The formation of galaxies

Investigation of how these luminous islands evolved and acquired their current forms may lead to insights into the large-scale structure of the universe as well.

Joseph Silk

Galaxies are beacons that cosmologists use to chart the universe out to its remotest depths. Yet unlike a lighthouse, built on a sturdy pillar of rock, a galaxy resembles an island of light constructed on quicksand in a continuous state of flux. Galaxies are not isolated objects, but are surrounded by close companions, with whom they often maintain intimate and ongoing relationships. The oldest stars in a galaxy are nearly as old as the universe itself, and a study of their properties yields clues about how galaxies evolved in the distant past. The galaxies themselves define space in the expanding universe. In this article, I will describe current ideas on how galaxies originated and acquired their observed forms, and how this knowledge promises to lead to insights into the evolution of the large-scale structure of the universe.

Galaxy formation

A galaxy is a star pile, with the stars orbiting in their mutual gravitational potential well. The average luminosity of a galaxy is about 4×10^{10} solar luminosities. The most numerous galaxies are the least luminous: the dwarf galaxies, which range in luminosity down to about $10^5\ L_\odot$. The most extreme dwarfs have very low surface brightness and are difficult to distinguish against the sky background. The luminous galaxies, on the other hand, have high surface brightness, and in the nearer systems one can often see magnificent regularities: elliptical galaxies, which are beautiful glowing spheroids of stars, and spiral galaxies, which consist of lesser spheroids surrounded by an extensive disk of stars and gas in which the spiral density pattern is actively being generated.

What determines the mass of a galaxy? What determines its shape—whether disk or spheroid, or a combination of the two?

A galaxy, by definition, consists predominantly of stars, so arguments about the rate and efficiency of star formation inevitably play a critical role in these issues. Unfortunately, star formation is not well understood, even in our own galaxy and at the present epoch. However, one can at least establish a phenomenological set of parameters to characterize star formation, and this provides sufficient ammunition for addressing the galaxy formation problem.

For a protogalactic cloud to fragment into stars requires both high density and low temperature, and efficient radiative cooling is essential if stars are to form rapidly in the protogalaxy. The mass of a galaxy may thus be considered to result from a coincidence between two time scales: the time t_g for gravitational collapse of the gas cloud and the time t_d for energy dissipation in the cloud. If the initial collapse takes place much faster than the cloud can dissipate energy, the free-fall collapse of the protogalactic cloud is halted by thermal pressure of the hot gas. When the mean density is low, star formation is not very efficient. Only if t_d is much less than t_g can energy be efficiently radiated and collapse proceed to very high density within a free-fall time. In this case, star formation is highly efficient.

Once stars form, they tend to retain their initial orbital energies and angular momenta, and their distribution defines the scale of the final galaxy. An estimate of this scale is therefore given by comparing t_g and t_d for a sequence of diffuse gas clouds at thermal equilibrium—that is, at the "virial" temperature, at which the kinetic (thermal) energy is half the potential (gravitational) energy. The dissipation time t_d is determined by radiative processes, and is therefore given in terms of the fine-structure constant α, or $1/137$, and other fundamental constants; the collapse time t_g is similarly given in terms of the gravitational constant G. The ratio of the two time scales involves the dimensionless gravitational coupling constant

$$\alpha_g = Gm_p{}^2/\hbar c = 6\times10^{-39}$$

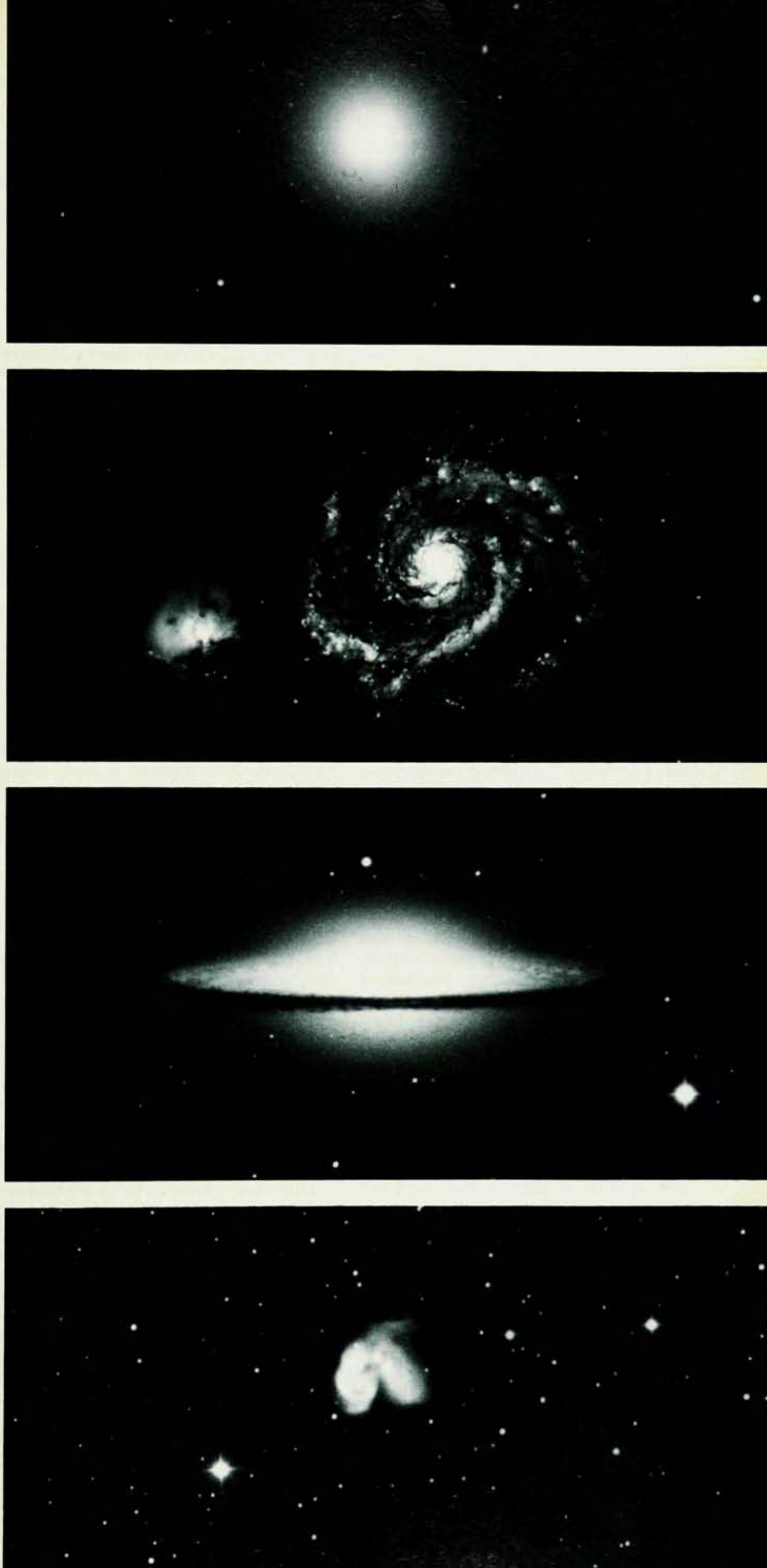
These time scales depend on the density and temperature in such a way that the condition that t_d be less than t_g for a primordial self-gravitating gas cloud of temperature greater than a rydberg is equivalent to the condition that the cloud mass not exceed

