symmetric, overlapping emission lines.

C. Composition

The spectrum of a SN II near the time of its maximum brightness is almost a smooth featureless continuum. Subsequently, as the temperature falls and the ionized hydrogen begins to recombine, the Balmer lines of hydrogen strengthen. Further cooling causes the spectrum to become much more complicated as many additional lines develop, from ions such as singly ionized calcium, singly ionized iron, and neutral sodium. Detailed analysis of the strengths of the spectral features during the early photospheric phase indicates that the composition of the outer layers of a SN II resembles that of the Sun and ordinary stars: hydrogen and helium are most abundant and only a small fraction of the matter is in the form of heavy elements. The deeper, slower-moving matter, which becomes observable at later times, shows strong lines of elements such as oxygen, calcium, and magnesium, and analysis reveals that the composition is dominated by such elements. In general, the deeper the layer, the heavier the elements of which it consists.

The spectra of SNe Ib lack hydrogen lines but they do develop strong lines of neutral helium. The composition structure is inferred to be much like that of a SN II, except for the absence of the outer hydrogen-rich layers. In SNe Ib, the temperature is not high enough to produce optical lines of helium by ordinary thermal excitation; a nonthermal source is required. This is provided by the radioactive decay of ^{56}Ni that was synthesized in the explosion, and then the decay of its daughter nucleus, ^{56}Co (Section III.A); γ rays from the radioactive decays scatter off thermal electrons to produce a population of fast (nonthermal) electrons, which in turn cause the excitation of helium. The spectra of SNe Ic do not have strong helium lines, either because the outer helium layer is absent (in which case the outer layers consist mainly of a mixture of carbon and oxygen) or because nonthermal excitation is ineffective. Otherwise the composition structure of SNe Ic resembles that of SNe II and Ib.

The spectral evolution of SNe Ia is quite different. Near the time of its maximum brightness the spectrum of a SN Ia is dominated by lines of singly ionized silicon, sulfur, and calcium. Weeks later the spectrum becomes dominated by overlapping lines from various ionization states of iron and cobalt. The inferred composition structure is very different from that of ordinary stellar matter; the outer layers of a SN Ia are a mixture of elements of intermediate mass, from oxygen to calcium, and the deeper layers consist mainly of iron and neighboring elements in the periodic table ("iron-peak" elements).

Because the line blending in supernova spectra is so severe, the process of extracting information from observed spectra usually entails calculating "synthetic spectra" of model supernovae and comparing them with observation. The parameters of the model are varied until the synthetic and observed spectra are in satisfactory agreement; at that point, the physical conditions and composition of the model are accepted as an approximation to the actual conditions and composition of the supernova. An example of a comparison of a synthetic spectrum and an observed one is shown in Fig. 7.

D. Polarization

Information about the shape of a supernova can be obtained by observing the extent to which its light is polarized, but the number of supernovae for which polarization measurements have been made is rather small. If a supernova is spherically symmetric, its light will be observed to be unpolarized, and this appears to be the case for

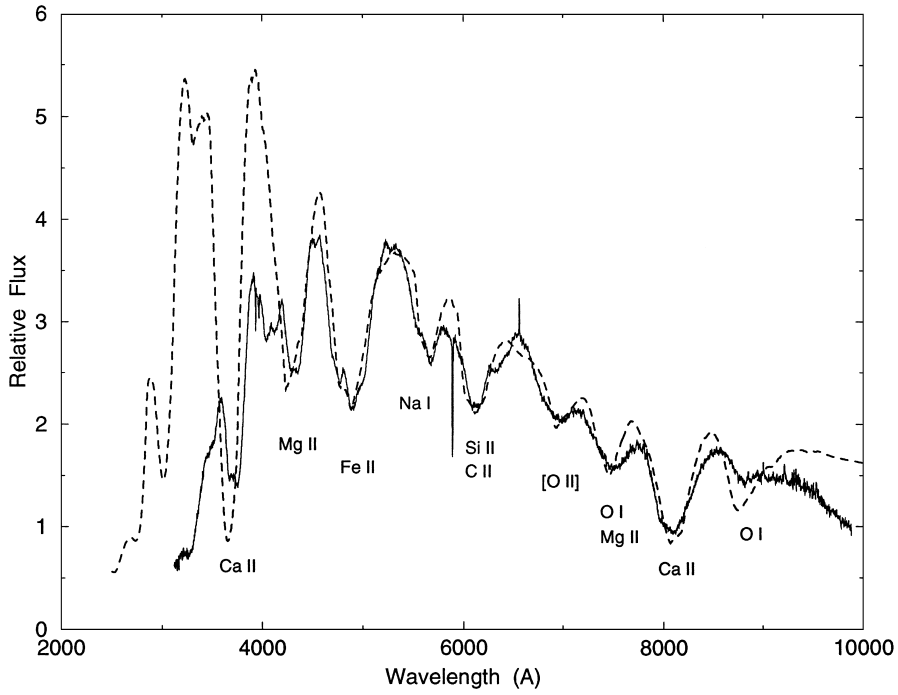


FIGURE 7 A synthetic spectrum (dashed line) is compared to an optical spectrum of the Type Ic SN 1994I obtained 4 days before the time of maximum brightness. The narrow absorption line near 5900 Å was produced by interstellar sodium atoms in the parent galaxy, M51. The synthetic spectrum has excess flux at short wavelengths because not all relevant spectral lines were included in the calculation. [Reproduced by permission from Millard, J., *et al.* (1999). *Astrophys. J.* **527,** 746.]

SNe Ia. The light from core-collapse supernovae, however, is observed to be polarized, which indicates that core-collapse supernovae are asymmetric. The observed degree of polarization, which is typically at the level of 1%, indicates that if the shape is ellipsoidal, then the axial ratio is about 1.5.

