exclusive online issue | no.24

SCIENTIFIC AMERICAN.com

EXTREME UNIVERSE

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

EXTREME UNIVERSE

ScientificAmerican.com
exclusive online issue | no. 24

Looking up at the heavens on a crisp autumn evening, it all seems so peaceful. But the serene beauty of the night sky belies the tumultuous nature of the cosmos. Light-years away, stars are being born, black holes are forming, and even the gas between the stars is a hotbed of activity.

In this exclusive online issue, leading authorities recount some of the most thrilling and bizarre discoveries about our universe that have been made in recent years. Explore the link between gamma-ray bursts and black holes. Learn how magnetized stars known as magnetars are altering the quantum vacuum. Tour the interstellar medium, with its landscape of gas fountains and bubbles blown by exploding stars. And find out why scientists are saying the cosmos is experiencing a kind of midlife crisis.

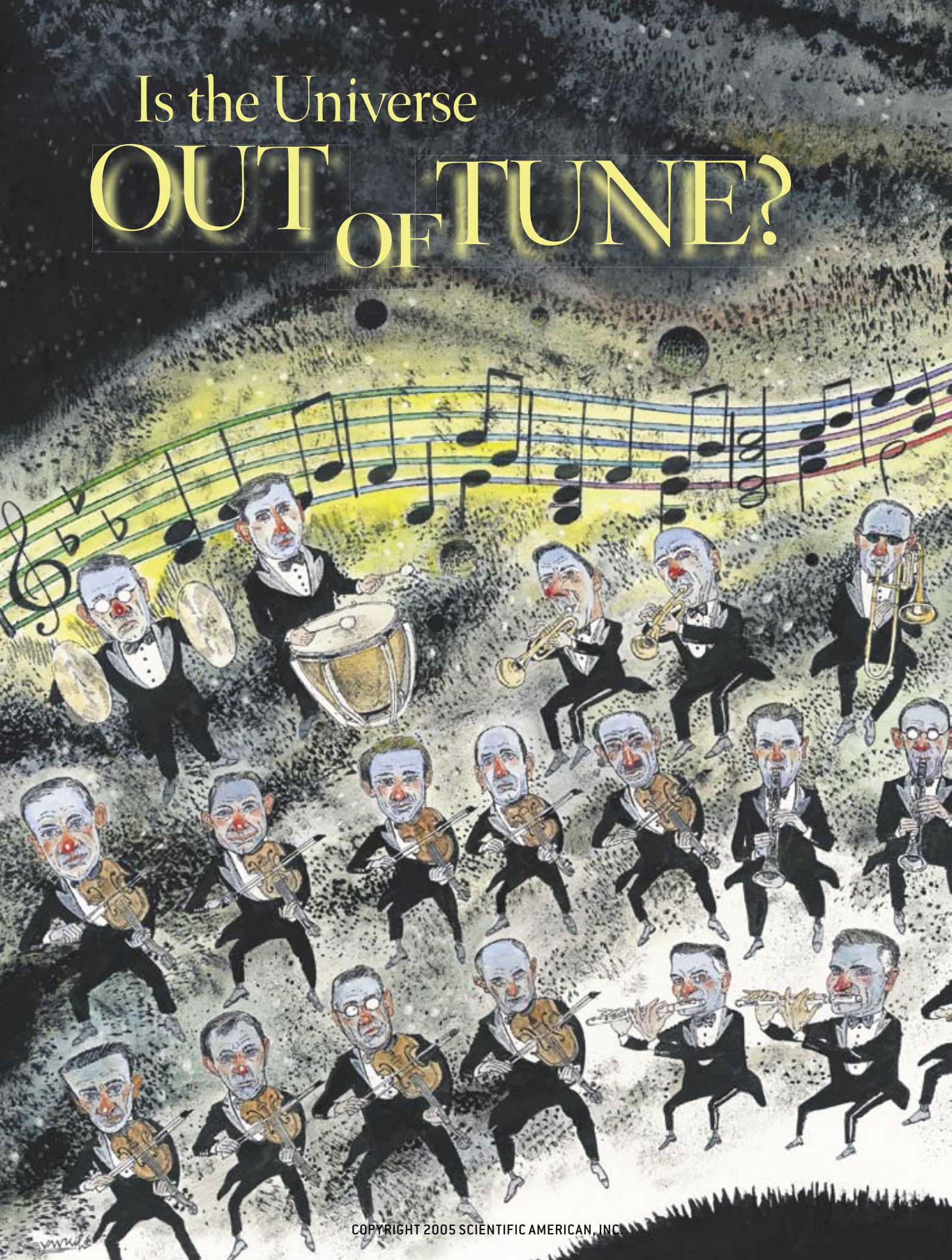
Other articles delve into even weirder phenomena. Jacob Beckenstein explains how the universe could be like a giant hologram. Glen Starkman and Dominik Schwarz listen to the "music" of the cosmic microwave background—and find it strangely out of tune. And Max Tegmark explains how cosmological observations imply that parallel universes really do exist. —*The Editors*