$$(m_p/m_e)^{1/2}\alpha^5/\alpha_g{}^2 \simeq 3\times10^{67} \text{ baryons}$$

To within factors of order unity this is precisely the luminous mass of a bright galaxy. A somewhat different argument may be applied to an opaque lump of matter, for which the require-

Joseph Silk is professor of astronomy at the University of California, Berkeley. His research interests include the interface of particle physics and cosmology.

0031-9228 / 87 / 0400-28-08 / $01.00



Galaxies come in a wide variety of sizes and shapes, depending on conditions of the universe in their vicinity at the time of their formation. From top to bottom: NGC 4486, an elliptical galaxy in Virgo; NGC 5194, the Whirlpool galaxy in Canes Venatici; NGC 4594, a spiral galaxy in Virgo—the dark band in this edge-on view is a result of absorption by gas and dust; NGC 4038 and 4039, a pair of interacting galaxies in Corvus. All photos are from the Palomar 200-inch telescope.

ment that gravity stably balance pressure (either thermal pressure or the pressure due to the degeneracy of a fermion system—as in a white dwarf) yields a characteristic mass of $\alpha_g^{-3/2}$, that is, about 10^{57} baryons, or roughly the mass of a star, $M_\odot$. It appears that astrophysical arguments can go a long way toward explaining the mass hierarchy of the astronomical universe!

As I have said, structures that formed by dissipation should have densities and temperatures that allow cooling within the structure's dynamical time scale. One can make this dissipation argument slightly more precise by using the actual cooling function for a primordial cloud of galactic mass. Consider an ensemble of clouds orbiting in a protogalaxy. The clouds occasionally collide, dissipate bulk kinetic energy in shocks and form stars, and the protogalaxy contracts to develop eventually into a normal galaxy. The gas density has to increase by some 20 orders of magnitude or more for stars to form. For star formation to be an efficient process, the shocks must be radiative, thereby guaranteeing strong cooling, large post-shock density enhancement and likely fragmentation into stars. The condition for a planar shock of speed V_s to be radiative can be expressed as

$$V_s\, t(n, V_s) < L$$

Here L is the cloud size, and the post-shock cooling time scale t depends on the cloud density n and the shock velocity. Because t is inversely proportional to n, this condition is equivalent to setting a lower bound on the surface density of the cloud expressed as a function of shock velocity, or equivalently, on the virial velocity dispersion σ_v of the protogalaxy. For cloud collisions to occur efficiently—that is, for a typical cloud to undergo a collision within a protogalaxy crossing time—the mean surface density of the protogalaxy must be at least comparable to that of an individual cloud.

The old spheroidal components of galaxies must have formed in these cloud shocks during the initial protogalactic collapse, and because stellar systems are at present evolving very slowly (the two-body relaxation time in a system of N stars is roughly $N/\log N$ crossing times) the mean density and stellar velocity dispersion in the old spheroidal population of stars represent fossilized conditions attained when the galaxy formed. The mean volume density observed in these star populations should therefore exceed the critical density given by the condition for a radiative shock and lie within the allowed region where cooling is efficient.

Indeed, estimates of the mean volume densities of the old spheroidal components based on their luminosities, velocity dispersions or rotation velocities and their effective radii lie well within the dissipative regime. Galaxies have evidently undergone considerable dissipation during the process of acquiring their observed binding energies. This has also been recognized in purely gravitational N-body simulations of elliptical galaxy formation, which fail to reproduce the high core densities or the concentration of metal-rich stars observed in the inner regions. Including gaseous dissipation in the simulations provides a means of accounting for these properties, but results in a far more complicated model—whose details are still poorly understood.

The observed surface densities and temperatures of galaxy clusters and groups, on the other hand, lie in the nondissipative region of the density–temperature parameter space. This tells us that on the scale of a galaxy cluster the binding energy is likely to be primordial in origin and has not been appreciably enhanced by dissipation. If we can apply a simple spherical collapse model to galaxy clusters, binding energy is conserved throughout the linear regime and we can infer the initial binding energy. One finds that as the cluster begins to bind gravitationally (hydrostatic equilibrium being attained when its kinetic and gravitational energies obey the virial theorem), it collapses by a factor of 2 from the radius of maximum expansion and its density changes by a factor of roughly 180. The inferred binding energy, corresponding to $(v/c)^2$ on the order of 10^{-5}, should be reflected in temperature fluctuations of the cosmic microwave background that trace the initial density fluctuations.