III. THERMONUCLEAR SUPERNOVAE

A. Nickel and Cobalt Radioactivity

When a star explodes its matter is heated suddenly, but expansion then begins to cause cooling, approximately adiabatically, with the temperature inversely proportional to the radius. Unless the initial radius of the progenitor star is much larger than that of the Sun, the ejected matter cools too quickly ever to attain the large (10^{15} cm), hot (10^4 K) radiating surface that is required to account for the peak optical brightness of a supernova. Some continuing source of energy must be provided to the expanding gas, to offset partially the adiabatic cooling. For SNe Ia, the continuing energy source is the radioactive decay of about 0.6 solar masse of ^{56}Ni, and then its daughter nucleus, ^{56}Co. The former has a half life of 6 days and decays by electron capture with the emission of a γ ray; the mean energy per decay is 1.7 MeV. Then ^{56}Co decays with a half-life of 77 days to produce stable ^{56}Fe, the most abundant isotope of iron in nature. The ^{56}Co decay is usually by electron capture with γ-ray emission, but sometimes it is by positron emission. The mean energy per decay is 3.6 eV, with 4%, on average, carried by the kinetic energy of the positrons.

Numerical hydrodynamical calculations have shown that the light curve of a SN Ia can be reproduced provided that a total ejected mass of somewhat more than 1 solar mass includes 0.6 solar mass of ^{56}Ni that was freshly synthesized by nuclear reactions during the explosion. During the first month after the explosion, the supernova is optically thick to the γ rays from radioactivity; they are absorbed and thermalized in the ejected matter and provide the continuing source of energy to keep it hot. As the supernova expands, the γ rays have a higher probability of escaping rather than being absorbed. The positrons from ^{56}Co decay can be trapped by even very small magnetic fields, long enough to deposit their kinetic energy into the gas by means of collisions before they mutually annihilate with electrons to produce more γ rays. The fraction of the γ-ray energy that is absorbed, together with the kinetic energy of the positrons, powers the tail of the SN Ia light curve. Because the fraction of the γ rays that is absorbed decreases with time, the optical light curve declines more rapidly than the 77-day half-life of ^{56}Co.

B. The Fate of Low-Mass Stars

The sites of SNe Ia indicate that they are produced by stars that were born with less than 8 solar masses. Such a star fuses hydrogen into helium in its core during its main-sequence phase, at a temperature of about 10^7 K. When the hydrogen in the core is exhausted, the star expands its envelope to become a red giant; during this phase the core contracts until it is able to fuse the helium to a mixture of carbon and oxygen, at 10^8 K. When helium is exhausted, the core contracts again. The carbon–oxygen core would fuse to heavier elements if the temperature approached 10^9 K, but this is not achieved because the density in the core becomes sufficiently high (10^6 g/cm^3) that gravity is balanced by the pressure of degenerate electrons—electrons that strongly resist further compression because their momentum distribution is determined by the Pauli exclusion principle. The star becomes a stable carbon–oxygen white dwarf that has a radius of only a few percent of the radius of the Sun, comparable to the radius of Earth. The maximum mass of a white dwarf is 1.4 solar masses—the Chandrasekhar mass. It appears that most, if not all, stars that are born with less than 8 solar masses manage to lose enough mass during their red giant phases, by means of stellar winds and planetary-nebula ejection, to become white dwarfs. Thus it is difficult to account for SNe Ia as the explosions of single stars.

C. Accreting White Dwarfs

A more promising explanation for the origin of SNe Ia appeals to the more complicated evolution of stars in close binary systems. A pair of stars forms, and the more massive one evolves first to become a white dwarf. At some point in the evolution of the second star, it expands and transfers matter to the surface of the white dwarf, eventually provoking it so explode. The time delay between the formation of the binary system and the explosion is determined by the initial mass of the less massive star. This model can account for the occurence of SNe Ia in elliptical galaxies (the second star has low mass, so binary systems that formed long ago are producing SNe Ia now) as well as for the correlation between the SN Ia rate and the recent rate of star formation in spirals (binaries in which the less massive star's mass is not so low also are producing SNe Ia now). A related possibility is that *both* members of the binary form white dwarfs, which gradually spiral together because of orbital decay caused by the emission of gravitational radiation, and eventually merge to temporarily assemble a super-Chandrasekhar mass. Whether such a configuration would explode as a SN Ia rather than collapse to a neutron star is not yet known. In this model, the spiral-in phase would add an additional time delay between formation and explosion.

Computer simulations of the response of a carbon–oxygen white dwarf to the accretion of matter lead to a variety of outcomes, depending on the initial mass of the white dwarf and the composition and rate of arrival of the accreted matter. Certain combinations of these parameters do lead to predictions that correspond well to the observed properties of SNe Ia. If the accretion rate of hydrogen-rich matter is of the order of 10^{-7} solar masses per year, fusion reactions in the accreted layers convert hydrogen to helium, and then helium to carbon and oxygen, thus producing an increasingly massive carbon–oxygen white dwarf. As its mass approaches 1.4 solar masses, the white dwarf contracts and heats until carbon begins to fuse near its center. The ignition of a nuclear fuel in a gas that is supported by the pressure of degenerate electrons causes a thermonuclear instability, because the increasing temperature of the nuclei fails to raise the pressure enough to cause a compensating expansion, while it does cause the fuel to burn at an increasing rate. The inner portions of the white dwarf are incinerated to a state of nuclear statistical equilibrium, becoming primarily ^{56}Ni because this is the most tightly bound nucleus that has equal numbers of protons and neutrons, and there is not enough time for weak interactions to change the proton-to-neutron ratio of the carbon–oxygen fuel. A nuclear flame front propagates outward through the white dwarf in about a second, heating and accelerating the matter, and the entire star is disrupted. The outer layers, because of their lower densities, are fused only to elements of intermediate mass, such as magnesium, silicon, sulfur, and calcium (Fig. 8). Thus the thermonuclear white dwarf model can account for the SN Ia composition structure that is inferred from spectroscopy. The uniformity, or near–uniformity, of the amount of mass ejected (1.4 solar masses) is responsible for the remarkable observational homogeneity of SNe Ia.

In this model of a SN Ia, the energy produced by fusion, less the energy needed to unbind the star, results in a final kinetic energy of about 10^{51} erg. However, for reasons given earlier, the explosion becomes optically bright only because of the smaller amount of energy that is released gradually by the radioactive decay of ^{56}Ni and ^{56}Co. In short, energy released promptly by nuclear fusion explodes the star, while energy released slowly by radioactivity makes the explosion shine. Because the final composition of a SN Ia is largely iron and neighboring elements in the periodic table, SNe Ia make an important contribution to the amount of iron-peak elements in the universe.