TABLE OF CONTENTS

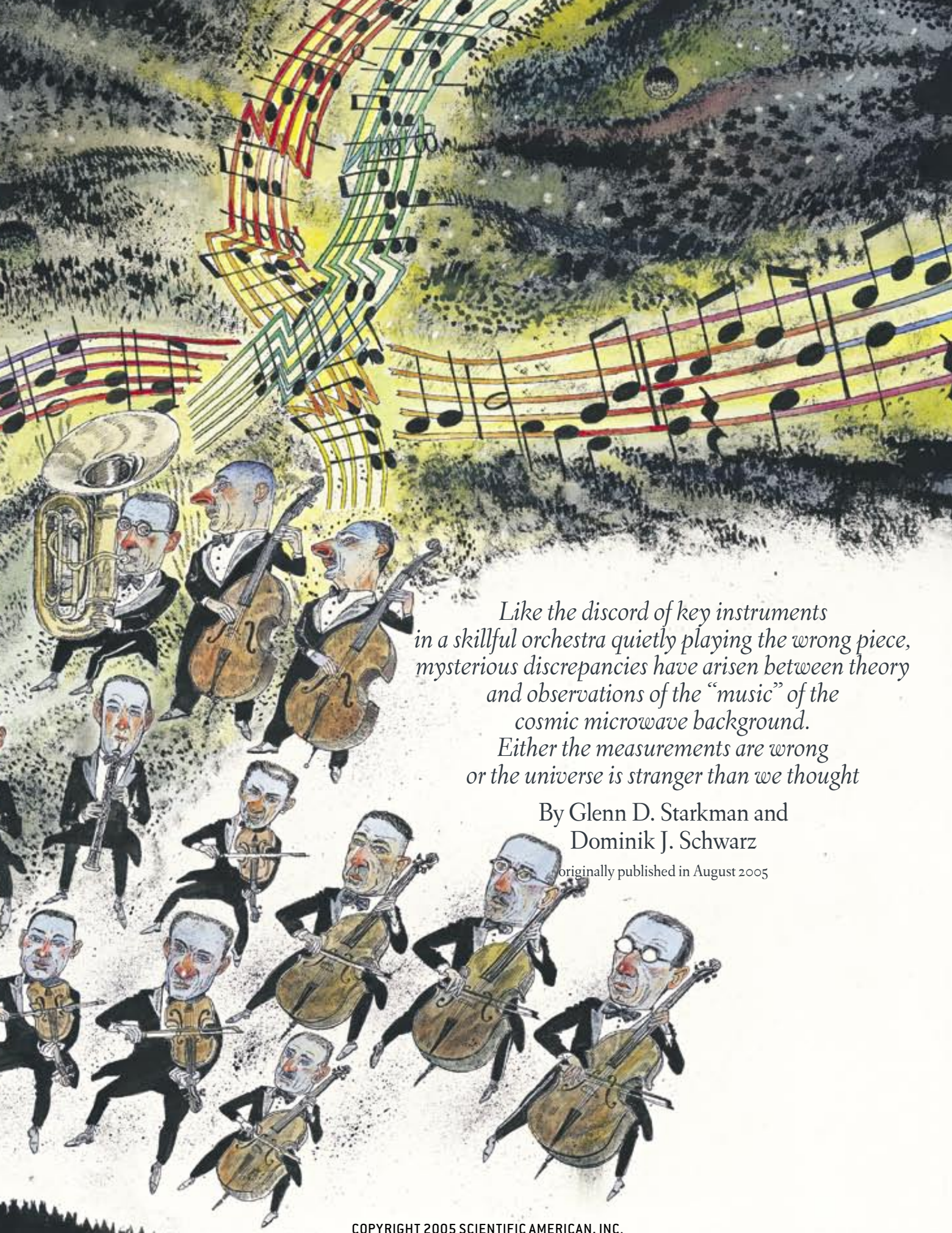
COVER IMAGE: NASA

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

Is the Universe OUT OF TUNE?



COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



Like the discord of key instruments in a skillful orchestra quietly playing the wrong piece, mysterious discrepancies have arisen between theory and observations of the "music" of the cosmic microwave background. Either the measurements are wrong or the universe is stranger than we thought

By Glenn D. Starkman and Dominik J. Schwarz

originally published in August 2005

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

IMAGINE a fantastically large orchestra

playing expansively for 14 billion years. At first, the strains sound harmonious. But listen more carefully: something is off key. Puzzlingly, the tuba and bass are softly playing a different song.

So it is when scientists "listen" to the music of the cosmos played in the cosmic microwave background (CMB) radiation, our largest-scale window into the conditions of the early universe. Shortly after the big bang, random fluctuations—probably thanks to the actions of quantum mechanics—apparently arose in the energy density of the universe. They ballooned in size and ultimately became the galaxy clusters of today. The fluctuations were a lot like sound waves (ordinary sound waves are oscillations in the density of air), and the "sound" ringing throughout the cosmos 14 billion years ago was imprinted on the CMB. Now we see a map of that sound drawn on the sky in the form of CMB temperature variations.

As with a sound wave, the CMB fluctuations can be analyzed by splitting them into their component harmonics—like a collection of pure tones of different frequencies or, more picturesquely, different instruments in an orchestra. Certain of those harmonics are playing more quietly than they should be. In addition, the harmonics are aligned in strange ways—they are playing the wrong tune. These bum notes mean that the otherwise very successful standard model of cosmology is flawed—or that something is amiss with the data.