Of course, I have presented only a very tentative sketch of what might have happened during the epoch of galaxy formation. A complementary approach to galaxy formation focuses on numerical simulations of protogalactic collapse, using ensembles of mass points designed to mimic both the gravitational interactions undergone by dark matter and by stars and the inelastic behavior of colliding gas clouds. Some success has been achieved, despite the crudeness of the approximations, in modeling the density profiles of the bright and dark halo components and in deriving galactic rotation curves.

From fluctuations to galaxies

To understand the distribution and properties of galaxies we need to know the properties of the density and temperature fluctuations in the universe before galaxies or clusters began to condense from the primordial matter. One of the most intriguing ideas to emerge from the application of particle physics to cosmology in recent years is the prediction that at a very early epoch (before about 10^{-35} sec after the Big Bang), the universe underwent a period of exponentially rapid growth, associated with the spontaneous breaking of the symmetry among the strong and electroweak interactions. (See PHYSICS TODAY, May 1983, page 17.) The universe we see is thus, in a sense, a crystallized bubble of the vacuum of a much larger space.

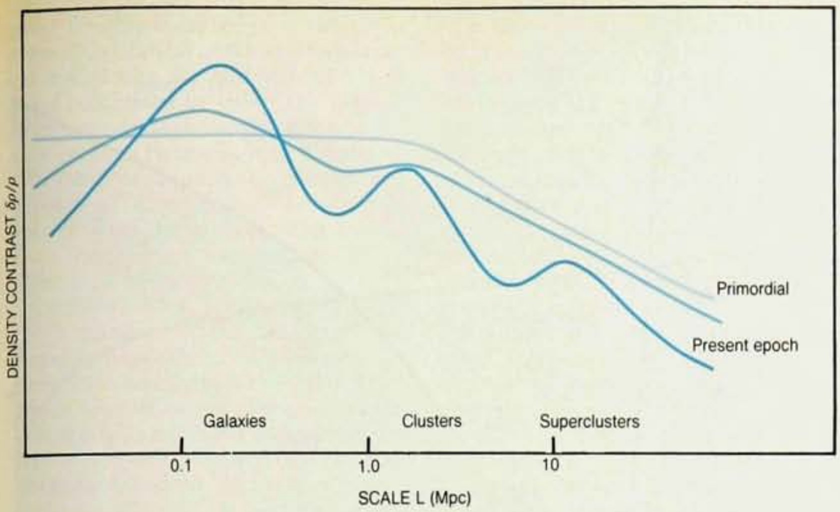
As a result of the inflation, quantum fluctuations are boosted up to macroscopic scales, resulting in primordial density fluctuations at about 10^{-35} sec after the Big Bang, when the temperature is about 10^{15} GeV. In the simplest models, these fluctuations have a spectrum that has no preferred scale, have random phases and are adiabatic, with a universal value of specific entropy, or photon-to-baryon ratio.

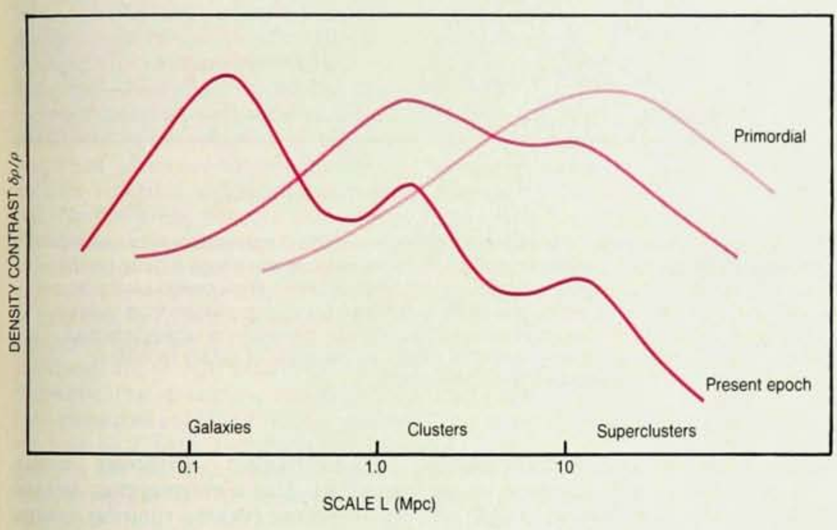
After the universe begins to expand normally, the infinitesimal density fluctuations grow in amplitude by accreting surrounding matter. After the Big Bang the dominant form of mass is at first the blackbody radiation, but after some 10^4 years, when the temperature has dropped to about 10 eV, the radiation has redshifted sufficiently that matter becomes the dominant form of mass, and gravity becomes the dominant force shaping the development of the universe. Small fluctuations grow freely from this stage onward, at least if the matter is predominantly in the form of cold, weakly interacting particles. The horizon at the epoch of equal matter and radiation densities corresponds to a length scale today of 10 Mpc. Any fluctuations on a larger scale that we now see must have entered the horizon more recently; such fluctuations subsequently continue to grow until the present epoch. Thus finally, when the fluctuations have become large, the spectrum of density fluctuations $\delta\rho/\rho$ ends up approximately constant over small scales L (reflecting the initial scale invariance), but proportional to L^{-2} on large scales (reflecting the amount of time these fluctuations have had to grow since entering our horizon), the very gradual transition occurring at 10 Mpc.

There is at least one intriguing test of the prediction that on large scales $\delta\rho/\rho$ is proportional to L^{-2}. The associated peculiar velocities, smoothed over scale L, must be roughly proportional to L^{-1}. I will describe below how these predicted amplitudes compare with the recently observed large-scale streaming motions. Whether the smaller-scale fluctuations actually survive into the matter-dominated era depends on the nature of the matter itself.

Inflation requires that the universe have precisely the critical density required to close it. This is usually expressed by saying that the ratio Ω of the mean density to the closure value must be 1. Any curvature resulting in a deviation of Ω from unity would have been almost entirely erased during the inflationary era. The data we have on the abundances of helium and deuterium indicate that the baryonic contribution to Ω cannot exceed 0.1. (Accounting for the observed abundances of the light elements, after suitable corrections for galactic nucleosynthesis, is one of the major successes of Big Bang cosmology.) It follows that at least 90%—and perhaps as much as 99%, given the uncertainties in the data—of the matter in the universe is nonbaryonic. This dark matter must be much more weakly interacting than baryons and remain relatively diffuse and uniform, and hence relatively unobservable, while the baryons contract to form galaxies and stars.