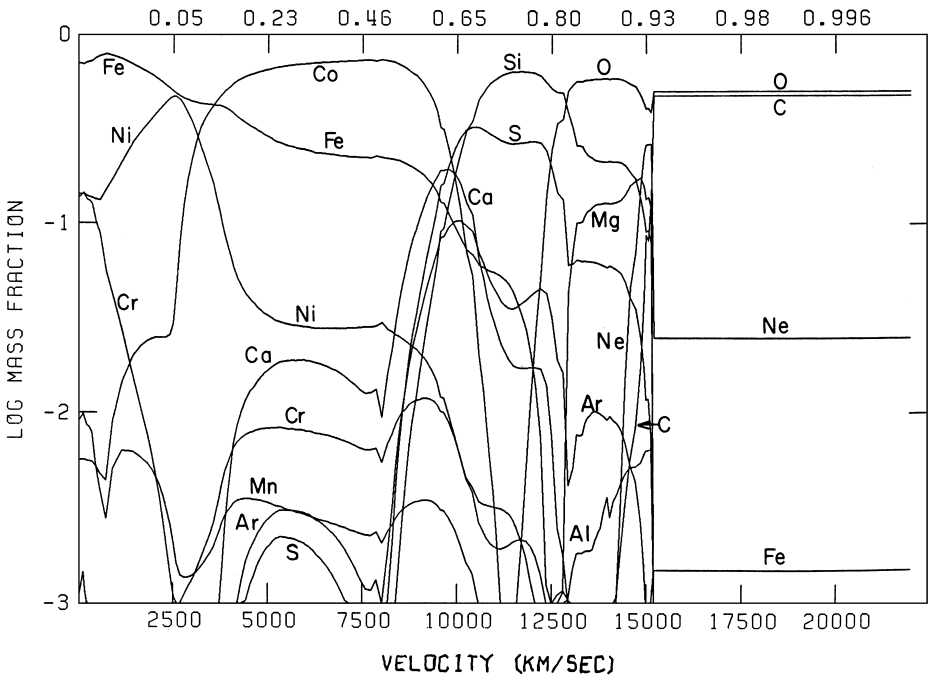


FIGURE 8 The calculated composition of an exploded white dwarf, displayed as element mass fraction plotted against ejection velocity. The fraction of the ejected mass that is interior to each velocity is shown at the top. This figure is for 32 days after the explosion; the composition below 10,000 km/sec changes with time owing to radioactive decays, and most of the cobalt eventually decays to iron. [Reproduced with permission from Branch, D., Doggett, J. B., Nomoto, K., and Thielemann, F. K. (1985). *Astrophys. J.* **294,** 619.]

IV. CORE-COLLAPSE SUPERNOVAE

A. Supergiant Progenitors

SNe II, SNe Ib, and SNe Ic occur in spiral arms and therefore come from massive stars. In its late evolution, a star having an initial mass in the range of 8 to about 30 solar masses becomes a red supergiant, developing a compact, dense, heavy-element core surrounded by an extended hydrogen-rich envelope that may swell to 1000 times the solar radius, approaching 10^{14} cm. If an instability in the core suddenly releases a large amount of energy, a shock wave may form and propagate outward, heating and ejecting the star's envelope. As it expands, the envelope cools nearly adiabatically. Nevertheless, because the initial radius of a red supergiant is so large, the envelope can attain the large, hot radiating surface that is required to produce the peak optical brightness of a supernova, without requiring a delayed input of energy from radioactivity. Hydrodynamical computer simulations of the response of a red supergiant to a sudden release of energy at its center predict light curves and other properties that are in good agreement with observations of SNe II-P, the most common kind of SN II. The light-curve plateau is a phase of diffusive release of thermal energy deposited in the envelope by the shock. To account for the observed peak luminosity, the expansion velocity, and the duration of the plateau phase, the progenitor needs to have a large radius and eject roughly 10 solar masses that carries a kinetic energy of 10^{51} erg. After the plateau phase, the light-curve tail is powered by ^{56}Ni and ^{56}Co decay. The γ rays are efficiently trapped and thermalized by the large ejected mass of a SN II-P, so the decline rate of the optical tail corresponds closely to the ^{56}Co half-life of 77 days. To account for the light curve of a SN II-L, the ejected mass should be only a few solar masses, and as in a SN Ia, γ rays increasingly escape so the tail declines somewhat faster than the 77-day half-life.

The progenitor of a SN Ib is a massive star that loses its hydrogen-rich envelope prior to core collapse, either by means of a stellar wind (if the initial mass is more than about 30 solar masses) or by mass transfer to a binary companion star. A star that lacks a hydrogen envelope has a smaller radius than a red supergiant, so the SN Ib light curve, like that of a SN Ia, is powered primarily by radioactive decay. Because the amount of ^{56}Ni that is ejected by an SN Ib is typically only 0.1 solar mass, a SN Ib is not as bright as a SN Ia. Similar statements apply to the progenitors of SNe Ic. Some of the SN Ic progenitors may be bare carbon–oxygen cores that lose most or all of

their helium layers prior to collapse, and others may eject helium but lack the nonthermal excitation that is needed to produce helium lines in optical spectra.

B. Collapse and Explosion

The basic explosion mechanism of SNe II, Ib, and Ic is the same. The sudden energy release at the center of the progenitor star is caused by the inevitable gravitational collapse of its highly evolved core. In massive stars, carbon is nondegenerate when it ignites, and it fuses nonexplosively to a mixture of oxygen, neon, and magnesium. Subsequent nuclear burning of these and heavier elements is unable to disrupt the increasingly tightly bound stellar core, and the core eventually becomes iron (Fig. 9). Iron is the ultimate nuclear "ash," from which no nuclear energy can be extracted. Instead, high-energy photons characteristic of the high core temperature photodisintegrate the iron endothermically, decreasing the energy and pressure in the core and causing it to collapse at a fraction of the speed of light, from the size of a planet to a radius of 50 km. During the collapse electrons are forced into nuclei to form a fluid of neutrons, and as the density exceeds the nuclear density (more than 10^{14} g cm^{-3}) the pressure rises enormously to resist further compression.