Scientists have constructed and corroborated the standard model of cosmology over the past few decades. It accounts for an impressive array of the universe's characteristics. The model explains the abundances of the lightest elements (various isotopes of hydrogen, helium and lithium) and gives an age for the universe (14 billion years) that is consistent with the estimated ages of the oldest known stars. It predicts the existence and the near homogeneity of the CMB and explains how many other properties of the universe came to be just the way they are.

Called the inflationary lambda cold dark matter model, its name derives from its three most significant components: the process of inflation, a quantity called the cosmological constant symbolized by the Greek letter lambda, and invisible particles known as cold dark matter.

According to this model, inflation was a period of tremendously accelerated growth that started in the first fraction of a second after the universe began and ended with a burst of radiation. Inflation explains why the universe is so big, so full of stuff and so close to being homogeneous. It also explains why the universe is not precisely homogeneous: because random quantum fluctuations in the energy density were inflated up to the size of galaxy clusters and larger.

The model predicts that after inflation terminated, gravity caused the regions of extra density to collapse in on themselves, ultimately forming the galaxies and clusters we see today. That process had to have been helped along by cold dark matter, which is made up of huge clouds of particles that are detectable only through their gravitational effects. The cosmological constant (lambda) is a strange form of antigravity responsible for the present speedup of the cosmic expansion [see "A Cosmic Conundrum," by Lawrence M. Krauss and Michael S. Turner; Scientific American, September 2004].

Overview/*Heavenly Discord*

- A theory known as the inflationary lambda cold dark matter model explains many properties of the universe very well. When certain data are analyzed, however, a few key discrepancies arise.
- The puzzling data come from studies of the cosmic microwave background (CMB) radiation. Astronomers divide the CMB's fluctuations into "modes," similar to splitting an orchestra into individual instruments. By that analogy, the bass and tuba are out of step, playing the wrong tune at an unusually low volume.
- The data may be contaminated, such as by gas in the outer reaches of the solar system, but even so, the otherwise highly successful model of inflation is seriously challenged.

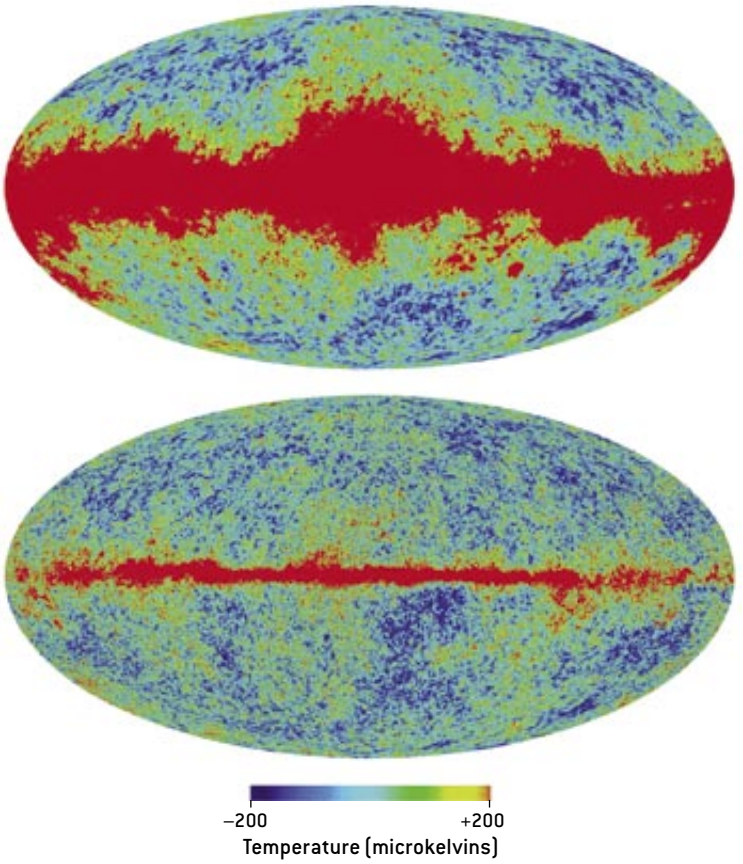
The Most Ancient Light

Despite the model's great success at explaining all those features of the universe, problems show up when astronomers measure the CMB's temperature fluctuations. The CMB is cosmologists' most important probe of the largest-scale properties of the universe. It is the most ancient of all light, originating only a few hundred thousand years after the big bang, when the rapidly expanding and cooling universe made the transition from dense opaque plasma to transparent gas. In transit for 14 billion years, the CMB thus reveals a picture of the early universe. Coming from the farthest reaches, that picture is also a snapshot of the universe at its largest size scale.

Arno Penzias and Robert Wilson of Bell Laboratories first detected the CMB and measured its temperature in 1965. More recently, the cutting edge of research has been studies of fluctuations in the temperature as seen when viewing different areas of the sky. (Technically, these fluctuations are called temperature anisotropies.) The differences in temperature across the sky reflect the universe's early density fluctuations. In 1992 the COBE (Cosmic Background Explorer) sat-

JACK UNRUH [*preceding pages*]

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



MICROWAVE SKY is measured in the K-band (23 gigahertz, *top*), the W-band (94 gigahertz, *bottom*) and three other bands (*not shown*) by the WMAP satellite. The entire sphere of the sky is projected onto the oval shape, like a map of the earth. The horizontal red band is radiation from the Milky Way. Such "foreground" radiation changes with wave band, allowing it to be identified and subtracted from the data, whereas the cosmic microwave background does not.

ellite first observed those fluctuations; later, the WMAP (Wilkinson Microwave Anisotropy Probe) satellite has made high-resolution maps of them.