The models for the dark matter fall into two categories, depending on the velocity dispersion, or temperature, of the dark matter at the epoch of matter–radiation equality. Any heavy particle (one with a mass of an MeV or more) would be cold, having negligible velocity dispersion at temperatures around 10 eV, and hence could clump gravitationally on all scales. Examples of cold dark matter candidates include the least massive particles predicted by





Density contrast $\delta\rho/\rho$ of primordial fluctuations as a function of size L. The upper graph shows the evolution of the fluctuation spectrum for a universe dominated by cold dark matter: Small-scale fluctuations develop first (light blue), while clusters and superclusters develop only recently (dark blue). The lower graph shows the evolution for a hot dark matter universe: The largest-scale structures collapse first (light red), and galaxies form from these by fragmentation (dark red). Galaxies form only recently in such a hot dark matter universe.

supersymmetric theories—such as the photino, higgsino or scalar neutrino. A light particle, on the other hand, such as a neutrino with a mass in the range 10–50 eV, would be quasirelativistic at matter–radiation equality, and hence is a candidate for hot dark matter. Because of the high velocity dispersion of such particles, pre-existing small-scale fluctuations are erased over time, and only the largest fluctuations, extending over 10 Mpc or more, can survive.

If the universe is dominated by cold dark matter, the smallest subgalactic structures, with a mass around $10^5 M_\odot$, are the first to acquire densities great enough to produce nonlinear effects, and collapse as a result. Hierarchical merging results in the formation of galaxies and eventually of galaxy clusters—a "bottom-up" scenario for the formation of structures in the universe. Hot dark matter, on the other hand, results in a "top-down" scenario for the evolution of structure: The first fluctuations to enter the nonlinear regime have the mass of a galaxy cluster—on the order of $10^{15} M_\odot$. These clouds collapse, typically into sheetlike pancakes or caustic surfaces, and fragment into galactic-mass objects.

The two rival scenarios, pancake fragmentation versus hierarchical clustering, have been hotly debated. The simplest versions of both scenarios have run into trouble when confronted with observation.

Effects of dark matter

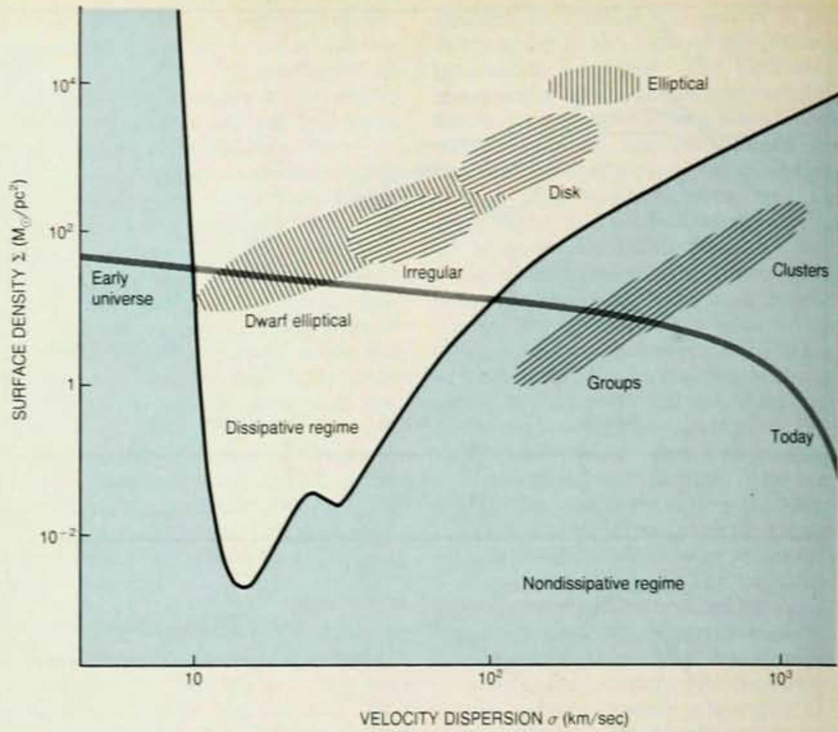
Consider first hot dark matter. Because large-scale structures form first, most galaxy formation in the top-down scenario must have been unacceptably recent to produce the observed large-scale galaxy correlations we see now. One can salvage the model by scenarios that involve triggering the early fragmentation of large structures.

In one scheme, there are rare "seeds" that lead to explosions liberating vast amounts of energy. The ensuing blast waves drive dense shells into the intergalactic medium, fragmenting it into pieces of galactic or subgalactic mass. The rarity of the seeds removes the constraint from the galaxy correlations, and the blast waves generate a matter distribution containing large bubblelike voids. Observational data from recent three-dimensional surveys of galaxy distributions show voids of up to 50 Mpc in diameter. (See PHYSICS TODAY, May, page 17.) Difficulties with the explosion theory arise in that the typical bubble size cannot exceed about 5 Mpc, unless the explosions start very early in the evolution—at a redshift larger than about 10. Appropriate sources of explosive energy release are difficult to come by. One recent exotic suggestion appeals to the discharge of superconducting cosmic strings in a weak primordial magnetic field. These strings are one-dimensional topological defects left over from the symmetry-breaking transition at the end of the grand unification epoch. While strings are generic in many models, some models involve a particular class of strings that are superconducting. The string tension drives motions of the strings at nearly the speed of light, electric currents develop, and in the presence of even a weak primordial magnetic field, superconducting strings carrying large currents are a prolific source of low-frequency photons. Once the currents have built up, these photons are released explosively, injecting energy into the intergalactic medium. A substantial fraction of the rest mass of a string can be released in this way. For the sort of string masses one typically assumes, about 10^{63}–10^{64} ergs would be released in a single string loop responsible for accreting the mass of a galaxy cluster. This considerably exceeds the energy release in even the most violent outbursts of radiogalaxies, and could, according to one recent speculation, be channeled into narrow jets, thereby providing a potential energy source for quasars.