The collapse releases a gravitational potential energy of more than 10^{53} erg. The resulting temperatures and densities are so high that neutrinos are created copiously, and because they interact with matter only weakly, they diffuse out of the core in seconds and carry away almost all of the energy. To produce a supernova, however, about 1% of the gravitational energy must somehow go into ejecting the matter outside the core. Just how this occurs is not yet understood. One possibility is a "prompt" hydrodynamical explosion. Owing to its infall velocity, the collapsing core overshoots its equilibrium density, rebounds ("bounces"), transfers energy mechanically to matter that is falling in more slowly from above, and produces an outgoing shock wave. The energy in the shock tends to be consumed, however, in photodisintegrating heavy elements just outside the core, causing the shock to stall before it can eject the outer layers of the star. A related possibility is a "delayed" hydrodynamical explosion, in which the small fraction of neutrinos that are absorbed in the matter just outside the core deposits enough energy to revive the stalled shock. Or it may be that the explosion mechanism is fundamentally different, e.g., a magnetohydrodynamical event in which the increasingly rapid rotation and the amplifying magnetic field of the collapsing core combine to cause the ejection of relativistic jets of matter along the rotation axis, which drive shocks into the envelope and cause ejection. This could be the reason for the significant departures from spherical symmetry that are implied by the observed polarization of the light from core collapse supernovae. It

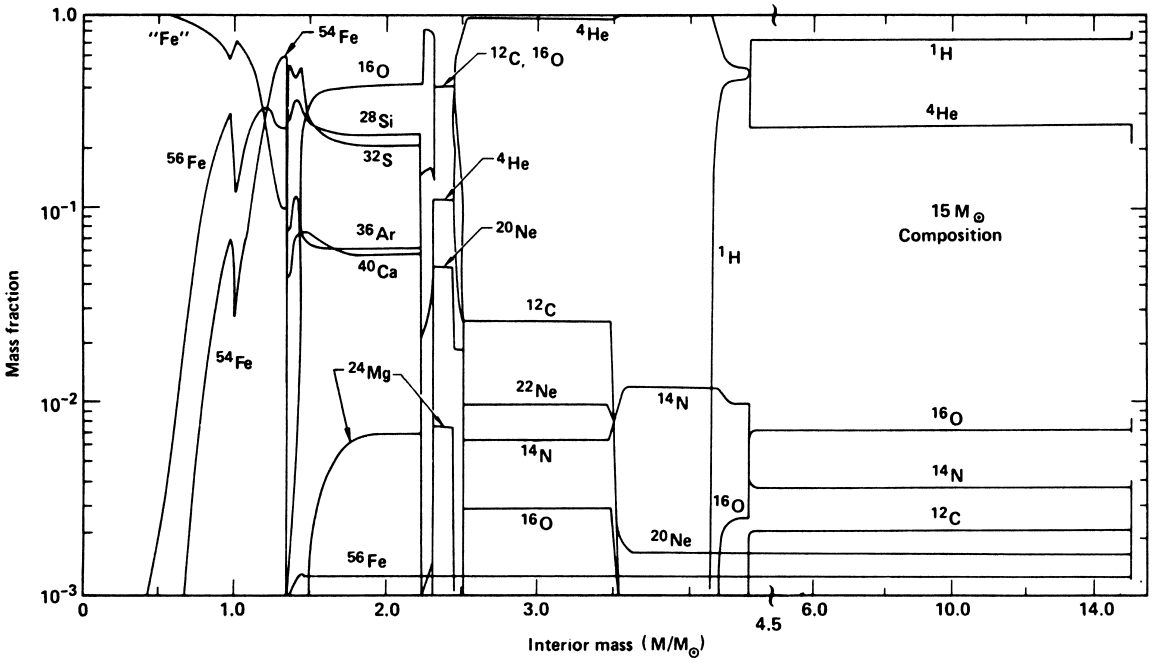


FIGURE 9 The calculated composition of a star of 15 solar masses, just before core collapse. The amount of mass interior to each point is plotted on the horizontal axis; note the change of scale at 4.5 solar masses. [Courtesy of S. E. Woosley, University of California at Santa Cruz, and T. A. Weaver, Lawrence Livermore National Laboratory.]

also could be related to the association of at least some core-collapse supernovae with at least some γ-ray bursts, which became apparent with the spatial and temporal coincidence of the peculiar Type Ic SN 1998bw and the γ-ray burst of April 25, 1998. In SN 1998bw and several other core-collapse events, the kinetic energy carried by the ejected matter appears to have been more than 10 times that of a typical supernova. Such "hyperenergetic" events may be produced by the collapse of the cores of very massive stars (more than 30 solar masses) to form black holes of about 7 solar masses.

Shock waves raise the temperature of the deep layers of the ejected matter high enough to cause fusion reactions. As a result of this "explosive nucleosynthesis" the innermost layers of the ejected matter are fused to ^{56}Ni and other other iron-peak isotopes, and the surrounding layers are fused to elements of intermediate mass. The low-density outer layers of the progenitor star—whether hydrogen, helium, or carbon and oxygen—are not fused. Core-collapse events are the main producers of intermediate-mass elements in nature, and they also make a significant contribution to the production of iron-peak elements.

Whether the collapsed star ends as a neutron star or a black hole depends on whether its final mass exceeds the maximum mass of a neutron star.

C. Circumstellar Interaction

The progenitor of a core-collapse supernova loses mass by means of a stellar wind. For a steady-state wind that has a constant velocity and a constant mass-loss rate, the density of the circumstellar matter is proportional to the mass-loss rate, inversely proportional to the wind velocity, and inversely proportional to the square of the distance from the star. The wind velocity ordinarily is comparable to the escape velocity from the photosphere of the star—of the order of 10 km/sec for a red supergiant and 1000 km/sec for a blue one—so the circumstellar matter of a red supergiant is much more dense than that of a blue one. When a core-collapse supernova occurs, both the ejected matter and the radiated photons interact with the circumstellar matter. The violent collision between the high-velocity ejecta and the circumstellar matter is called a hydrodynamic interaction. The interaction between the supernova photons and the circumstellar matter is called a radiative interaction. These interactions can have observable consequences in various parts of the electromagnetic spectrum.