Models such as the lambda cold dark matter model cannot calculate the exact pattern of the fluctuations. Yet they can predict their statistical properties, similar to predicting their average size and the range of sizes they span. Some of these statistical features are predicted not only by the lambda cold dark matter model but also by numerous other simple inflationary models that physicists have considered at one time or another as possible alternatives. Because such properties arise in many different inflationary models, they are considered "generic" predictions of inflation; if inflation is true at all, these predictions hold irrespective of the finer details of the model. To falsify one of them would be to challenge the scenario of inflation in the most serious way a scientific theory can be challenged. That is what the anomalous CMB measurements may do.

The predictions are best expressed by first breaking down the temperature fluctuations into a spectrum of modes called spherical harmonics, much as sound can be separated into a spectrum of notes [*see box on page 7*]. As mentioned earlier, we can consider the density fluctuations, before they grow into galaxies, to be sound waves in the universe. If this breakdown into modes seems mysterious, recall the orchestra analogy: each mode is a particular instrument, and the whole map of temperatures across the sphere of the sky is the complete sound produced by the orchestra.

The first of inflation's generic predictions about the fluctuations is "statistical isotropy." That is, the CMB fluctuations neither align with any preexisting preferred directions (for example, the earth's axis) nor themselves collectively define a preferred direction.

Inflation further predicts that the amplitude of each of the modes (the volume at which each instrument is playing, if we think about an orchestra) is random, from among a range of possibilities. In particular, the distribution of probabilities follows the shape of a bell curve, known as a Gaussian. The most likely amplitude, the peak of the curve, is at zero, but in general nonzero values occur, with decreasing probability the more the amplitude deviates from zero. Each mode has its own Gaussian curve, and the width of its Gaussian distribution (the wider the base of the "bell") determines how much power (how much sound) is in that mode.

Inflation tells us that the amplitudes of all the modes should have Gaussian distributions of very nearly the same width. This property comes about because inflation, by stretching the universe exponentially, erases, like a pervasive cosmic iron, all traces of any characteristic scales. The resulting power spectrum is called flat because of its lack of distinguishing features. Significant deviations from flatness should occur only in those modes produced at either the end or the beginning of inflation.

Missing Notes

SPHERICAL HARMONICS represent progressively more complicated ways that a sphere can vibrate in and out. As we look closer at the harmonics, we begin to see where the observations run into troubling conflicts with the model. These modes are convenient to use, because all our information about the distant universe is projected onto a single sphere—the sky. The lowest note (labeled l=0) is the monopole—the entire sphere pulses as one. The monopole of the CMB is its average temperature—just 2.725 degrees above absolute zero [*see box on page 7*].

The next lowest note (labeled l=1) is the dipole, in which the temperature goes up in one hemisphere and down in the other. The dipole is dominated by the Doppler shift of the solar system's motion relative to the CMB; the sky appears slightly hotter in the direction the sun is traveling.

In general, the oscillation for each value of l (0, 1, 2 ...) is called a multipole. Any map drawn on a sphere, whether it be the CMB's temperature or the topography of the earth, can be broken down into multipoles. The lowest multipoles are the largest-area, continent- and ocean-size undulations on our temperature map. Higher multipoles are like successively smaller-area plateaus, mountains and hills (and trenches and valleys) inserted in orderly patterns on top of the larger fea-

NASA/WMAP SCIENCE TEAM

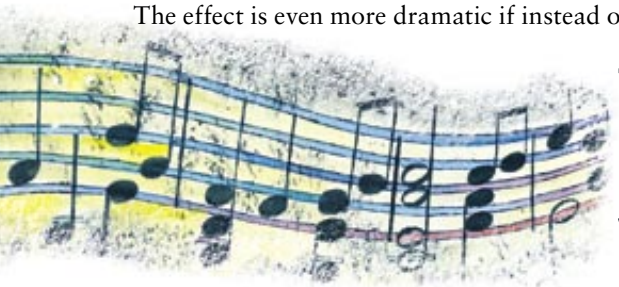
COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

tures. The entire complicated topography is the sum of the individual multipoles.

For the CMB, each multipole l has a total intensity, C_l—roughly speaking, the average heights and depths of the mountains and valleys corresponding to that multipole, or the average volume of that instrument in the orchestra. The collection of intensities for all different values of l is called the angular power spectrum, which cosmologists plot as a graph.

The graph begins at C_2 because the real information about cosmic fluctuations begins with l=2. The illustration on page 54 shows both the measured angular power spectrum from WMAP and the prediction from the inflationary lambda cold dark matter model that most closely matches all the measurements. The measured intensities of the two lowest-l multipoles, C_2 and C_3, the so-called quadrupole and octopole, are considerably lower than the predictions. The COBE team first noticed this deficiency in the low-l power, and WMAP recently confirmed the finding. In terms of topography, the largest continents and oceans are mysteriously low and shallow. In terms of music, we are missing bass and tuba.