Cosmic strings need not be superconducting to help hot dark matter play an important role. A string loop can act as a seed accreting matter gravitationally and thereby allow galaxy formation to occur before the hot dark matter develops any large-scale structures with clearly defined boundaries. Such string-seeded galaxy formation could probably explain the clustering of galaxy clusters: String loops form as a consequence of self-intersections of infinite cosmic strings, and the loops in turn can self-intersect and subdivide; these self-interactions and the types of motion that strings undergo apparently result in a clustering of galaxies with a fractal dimension that agrees with what is observed.

The string model is testable, because the number of rich clusters uniquely specifies the string mass parameter. Observable implications of the models include a gravitational wave background from early loop decays that is capable of affecting the timing of millisecond pulsars, gravitational lensing of background galaxies and quasars, and generation of a unique signature of fine-scale anisotropy in the cosmic microwave background radiation. None of these effects have yet been seen, but all are capable of providing a test of the string hypothesis.

If the universe is dominated by cold dark matter, the first bound structures on subgalactic scales form at an early epoch, so strings are not required for galaxies to form. Objects of the mass of a dwarf galaxy form at a redshift of about 10. Larger galaxies form more recently, with the outermost halos of the largest galaxies only forming at about the present epoch. Secondary infall of dark matter provides a potential well that can account for the rotation curves of spiral galaxies.



Density–temperature parameter space for galaxies. In this diagram density is represented by mean surface density Σ and temperature by mean velocity dispersion σ. The hatched areas indicate observed ranges for various kinds of structures. The colored area indicates the range of parameters for which gravitational collapse takes place faster than radiative cooling (t_g less than t_c); in this nondissipative regime, star formation is largely inhibited. The gray curve shows the evolution of cold dark matter at the onset of galaxy formation, increasing in temperature and decreasing in density.

If there is indeed enough cold dark matter to make the value of Ω equal to 1, how can it be that we find a value of only 0.1–0.3 based on the measured dynamics of galaxies? Clearly, the cold dark matter cannot have avoided clustering on all scales. A popular resolution of this apparent paradox requires the introduction of a biasing threshold. If luminous galaxies form only in the highest-density peaks of the fluctuation spectrum, one can expect galaxies to cluster because, just as high mountain peaks cluster in mountain ranges, so are the high-density peaks in a distribution of random-phase fluctuations more highly clustered than the rms density fluctuations. In that case, most of the matter in the universe is more uniformly distributed than the luminous galaxies, and a measurement of the density based on luminous objects necessarily underestimates the true value of the density. If galaxies form only in peaks 2–3 times the rms fluctuation level, one finds that only a modest fraction of the mass density of the universe is now in the form of luminous galaxies, and the dynamic measurements of Ω should yield a value of 0.2–0.3. In other words, if galaxy formation is biased toward rare peaks, one obtains an acceptable model of large-scale structure. Indeed, maps of the galaxy distribution based on simulations of the cold dark matter hypothesis are virtually indistinguishable from the real thing, insofar as one finds galaxy groups and clusters, large voids, and filaments and sheets of galaxies.

Several theoretical explanations of why galaxy formation might be biased have been put forth. One involves systematic disruption of the shallower potential wells by supernova explosions driven by dying massive stars. The most common fluctuations are the most weakly bound, so the disruptions result in luminous galaxies' forming only from the rarer fluctuations. Another possibility is that the regions slightly denser than average behave locally like a universe with Ω larger than 1; fluctuations grow much faster in these regions than in the slightly underdense regions, which behave as if Ω were less than 1. (For $\Omega = 1$ there is an exact balance between expansion and gravitational collapse.) Hence, if one catches the universe at exactly the right moment, only a subset of fluctuations will have collapsed, thereby providing a more or less natural biasing mechanism. Problems remain, however. One is overmerging: In locally overdense regions, runaway collapse can produce objects more massive than galaxies. These could be groups of galaxies if merging of the baryonic cores of galaxies can be avoided; more complex numerical simulations incorporating both baryons and dark matter are needed to verify this. A more serious issue is that while theories involving cold dark matter successfully reproduce the correlations of the galaxies, they cannot account for the observed clustering of galaxy clusters. There simply is insufficient power in a biased distribution of cold dark matter, in which the rms mass fluctuations have unit amplitude on a scale of about 3 Mpc, to reproduce the cluster correlations, which have unit amplitude on a scale of about 40 Mpc. However, the Abell clusters, on which these correlations are based, are themselves an observationally biased sample, derived from galaxy counts. It remains to be seen whether clusters selected from three-dimensional surveys reveal the same degree of correlation.

The recent report (see PHYSICS TODAY, November, page 17) of large-scale streaming motions—amounting to some 600 km/sec over distances of order 50 Mpc observed in a sample of some 400 elliptical galaxies—raises another difficulty for the idea of a universe dominated by cold dark matter. Such large-scale streaming motions are impossible to reconcile with a distribution of cold dark matter consistent with the biased distribution of luminous galaxies—the large-scale velocities in these models are expected to be no more than about 100 km/sec.

On the other hand, the observation of the streaming motion is itself not perfectly unambiguous. Accurate distance measurements for remote galaxies are required to establish the streaming motion, and these measurements rely on empirical correlations between distance-dependent quantities (luminosity or effective radius) and a distance-independent quantity (central stellar velocity dispersion for ellipticals or maximum rotation velocity for spirals). The theoretical basis for these correlations is not well understood, but they seem to provide one of the best empirical distance-measuring tools both for elliptical and for spiral galaxies. Nevertheless, the systematic uncertainties in the large-scale streaming velocities are still large. It may be, for example, that a few smaller aggregates of galaxies, each with a coherent but randomly directed streaming motion, could account for the observed signal. Moreover, the streaming has thus far been measured only for elliptical galaxies, and as I have mentioned, spiral galaxies are generally seen in lower-density regions of the universe than ellipticals—that is, more of them are outside of clusters. Independent confirmation of the streaming for a sample of spiral galaxies would thus be very useful.