Hydrodynamic circumstellar interaction has been detected most often by observations at radio wavelengths. The spectrum of the radio emission indicates that it is synchrotron radiation from electrons that are accelerated to relativistic velocities in a very hot (up to 10^9 K) interaction region. The higher the density of the circumstellar matter, the longer the radio "light curve" takes to rise to its peak, because initially the circumstellar matter outside the interaction region, which has been ionized by the X-ray and ultraviolet flash, absorbs the radio photons by free–free processes. Thermal X-rays from the hot interaction region also have been detected in a smaller number of cases. The fraction of a supernova's total luminosity that is emitted in the radio and X-ray bands ordinarily is very small, but observation of the radio and X-ray emission combined with modeling of the interaction provides valuable information on the mass-loss rate of the progenitor star. Rates in the range of 10^{-6} to 10^{-4} solar masses per year have been inferred. Optical light radiated from the interaction region is responsible for the slow decay rate of the light curves of SNe IIn. Unlike most supernovae, which seldom are followed observationally for more than a year after explosion, some circumstellar-interacting supernovae are observed decades after explosion, at optical, radio, and X-ray wavelengths.

Radiative interactions are manifested most clearly by relatively narrow emission and absorption lines that appear in optical spectra (which cause the supernova to be classified Type IIn) and, especially, in ultraviolet spectra [which can only be obtained from above Earth's atmosphere with instruments such as the *Hubble Space Telescope* (*HST*); Fig. 10]. The widths of the circumstellar lines provide information on the wind velocity of the progenitor, and analysis of the line strengths provides information on the relative abundances of the elements in the wind. The ionization state of the circumstellar matter also provides information on the amount of ionizing radiation that was emitted by the explosion, something that is not observable directly.

Some supernovae emit excess infrared radiation, well beyond the amounts expected from their photospheres, by thermal radiation from cool (1000 K), small (10^{-5} cm) solid particles (dust grains) in the circumstellar medium. The grains are heated by absorbing optical and ultraviolet radiation from the supernova and respond by emitting in the infrared. Infrared observations provide information on the nature and spatial distribution of the dust grains.

D. Supernova 1987A

SN 1987A was discovered during the predawn hours of February 24, 1987, as a star of the fifth apparent magnitude in place of a previously inconspicuous twelfth-magnitude star known as Sanduleak (Sk) −69 202 (star number 202 near −69° declination in a catalog of stars published by N. Sanduleak in 1969). Sk −69 202 was the first supernova progenitor star whose physical properties could be determined from observations that had been made prior to

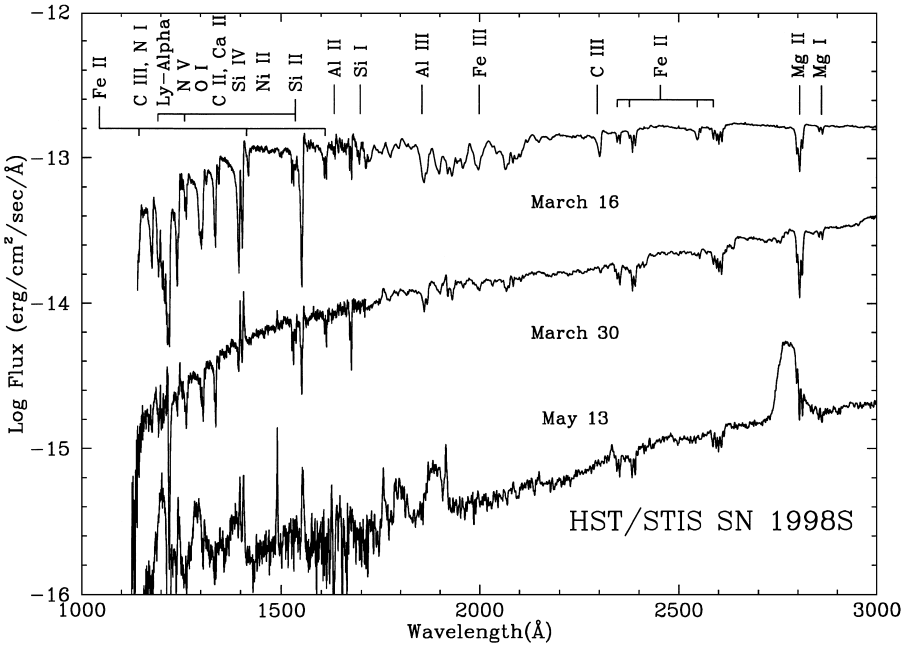


FIGURE 10 Ultraviolet spectra of the Type IIn SN 1998S obtained with the *Hubble Space Telescope*. [Courtesy of P. Challis, Harvard–Smithsonian Center for Astrophysics and Supernova INtensive Study (SINS) team.]

its explosion. It was a supergiant star of spectral type B3, with an effective temperature of 16,000 K and a luminosity 10^5 times that of the Sun. The luminosity indicates that the mass of its core of helium and heavier elements was 6 solar masses, which in turn implies that the initial mass of the star was about 20 solar masses. By the time of the explosion, presupernova mass loss had reduced the mass to about 14 solar masses. From the luminosity and effective temperature, the radius of Sk −69 202 was inferred to be 40 times that of the Sun.

As discussed above, intrinsically luminous SNe II are the explosions of red supergiants, which have very large radii. It is well understood why an explosion of a less extended star such as Sk −69 202 would produce a subluminous supernova (see Section IV.A). A more difficult question is, Why did Sk −69 202 explode as a relatively compact blue supergiant rather than as a very extended red one? Observations of the circumstellar matter of SN 1987A indicate that its progenitor did go through a red supergiant phase, but it ended some 40,000 years prior to the explosion. There are several possible reasons why the progenitor star contracted and heated to become a less extended blue supergiant, including the possibility that the evolutionary history of Sk −69 202 involved a *merger* of two stars in a close binary system.