The effect is even more dramatic if instead of looking at the total intensities (the C_l's) one looks at the so-called angular correlation function, $C(\theta)$. To understand this function, imagine we look at two points in the sky separated by an angle θ and examine whether they are both hotter (or both colder) than average, or one is hotter and one colder. $C(\theta)$ measures the extent to which the two points are correlated in their temperature fluctuations, averaged over all the points in the sky. Experimentally we find that the $C(\theta)$ for our universe is nearly zero at angles greater than about 60 degrees, which means that the fluctuations in directions separated by more than about 60 degrees are completely uncorrelated. This result is another sign that the low notes of the universe that inflation promised are missing.



The absence of large-angle power is in striking disagreement with most inflationary theories.

This lack of large-angle correlations was first revealed by COBE, and WMAP has now confirmed it. The smallness of $C(\theta)$ at large angles means not only that C_2 and C_3 are small but that the ratio of the values of the first few total intensities—up to at least C_4—are also unusual. The absence of large-angle power is in striking disagreement with *all* generic inflationary models.

This mystery has three potential solutions. First, the unusual results may be just a meaningless statistical fluke. In particular, uncertainties in the data may be larger than have been estimated, which would make the observed results less improbable. Second, the correlations may be an observational artifact—an unexpected physical effect that has not been compensated for in the WMAP team's analysis of its data. Finally, they may indicate a deeper problem with the theory.

Several authors have championed the first option. George Efstathiou of the University of Cambridge was first, in 2003, to raise questions about the statistical methods used to extract the quadrupole strength and its uncertainty, and he claimed that the data implied a much larger uncertainty. Since then, many others have looked at the methods by which the WMAP team extracted the low-l C_l and concluded that uncertainties caused by the emissions of our own Milky Way galaxy are larger than what researchers originally inferred.

Mysterious Alignments

TO ASSESS THESE DOUBTS about the significance of the discrepancy, several groups have looked beyond the information contained in the C_l's, which represent the total intensity of a mode. In addition to C_l, each multipole holds directional information. The dipole, for instance, has the direction of the hottest half of the sky. Higher multipoles have even more directional information. If the intensity discrepancy is indeed just a fluke, then the directional information from the same data would be expected to show the correct generic behavior. That does not happen, however.

The first odd result came in 2003, when Angelica de Oliveira-Costa, Max Tegmark, both then at the University of Pennsylvania, Matias Zaldarriaga of Harvard University and Andrew Hamilton of the University of Colorado at Boulder noticed that the preferred axes of the quadrupole modes, on the one hand, and of the octopole modes, on the other, were remarkably closely aligned. These modes are the same ones that seemed to be deficient in power. The generic inflationary model predicts that each of these modes should be completely independent—one would not expect any alignments.

Also in 2003 Hans Kristian Eriksen of the University of Oslo and his co-workers presented more results that hinted at alignments. They divided the sky into all possible pairs of hemispheres and looked at the relative intensity of the fluctuations on the opposite halves of the sky. What they found contradicted the standard inflationary cosmology—the hemispheres often had very different amounts of power. But what was most surprising was that the pair of hemispheres that were the most different were the ones lying above and below the ecliptic, the plane of the earth's orbit around the sun. This result was the first sign that the CMB fluctuations, which were supposed to be cosmological in origin, with some contamination by emission in our own galaxy, have a solar system signal in them—that is, a type of observational artifact.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

Detecting Harmonics in the Heavenly Music

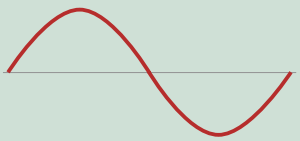
ALISON KENDALL

When scientists say that certain instruments in the cosmic microwave background (CMB) seem to be quietly playing off key, what do they mean—and how do they know that?

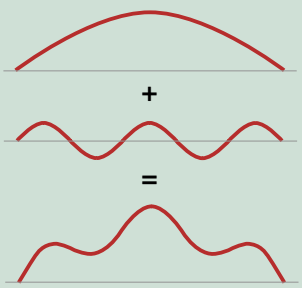
CMB researchers study fluctuations in temperature measured in all directions in the sky. They analyze the fluctuations in terms of mathematical functions called spherical harmonics. Imagine a violin string. It can sound an infinite number of possible notes, even without a finger pressing it to shorten it. These notes can be labeled *n,* the number of spots (called nodes) on the string other than its ends that do not move when the note is sounded.

The lowest note, that is, no node (n=0), is called the fundamental tone. The entire string, except for the ends, moves back and forth in unison (*below*).

The note with a single node in the middle (n=1) is the first harmonic oscillation. In this case, half of the string moves one way while the other half moves the other (*below*). If you sing *do-re-mi-fa-so-la-ti-do,* the final *do* is the first harmonic to the fundamental tone of the first *do*. The note with two equally spaced nodes is the second harmonic, and so on.



Any complicated way that the string vibrates can be broken down into its component harmonics. For example, we can consider the vibration below as the sum of the fundamental tone (n=0) and the fourth harmonic (n=4). Note that the fourth harmonic has a lower amplitude (its waves are shallower) in the sum than the fundamental tone. In the orchestra analogy, instrument number four is playing more softly than instrument number zero. In general, the more irregular the vibration of the string, the more harmonics are needed in the sum.