One should bear in mind that the motivation for nonbaryonic dark matter—cold or hot—rests on the belief that Ω equals 1 as a result of the inflationary expansion just after the Big Bang. If the universe did not undergo such an expansion, the value of Ω reflects the initial conditions of the Big Bang, and there is a serious possibility that Ω is more nearly 0.1, as the dynamical measurements have shown out to about 20 Mpc. In that case, primordial nucleosynthesis predictions are entirely consistent with the dark matter's comprising only low-mass stars. Indeed, some indication that galaxy halos could be predominantly baryonic has come from studying rotation curves of disk galaxies, which reveal little evidence of any transition from a baryon-dominated spheroid and disk to a dark-matter-dominated halo. While this result is consistent with the hypothesis that the halo matter is stellar, other explanations are possible. The halo of our own galaxy appears to be truncated at about 50 kpc, a result that is difficult to understand in the context of a universe dominated by cold dark matter. An intriguing clue that the halo matter might actually consist of low-mass stars has come from studies of globular star clusters, whose luminosity functions appear to rise steeply toward low stellar masses as the cluster metallicity decreases. These globular cluster observations sample only a very limited range of the stellar mass function. However, were this result to be representative of stars in the halo, and with only a modest extrapolation, the outer regions of a galaxy would have sufficient mass per unit of luminosity to account for practically all of the halo mass inferred from studies of disk galaxy rotation curves. (See the box on page 34.)

Shapes of galaxies

Galaxies come in a wide range of sizes and shapes—from small, rather irregular blobs, such as the Magellanic Clouds, to large, beautiful spiral systems, such as the Andromeda nebula. In size they range from perhaps 2×10^3 parsecs in diameter with a mass of 10^7 $M_\odot$ for the dwarf galaxies; through 3×10^4 parsecs and 10^{11} $M_\odot$ for the gas-rich spirals, such as our own; to 10^5 parsecs and 10^{12} $M_\odot$ for the giant ellipticals.

A striking feature of the distribution of galactic binding energies is the marked dependence on galactic morphology. Elliptical galaxies are denser (in terms of the surface density Σ or volume density n) and hotter (in terms of the velocity dispersion σ_v) than spiral and barred spiral galaxies. The spiral galaxies are generally richer in gas than the spherical and elliptical galaxies. There are clearly more fundamental parameters than the observed density and energy that help determine morphology. Two candidates for such a role are specific angular momentum and initial density.

A protogalactic cloud can acquire angular momentum from neighboring clouds that exerted tidal torques at a time when the clouds were still expanding density fluctuations that possessed appreciable quadrupole moments and were large enough to have an appreciable nonlinear gravitational interaction. Including such torques on the baryonic matter in simulations of the dissipative collapse of a protogalaxy within a dark halo results in a rotationally supported disk. The dimensionless angular momentum λ (approximately equal to one-third of the ratio of mean rotational speed to velocity dispersion for an elliptical) is predicted to be 0.07 with a broad dispersion. The observed values for luminous elliptical galaxies fall within this range—one of the rare success stories of galaxy formation theory. Spiral galaxies have a value of λ close to 1. One can account for this by dissipative collapse of baryons through a dark halo, with λ increasing proportionately to the radial collapse factor until the galaxy attains rotational support.

Most luminous ellipticals are not rotationally supported, and their lower values of λ together with their high central surface brightnesses are likely to be a consequence of dissipation following mergers between gas-rich disk galaxies. There is ample evidence that mergers are going on today, and they were presumably far more frequent at earlier epochs. The discovery of multiple shells around some ellipticals is strong evidence for survival of relics of a past merger, and multiple nuclei in a few ellipticals provide marginal evidence for more recent mergers.

However, mergers between conventional disk galaxies probably cannot account for the observed frequencies of spheroidal galaxies, of the most massive elliptical galaxies or of the dwarf ellipticals. Moreover, one of the most striking characteristics of ellipticals is their spatial clustering: Ellipticals are found preferentially in high-density regions, such as rich cluster cores, lenticular galaxies are found mostly in regions of intermediate density; and spirals are generally in low-density regions. Yet mergers are not occurring today in clusters: The relative velocities of galaxies are too large for colliding galaxies to merge together. The fundamental reason for morphological differences between galaxies must take account of environmental effects, and the density of the region in the universe where the galaxy formed is the obvious culprit. To account for morphology by merging, one needs an abundant gas supply: This allows the freedom to dissipate energy and form large, dense spheroids. Gas-rich protogalaxies interacting during the initial phase of cluster formation, before large relative velocities developed, seem ideal candidates for this role. The problem one immediately encounters, however, is that at the epoch of galaxy formation, galaxy clustering could not have been very pronounced in any bottom-up scheme of structure formation. One might expect a forming galactic cloud to turn itself into stars within a collapse time of 10^8 years—long before the 10^9 years that must elapse before galactic clustering occurs. Somehow protogalaxies must be capable of preserving their gas-rich char-

The 'dark matter' problems

Cosmologists tell us that 90% or more of the universe's mass may be invisible. Because of the attractiveness of the inflationary scenario for the Big Bang, theorists have already launched a bandwagon of exotic particles to try to explain both the so-called missing mass problem and galaxy formation. The term "missing mass" is a misnomer: The problem is "missing light" or "hidden mass." Apart from the cosmological "dark matter" problem, discussed in the accompanying article, there are at least two more dark matter problems associated with galaxy systems.

Recently Jack Hills (Los Alamos) has argued that much mass in the plane of the Milky Way Galaxy is in the form of "gray" dwarfs: low-luminosity, low-mass objects. The kinematics of nearby stars implies a total mass in the Galactic plane larger than the observed baryonic mass. Apart from stars and interstellar matter composed of gas and dust, Hills says that dim Jupiter-size objects are the most probable contributors to mass in the Galactic disk.