Within days of the optical discovery of SN 1987A, it was realized that a neutrino burst had been recorded by undergound neutrino detectors in Japan and the United States, a few hours before the optical discovery. Because neutrinos interact only weakly with matter (each person on Earth was harmlessly perforated by 10^{14} neutrinos from SN 1987A), only a total of 18 neutrinos, with arrival times spread over 12 sec, was detected with certainty. The total neutrino energy emitted by SN 1987A was 2×10^{53} erg, and the mean neutrino energy was 12 MeV, as expected, so the neutrino signal confirmed the basic theory of core collapse. It also provided valuable information for particle physics, including an upper limit of 16 eV to the mass of the electron neutrino. The 12-sec duration of the neutrino burst suggested that the core initially formed a neutron star, because black-hole formation would have terminated the burst earlier. However, the neutron star has not yet been detected directly. It is possible that "fall-back" of matter that was ejected very slowly has caused the neutron star to collapse to a black hole.

The spectrum of SN 1987A has been observed far into the nebular phase, and over a very broad range of wavelengths. Analysis of optical and infrared spectra of SN 1987A has established that, as expected, the composition varied from primarily hydrogen and helium in the outermost layers to mainly heavier elements in the innermost layers. The brightness during the tail of the light curve indicated that the amount of ejected ^{56}Ni was 0.07 solar mass. The optical and infrared spectra, as well as the unexpectedly early leakage of γ rays and X-rays (produced by γ rays that transferred much of their energy to electrons), indicated that a sustantial amount of composition mixing (light elements into the deeper layers and heavier

elements, including ^{56}Ni, into the outer layers) took place. The element mixing process still is not well understood. An excess of infrared radiation that began to develop a year after explosion indicated that a small fraction of the heavy elements in the ejected matter had condensed into dust grains.

SN 1987A showed various signs of early circumstellar interaction, at ultraviolet and radio wavelengths, but the interaction was weak because the wind of a blue supergiant is fast and the circumstellar density is correspondingly low. After a few years, however, it became clear that SN 1987A was at the center of a ring of higher density matter, which appears elliptical on the sky but actually is circular, and tilted with respect to our line of sight by 42°. The matter in the ring, which is expanding at 10 km/sec, presumably originated from the slow, dense wind of Sk −69 202 when it was in its red supergiant phase. The reason that it is confined to a ring rather than a spherical shell is not well understood. The later discovery of two larger, fainter rings (Fig. 11), as well as additional complexity in the circumstellar environment, is suggestive of a complicated binary-star history of Sk −69 202 involving interacting stellar winds. The time variation of ultraviolet emission lines emitted by ions in the inner ring, as they recombined following ionization by the supernova radiation, indicated that the radius of the ring is 0.6 light year; when combined with the observed angular size of the ring (0.9 sec of arc), this provides a valuable independent confirmation that the distance to the LMC is 50 kpc. In the late 1990s, the first signs of an imminent collision between the high-velocity supernova ejecta and the slowly expanding inner ring began to appear. This interaction will cause the emission from SN 1987A to increase dramatically, across the electromagnetic spectrum, and provide an opportunity to probe further the distribution of the circumstellar matter of SN 1987A.

V. SUPERNOVA REMNANTS

The matter ejected by a supernova sweeps up interstellar gas. After a time of the order of a century, the mass of the swept-up matter becomes comparable to the mass of the ejected matter, and the latter begins to decelerate, with energy being converted from kinetic into other forms. Thus, a hot region of *interstellar* interaction develops, and an

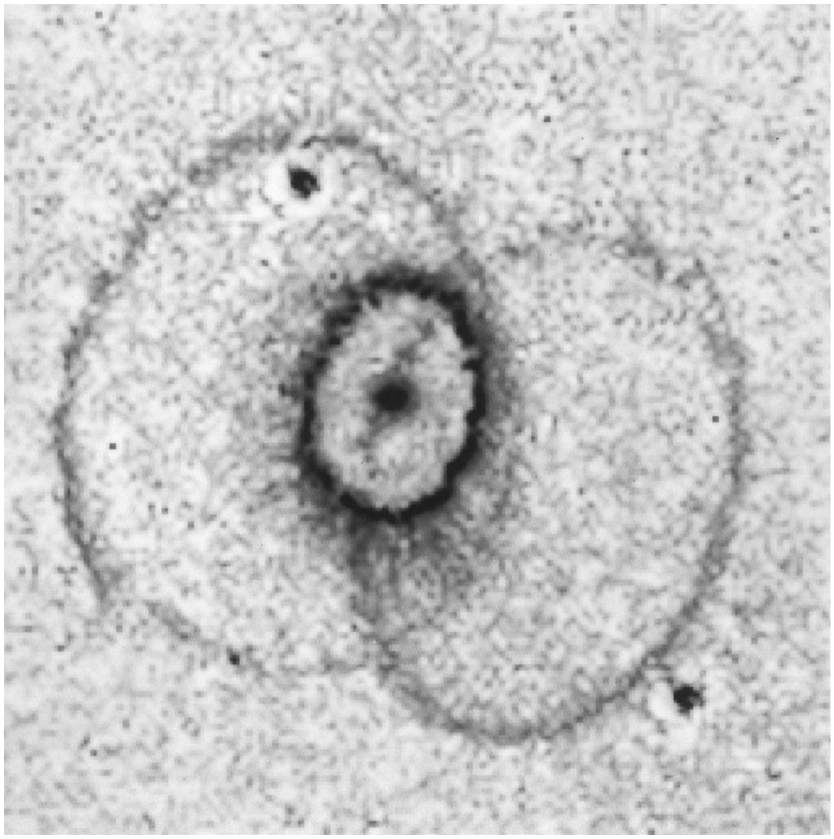


FIGURE 11 A *Hubble Space Telescope* image of SN 1987A (center), its inner ring, and its two fainter outer rings. [Reproduced by permission from Burrows, C. J., *et al.* (1995). *Astrophys. J.* **452,** 680.]