Now let us examine spherical harmonics—denoted Y_{lm}—in which the modes occur around a spherical "drum." Because the surface of the sphere is two-dimensional, we now need two numbers, l and m, to describe the modes. For each value of l (which can be 0, 1, 2, ...), m can be any whole number between $-l$ and l. The combination of all the different notes with the same value of l and different values of m, each with its respective amplitude (or in audio terms, the volume), is called a multipole.

We cannot easily draw the spherical harmonics as we drew the violin string. Instead we present a map of the sphere colored according to whether a given region is at a higher or lower temperature than the average. (The map's shape comes from being stretched flat, just like maps of the earth hung in schoolrooms.) The monopole, or l=0, is the entire spherical drum pulsing as one (*below*).

The dipole (l=1) has half the drum pulsing outward (*red*) and half pulsing in (*blue*). There are three dipole modes (m= –1, 0, 1) in the three perpendicular directions of space (in and out of the page, up and down, and left and right).

The regions of green color are at the average temperature; these nodal lines are the analogues of nodes on the violin string. As l increases, so does the number of nodal lines.

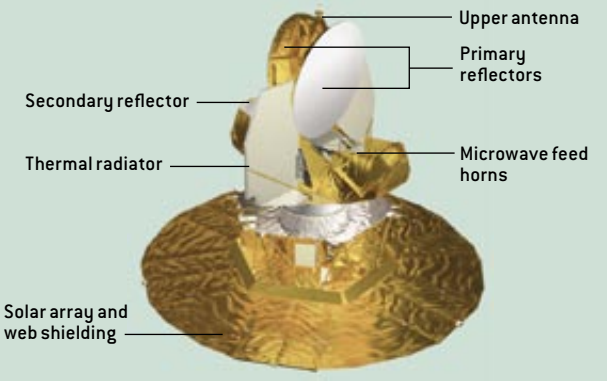
The quadrupole (l=2) has five modes, each with a more complicated pattern of oscillations or temperature variations on the sphere (*below*).

We can break down any pattern of temperature distributions on a spherical surface into a sum of these spherical harmonics, just as any vibration of the violin string can be broken down into a sum of harmonic oscillations. In the sum, each spherical harmonic has a particular amplitude, in essence representing the amount of that harmonic that is present or how loudly that cosmic "instrument of the orchestra" is playing. —*G.D.S. and D.J.S.*

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

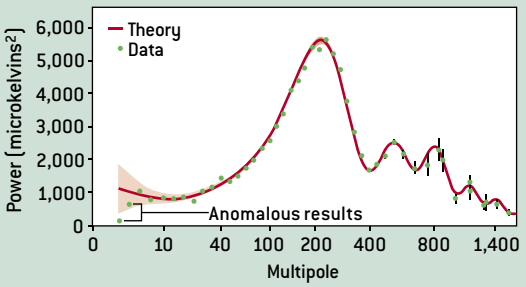
MYSTERIES FROM WMAP

WMAP SATELLITE produces data that are mysterious in three ways.



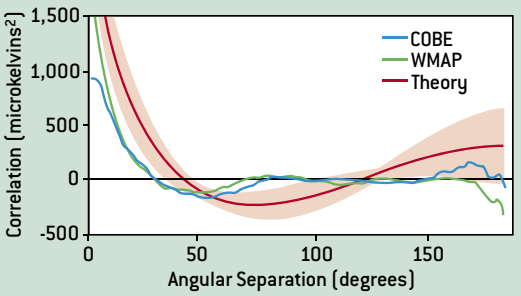
1 ANGULAR POWER SPECTRUM

Most of the WMAP measurements, like those from earlier experiments, are in excellent agreement with values predicted from the inflationary lambda cold dark matter model. But the first two data points (multipoles)—the quadrupole and octopole—are anomalously low in power.



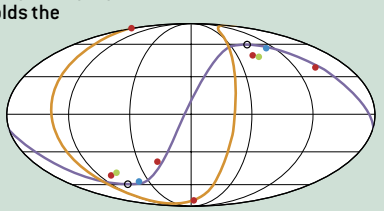
2 ANGULAR CORRELATION FUNCTION

This function relates data from points in the sky separated by a given angle. The data curves from COBE and WMAP should follow the theoretical curve. Instead they are virtually zero beyond about 60 degrees.



3 ALIGNMENT OF THE FIRST TWO MULTIPOLES

The quadrupole (*blue*) and octopole (*red*) should be randomly scattered, but instead they clump close to the equinoxes (*open circles*) and the direction of the solar system's motion (dipole, *green*). They also lie mostly on the ecliptic plane (*purple*). Two are on the supergalactic plane that holds the Milky Way and most of its neighboring galaxies and galactic clusters (*orange*). The probability of these alignments occurring by chance is less than one in 10,000.