Another dark matter problem is revealed by flat galactic rotation curves: Visible matter far from the center of a galaxy has orbital speeds much greater than one would expect for Keplerian orbits, given the luminous mass within the orbits. Extensive dark matter galactic halos have been invoked to account for these observations. The dark material may extend out far enough that tidal forces between galaxies in a cluster shred off the halos, yielding a smooth distribution of dark matter within the cluster. This dark matter can be assayed by applying the virial theorem to clusters of galaxies, with the assumption that such systems are dynamically relaxed. Whatever its origin and nature, the virial mass is indeed much in excess of the total visible galactic mass in a cluster of galaxies.

Can finding the mass of our own Galaxy cast any light on the dark matter problem of flat rotation curves? The Sun's motion yields only the Galactic mass located within the Sun's orbit about the Galactic center. One can use orbits of stars as probes to map the Galactic mass distribution—whether the matter is visible or not.

Blane Little (University of Toronto) and Scott Tremaine (Canadian Institute for Theoretical Astrophysics) have carried out a statistical study of about a dozen distant star clusters and satellite galaxies of our own Galaxy. Because both distance and radial (line of sight) velocity are known for these objects, Little and Tremaine can calculate the Galactic mass M as 2.4×10^{11} solar masses (with a statistical range of $1.7\text{–}3.7\times10^{11}\ M_\odot$) with M less than about $5.2\times10^{11}\ M_\odot$ at the 95% confidence level. This result is about two to three times *smaller* than predicted from a very extended halo. Little and Tremaine suggest that the Galaxy's massive dark halo extends to less than 50 kpc from the Galactic center. Their method assumes that on average motions are noncircular with an isotropic distribution of velocity components—consistent with sampling a dynamically old system.

Bruce Carney (University of North Carolina) and David Latham (Harvard–Smithsonian Center for Astrophysics) look at the stellar equivalent to Solar System comets: population II stars on markedly eccentric orbits crossing the Sun's Galactic orbit. Such objects are only marginally bound to the Galaxy. Carney and Latham have selected extremely high-velocity star candidates from surveys of objects showing large proper motion, that is, angular displacement perpendicular to the line of sight. Their analysis uses a star's three-dimensional velocity vector whereas Little and Tremaine rely on a statistical study of the line-of-sight velocity components. Carney and Latham find an escape velocity for the Galaxy of 525 km/sec, which implies a Galactic mass of at least 5×10^{11} $M_\odot$—close to Little and Tremaine's upper limit.

Tony Tyson (AT&T Bell Labs) suggests that "the average gravitational lens distortion of background galaxy images by foreground galaxies is an independent, nonkinematical measurement of average galaxy mass distribution. The data can be made consistent with the kinematic and photometric observations leading to dark halos if we assume that most dark halos are truncated at 30–60 kpc. This constrains the nature of the dark matter associated with galaxies."

—Per H. Andersen

acteristics until galaxy clustering develops.

In any model of galaxy formation, gas-rich protogalaxies almost certainly can survive until galaxy clustering develops—provided that the probability of forming massive stars was enhanced in the early universe. The trick is to suppress low-mass star formation (it is the long-lived stars of mass less than $3\,M_\odot$ and their remnants that are responsible for locking up the gas in a galaxy for long periods) while enhancing the rate of star formation per unit mass above what we now see in normal spirals. This phenomenon, known as bimodal star formation, is believed to occur in regions of vigorous star formation, such as the inner spiral arms of our galaxy and in starburst galaxies. It may be the reason why such galaxies as the Milky Way have retained significant amounts of gas since their formation. This phenomenon is also suspected to have occurred when the oldest nearby stars formed: The chemical composition of the solar neighborhood directly reflects the effects of massive, short-lived stars, which are heavy-element factories; low-mass stars like the Sun, which live for a Hubble time, have provided very little material to the interstellar gas.

Once galaxies begin to form, tidal interactions and galaxy mergers severely alter star formation in the denser regions. In a low-density region the gas supply may last for at least 10^{10} years, and disks slowly develop by dissipative infall. In the denser regions, star formation is likely to be stimulated and then truncated as the gas reservoir in the outlying regions of a protogalaxy is disrupted. Spirals, which have a relatively low rate of star formation and are gas rich, form in the low density environments, whereas ellipticals, which have had a high past rate of star formation and are gas poor today, form in denser environments. In these dense environments, tidally enhanced dissipation and ensuing star formation are likely to produce a final galaxy of higher binding energy and mean density—as is in fact inferred for luminous ellipticals from the observations. Precisely why tidal interactions and merging should enhance star formation efficiency is a matter of conjecture. The gas response, at least in a cold system such as a disk, is very nonlinear, and an enhanced rate of cloud collision not implausibly triggers a burst of star formation. Perhaps massive stars dominate because of the high thermal energy in a vigorously stirred system.