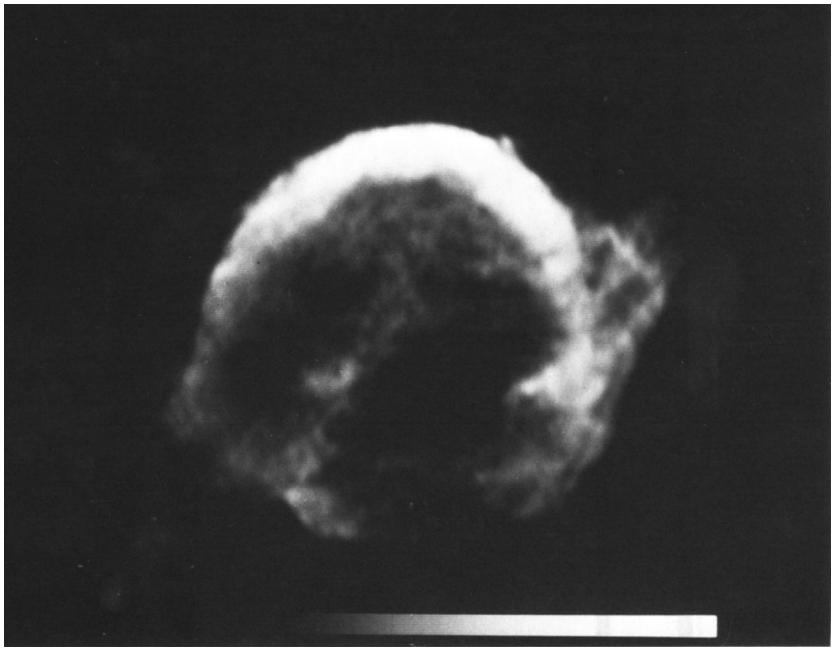


FIGURE 12 A radio image of the shell-type remnant of Kepler's supernova of 1604. The image was made at a wavelength of 20 cm using the Very Large Array radiotelescope at the National Radio Astronomy Observatory. [Reproduced by permission from Matsui, Y., *et al.* (1984). *Astrophys. J.* **287,** 295.]

extended "shell-type" supernova remnant (SNR) forms. The remnants of the supernovae of 1006, 1572 (Tycho Brahe's supernovae), 1604 (Johannes Kepler's), and 1680 (Cas A) are well-observed examples of shell-type SNRs (Fig. 12).

Galactic supernova remnants are among the brightest radio sources on the sky. The emission is synchrotron radiation produced by relativistic electrons that are accelerated in the hot interaction region. Because radio waves are unaffected by interstellar dust grains, radio SNRs are useful for statistical studies of the spatial distribution and rate of supernovae in the Galaxy. Thermal X-rays also are detected from some SNRs, but most SNRs are not detected in this way because of absorption by interstellar matter; the X-ray spectra provide important information on the composition of the ejected matter.

The famous Crab Nebula, the remnant of the supernova of 1054, is the prototype of a less common kind of supernova remnant, called a "filled center" SNR, or a *plerion* (from the Greek word for full). The emission from the Crab includes spectral lines from matter that is expanding at only about 1000 km/sec. Another component of emission, which extends from the radio through the optical to the X-ray region, is nonthermal synchrotron radiation. This indicates the presence of highly relativistic electrons and the need for continued energy input into the central regions of the remnant; the source of this is a pulsar that rotates 30 times per second and generates the relativistic electrons. Searches for signs of high-velocity matter associated with the Crab Nebula have not yet been successful.

Some SNRs also can be detected in relatively nearby external galaxies, but because of their small angular sizes they cannot be studied in as much detail.

VI. SUPERNOVAE AS DISTANCE INDICATORS FOR COSMOLOGY

A. The Extragalactic Distance Scale

The universe is expanding. The radial velocity of a galaxy relative to us is proportional to the distance of the galaxy from us; thus the cosmic expansion can be represented by the "Hubble law": $v = H_0 d$, where v is the radial velocity (ordinarily expressed as km/sec), d is the distance (as Mpc), and H_0 is the Hubble constant (as km/sec/Mpc) at the present epoch. The reciprocal of the Hubble constant gives the Hubble time—the time since the Big Bang origin of the expansion, assuming deceleration to have been negligible. Radial velocities are easily measured, so determining the value of H_0 reduces to the problem of measuring distances to galaxies. In practice, because galaxies have random motions, it is necessary to determine distances to fairly remote galaxies, at distances of hundreds of megaparsecs. Measuring the distances to supernovae provides an alternative to the classical approach of estimating the distances to galaxies themselves.

There are several ways to estimate distances to supernovae that require some physical understanding of the events but that have the advantage of being independent of all other distance determinations in astronomy. For example, the observed properties of SNe Ia indicate that they eject about 0.6 solar mass of ^{56}Ni, so the peak luminosity of an SN Ia can be calculated from the decay rate of ^{56}Ni and compared with the observed flux to yield the distance. Alternatively, the luminosity of a supernova can be calculated from its temperature (as revealed by its spectrum) and the radius of the photosphere (as the product of the expansion velocity and the time elapsed since explosion). Both methods usually have given longer distances and therefore lower values of H_0, near 60 km/sec/Mpc, than most other methods.

Another important approach takes advantage of the near homogeneity of the peak absolute magnitudes of SNe Ia, to use them as "standard candles" for distance determinations. The top panel in Fig. 13 illustrates the near-homogeneity, and the bottom panel shows how well SNe Ia can be standardized even further by correcting the peak magnitudes with the help of an empirical relation between the peak brightness and the rate of decline of the light curve. The *HST* has been used to find Cepheid variable stars in nearly a dozen galaxies in which SNe Ia were observed in the past. The period–luminosity relation for Cepheids gives the distance to the parent galaxy; the distance and the apparent magnitude of the supernova gives the absolute magnitude of the supernova ($M_V = -19.5$); and that absolute magnitude, when assigned to SNe Ia that appear in remote galaxies, gives their distances. In this way, H_0 has been determined to be about 60 km/sec/Mpc, in good agreement with the supernova physical methods. The corresponding Hubble time is 17 Gyr.