Meanwhile one of us (Starkman), together with Craig Copi and Dragan Huterer, then both at Case Western Reserve University, had developed a new way to represent the CMB fluctuations in terms of vectors (a mathematical term for arrows). This alternative allowed us (Schwarz, Starkman, Copi and Huterer) to test the expectation that the fluctuations in the CMB will not single out special directions in the universe. In addition to confirming the results of de Oliveira-Costa and company, we revealed some unexpected correlations in 2004. Several of the vectors lie surprisingly close to the ecliptic plane. Within that plane, they sit unexpectedly close to the equinoxes—the two points on the sky where the projection of the earth's equator onto the sky crosses the ecliptic. These same vectors also happen to be suspiciously close to the direction of the sun's motion through the universe. Another vector lies very near the plane defined by the local supercluster of galaxies, termed the supergalactic plane.

Each of these correlations has less than a one in 300 chance of happening by accident, even using conservative statistical estimates. Although they are not completely independent of one another, their combined chance probability is certainly less than one in 10,000, and that reckoning does not include all the odd properties of the low multipoles.

Some researchers have expressed concern that all these results have been derived from maps of the full CMB sky. Using the full-sky map might seem like an advantage, but in a band around the sky centered on our own galaxy the reported CMB temperatures may be unreliable. To infer the CMB temperature in this galactic band, one must first strip away the contributions of the galaxy. Perhaps the techniques that the WMAP team or other groups have used to remove the galactic thumbprints are not reliable enough. Indeed, the WMAP team cautions other researchers against using its full-sky map; for its own analysis, it uses only those parts of the sky outside the galaxy. When Uros Seljak of Princeton University and Anze Slosar of the University of Ljubljana excluded the galactic band, they found that the statistical significance of some of these alignments declined at some wavelengths. Yet they also found that the correlations increased at other wavelengths. Our own follow-up work suggests that the effects of the galaxy cannot explain the observed correlations. Indeed, it would be very surprising if a misunderstanding of the galaxy caused the CMB to be aligned with the solar system.

The case for these connections between the microwave

THE AUTHORS

GLENN D. STARKMAN and *DOMINIK J. SCHWARZ* first worked together in 2003, when they were at CERN near Geneva. Starkman is Armington Professor at the Center for Education and Research in Cosmology and Astrophysics in the departments of physics and astronomy at Case Western Reserve University. Schwarz has done research on cosmology since he graduated from the Vienna University of Technology in Austria. He recently accepted a faculty position at the University of Bielefeld in Germany. His main scientific interests are the substance of the universe and its early moments.

NASA/WMAP SCIENCE TEAM (*image*); ALISON KENDALL (*graphs*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

background and the solar system being real is strengthened when we look more closely at the angular power spectrum. Aside from the lack of power at low *l*, there are three other points—*l*=22, *l*=40 and *l*=210—at which the observed power spectrum differs significantly from the spectrum predicted by the best-fit lambda cold dark matter model. Whereas this set of differences has been widely noticed, what has escaped most cosmologists' attention is that these three deviations are correlated with the ecliptic, too.

Two explanations stand out as the most likely for the correlation between the low-*l* CMB signal and features of the solar system. The first is an error in the construction or understanding of the WMAP instruments or in the analysis of the WMAP data (so-called systematics). Yet the WMAP team has been exceedingly careful and has done numerous cross-checks of its instruments and its analysis procedure. It is difficult to see how spurious correlations could accidentally be introduced. Moreover, we have found similar correlations in the map produced by the COBE satellite, which used different instruments and analysis and so would have had mostly independent systematics.

The results could send us back to the drawing board about the early universe.

A more probable explanation is that an unexpected source or absorber of microwave photons is contaminating the data. This new source should somehow be associated with the solar system. Perhaps it is some unknown cloud of dust on the outskirts of our solar system. But this explanation is itself not without problems: How does one get a solar system source to glow at approximately the wavelength of the CMB brightly enough to be seen by CMB instruments, or to absorb at CMB wavelengths, yet remain sufficiently invisible in all other wavelengths not to have yet been discovered? We hope we will be able eventually to study such a foreground source well enough to decontaminate the CMB data.

Back to the Drawing Board?

AT FIRST GLANCE, the discovery of a solar system contaminant in the CMB data might appear to solve the conundrum of weak large-scale fluctuations. Actually, however, it makes the problem even worse. When we remove the part that comes from the hypothetical foreground, the remaining cosmological contribution is likely to be even smaller than previously believed. (Any other conclusion would require an accidental cancellation between the cosmic contribution and our supposed foreground source.) It would then be harder to claim that the absence of low *l* power is just a statistical accident. It looks like inflation is getting into a major jam.

A statistically robust conclusion that less power than expected exists on large scales could send us back to the drawing board about the early universe. The current alternatives to generic inflation are not terribly attractive: a carefully designed inflationary model could produce a glitch in the power spectrum at just the right scale to give us the observed absence of large-scale power, but this "designer inflation" stretches the limits of what we look for in a compelling scientific theory—an exercise akin to Ptolemy's addition of hypothetical epicycles to the orbits of heavenly bodies so that they would conform to an Earth-centered cosmology.