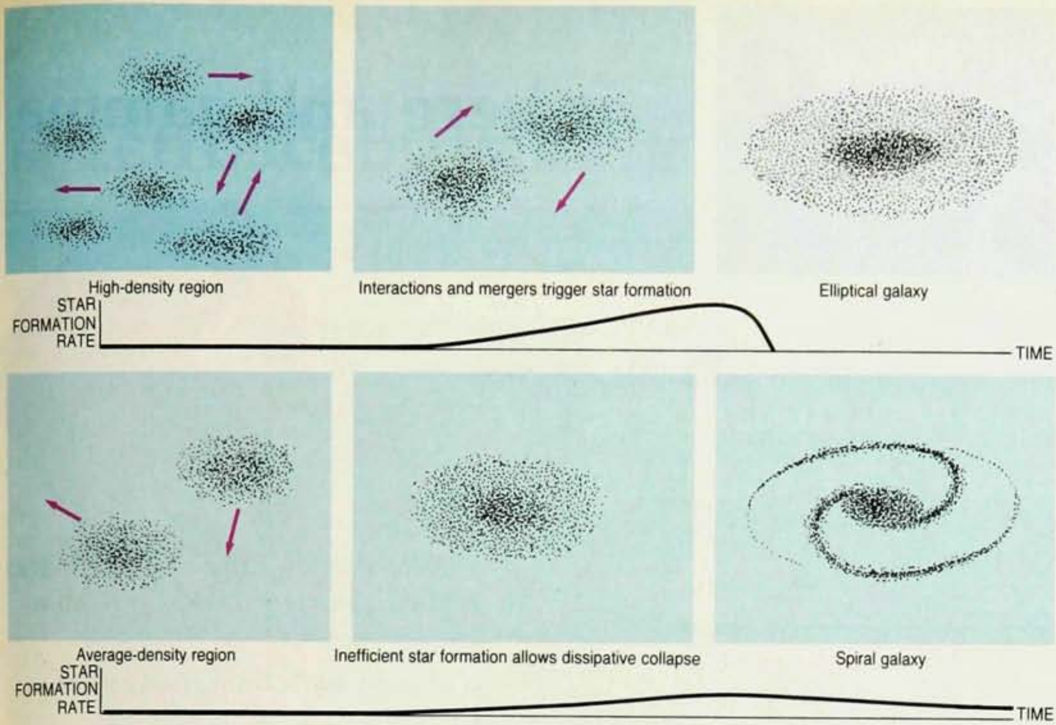
Dissipation masks much of the influence of the initial conditions on the luminous regions of galaxies. One has to look at dark halo structures to assess the influence of the primordial fluctuation spectrum. A recent intriguing result is that the cold dark matter spectrum, or at least a fluctuation spectrum with a similar slope over the relevant range, has been found from N-body simulations to produce dark halos with flat rotation curves of more or less the required amplitude and spatial frequency. This represents one of the most noteworthy successes of the cold dark matter theory, and one that rival theories are hard pressed to surpass.

Future goals

This brief sketch of galaxy formation is necessarily very speculative. The inputs from both particle physics and astrophysics are highly uncertain. There is no compelling experimental evidence for inflation, nor has a single protogalaxy been definitively discovered.

Despite these uncertainties, some pieces of the cosmic jigsaw puzzle are falling into place:

▶ It is now apparent that for the



Galaxy morphology. Elliptical galaxies (top) most likely form as a consequence of mergers of smaller subsystems, and thus are likely to form in dense environments where star formation is efficient; the violent dynamical interactions of the forming stellar systems result in smooth, gas-poor elliptical galaxies. Disk galaxies (bottom) form in low-density, relatively quiescent, gas-rich regions, where star formation remains relatively slow and inefficient; the continuing infall of gas and gaseous dissipation result in formation of a galactic disk. The graphs show the rates at which stars form in each case.

galaxies we can see, tidal interactions and mergers play an important role in history and morphology as well as in star formation. If this is true in our vicinity, one cannot doubt that it was dominant at earlier epochs in the universe.

▶ The rotation curves of galaxies are testimony to the prevalence of diffuse dark matter. The masses and detailed morphologies of galaxies can be elegantly explained by the physics of gaseous dissipation, and the clustering of galaxies can be quantitatively understood if the universe is dominated by cold dark matter.

How to improve the situation? Verification that Ω equals 1, for example, would confirm a key inflationary prediction. An observation of fluctuations in the microwave background would provide a probe of the initial conditions from which large-scale structure developed. Current upper limits on possible temperature fluctuations are lower than 0.01%. The cold dark matter theory predicts that fluctuations at the level of a few parts in 100 000 should be present. Anisotropies with spatial correlations on the scale of 1° or 2° sample the primordial spectrum of fluctuations that emerged from the inflationary epoch, because causal effects could not have been operative on such a scale in the standard Big Bang. Ongoing experiments should be capable of detecting such anisotropies.

Evidence that luminous galaxy formation occurred in the recent past would corroborate the picture of a universe dominated by cold dark matter. The theory predicts that a protogalaxy should be an extended collection of merging clouds, peaking in luminosity at a redshift of 1 or 2, but in which stars have been forming ever since a redshift of 10 or so. At least one protogalaxy may have already been discovered, namely a very extended gas cloud at a redshift of 1.8, equal in mass to a large galaxy and profusely emitting Lyman-alpha photons. While the search for protogalaxies is being vigorously pursued, there are already tantalizing indications of evolution in remote regions of the universe. One noteworthy example has come from studies of Lyman-alpha absorption clouds, seen in the spectra of high-redshift quasars. (Quasars, because of their luminosity, are used as beacons with which to study intervening gas-rich systems.) Apparently there is an abundant population of small clouds, progressively more frequent in the distant reaches of the universe, that produce a forest of weak Lyman absorption lines. Even the strong Lyman-alpha absorptions seen in quasar spectra are produced in galaxies that are more gas rich and more extended than their contemporary counterparts.

The deep redshift surveys now being undertaken may verify the reality of the large voids, of the huge superclusters and filaments of galaxies, and of the clustering and the peculiar motions or cosmic drift of the galaxy clusters. These results all may challenge the simple schemes for the generation of large-scale structure I have described here. It remains to be seen whether more exotic models—such as the non-Gaussian fluctuations driven by cosmic strings—are needed to account for the new observations. The standard model of a universe dominated by cold dark matter, however, clearly has many successes to its credit. It is always possible that a new theory will arise, phoenixlike, from the ashes of cold dark matter, but my best guess is that the current theory, warts and all, will be incorporated without major modifications into any future refinement.

Bibliography

- J. Bardeen, J. R. Bond, N. Kaiser, A. S. Szalay, Astrophys J. **304**, 15 (1986).
- G. Blumenthal, S. Faber, J. Primack, M. Rees, Nature **301**, 584 (1984).
- J. P. Ostriker, C. Thompson, E. Witten, Phys. Lett. **181**, 243 (1986).
- J. Silk, Astrophys. J. **297**, 1 (1985).
- S. D. M. White, C. Frenk, M. Davis, G. Efstathiou, Astrophys. J. **313**, 505 (1987).