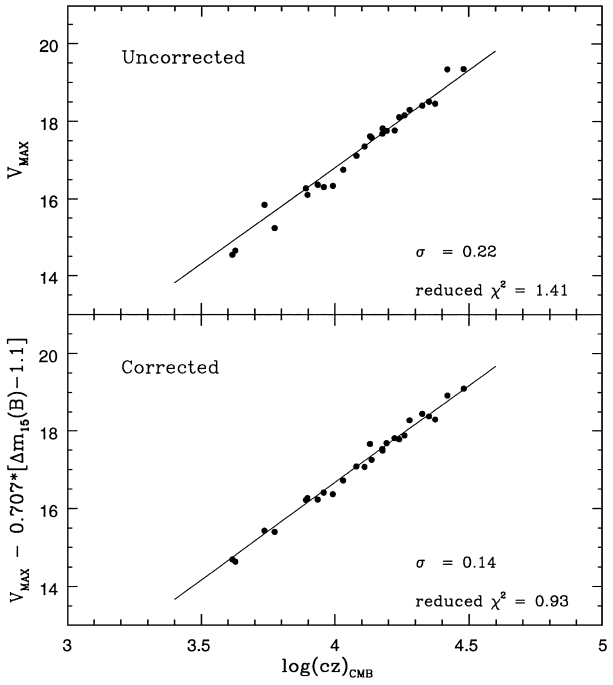


FIGURE 13 A "Hubble diagram" for a sample of well-observed SNe Ia. Top: The apparent visual magnitude of the supernova at maximum brightness is plotted against the logarithm of the parent-galaxy radial velocity (km/sec). If SNe Ia had identical peak absolute magnitudes, the points would fall on the straight line. Bottom: A correction for a relation between the peak magnitudes and the light-curve decay rate has been applied, to reduce the scatter about the straight line. [Reproduced by permission from Hamuy, M., *et al.* (1996). *Astron. J.* **112,** 2398.]

B. Deceleration and Acceleration

The self-gravity of the mass in the universe causes a deceleration of the cosmic expansion. If the matter density and the deceleration exceeded a critical amount, the expansion eventually would cease, and be followed by contraction and collapse. A way to determine the amount of deceleration that actually has occurred is to compare the expansion rate in the recent past, based on observations of the "local" universe, to the expansion rate in the remote past, based on extremely distant objects ("high-redshift" objects, because a large distance means large radial velocities and Doppler shifts toward longer wavelengths). Because SNe Ia are such excellent standard candles, and because they are so luminous that they can be seen at high redshifts, they are suitable for this task. This has been the main motivation for searching for batches of remote SNe Ia, beginning in the mid-1990s. Results based on the first dozens of high-redshift SNe Ia to have been discovered (which actually occurred deep in the past, when the distances between galaxies were only one-half to two-thirds of what they are now) indicate that deceleration is insufficient to bring the expansion to a halt. Evidently, the universe will expand forever. In fact, the SN Ia data indicate that although the cosmic expansion was decelerating in the past, it is *accelerating* now. An acceleration would require the introduction of Einstein's notorious "cosmological constant," or some other even more exotic agent that would oppose the deceleration caused by gravity. The acceleration is inferred from the observation that high-redshift SNe Ia are 20 to 50% dimmer, observationally, than they would be in the absence of acceleration. Observations of many more SNe Ia, and an improved understanding of the physics of SNe Ia, are needed to draw a firm conclusion that acceleration really is occurring.

VII. PROSPECTS

Our understanding of supernovae began to develop only after about 1970. Observations with modern detectors,

primarily in the optical region but increasingly in other wavelength bands as well, provided high-quality data for comparison with the results of detailed numerical models. A basic understanding of the nature of supernova spectra and light curves, and detailed computer simulations of the explosions, emerged from the interplay between observation and theory. SN 1987A attracted much attention to the study of supernovae. During the 1990s dramatic progress was made in using SNe Ia to measure the Hubble constant and to probe the cosmic deceleration, and the discovery of an apparent acceleration has caused excitement among cosmologists and particle physicists, and even more interest in supernovae.

Further important observational advances are forthcoming. Ground-based optical astronomy will contribute automated searches with supernova-dedicated telescopes, to increase the discovery rate greatly. New powerful instruments such as the *Next Generation Space Telescope* (the *NGST*, successor to the *HST*), and perhaps even a large orbiting supernova-dedicated observatory, will provide observations of both nearby and very remote events. New sensitive instruments also will detect large numbers of neutrinos, and perhaps even gravitational waves, from supernovae in our Galaxy and the nearest external galaxies. The flow of observational data will stimulate further theoretical modeling. Improved physical data will be fed into new generations of fast computers to produce increasingly realistic models of the explosions. Astronomers will intensify their efforts to understand better which kinds of stars produce supernovae, and how they do it; to evaluate more precisely the role of supernovae in the nuclear evolution of the cosmos; and to exploit supernovae to probe the history and future expansion of the universe with ever-greater precision. These are grand goals—and they can be achieved.

SEE ALSO THE FOLLOWING ARTICLES

COSMOLOGY • GAMMA-RAY ASTRONOMY • GRAVITATIONAL WAVE ASTRONOMY • INFRARED ASTRONOMY • NEUTRINO ASTRONOMY • NEUTRON STARS • STELLAR SPECTROSCOPY • STELLAR STRUCTURE AND EVOLUTION • ULTRAVIOLET SPACE ASTRONOMY

BIBLIOGRAPHY

Arnett, D. (1996). "Supernovae and Nucleosynthesis," Princeton University Press, Princeton NJ.

Barbon, R., Buondi, V., Cappellaro, E., and Turatto, M. (1999). *Astron. Astrophys. Suppl. Ser.* **139,** 531.

Branch, D. (1998). *Annu. Rev. Astron. Astrophys.* **36,** 17.

Filippenko, A. V. (1997). *Annu. Rev. Astron. Astrophys.* **35,** 309.

Mann, A. K. (1997). "Shadow of a Star: the Neutrino Story of Supernova 1987A," W. H. Freeman, San Francisco.

Ruiz–Lapuente, P., Canal, R., and Isern, J. (eds.) (1997)."Thermonuclear Supernovae," Kluwer, Dordrecht.