One possibility is that the universe has an unexpectedly complex cosmic topology [see "Is Space Finite?" by Jean-Pierre Luminet, Glenn D. Starkman and Jeffrey R. Weeks; SCIENTIFIC AMERICAN, April 1999]. If the universe is finite and wrapped around itself in interesting ways, like a doughnut or pretzel, then the vibrational modes it allows will be modified in very distinctive ways. We might be able to hear the shape of the universe, much as one can hear the difference between, say, church bells and wind chimes. For this purpose, the lowest notes—the largest-scale fluctuations—are the ones that would

most clearly echo the shape (and the size) of the universe. The universe could have an interesting topology but have been inflated precisely enough to take that topology just over the horizon, making it not just hard to see but very difficult to test.

Is there hope to resolve these questions? Yes, we expect more data from the WMAP satellite, not only on the temperature fluctuations of the sky but also on the polarization of the received light, which may help reveal foreground sources. In 2007 the European Space Agency will launch the Planck mission, which will measure the CMB at more frequency bands and at higher angular resolution than WMAP did. The higher angular resolution is not expected to help solve the low-*l* puzzle, but observing the sky in many more microwave "colors" will give us much better control over systematics and foregrounds. Cosmological research continues to bring surprises—stay tuned. SA

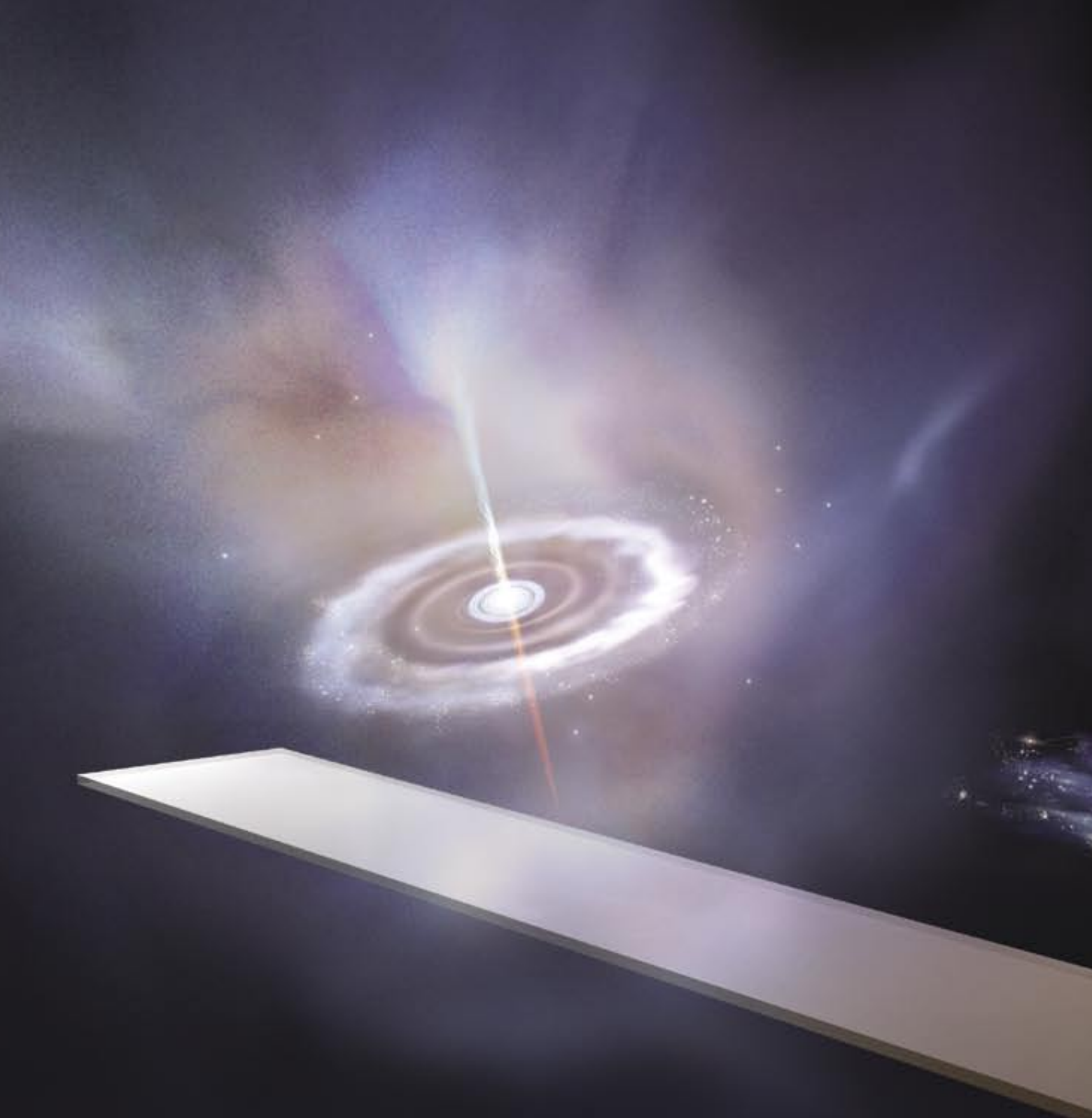
MORE TO EXPLORE

First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results. C. L. Bennett et al. in *Astrophysical Journal Supplemental,* Vol. 148, page 1; 2003.

The Cosmic Symphony. Wayne Hu and Martin White in *Scientific American,* Vol. 290, No. 2, pages 44–53; February 2004.

The WMAP Web page is at **http://wmap.gsfc.nasa.gov/**

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



COSMIC DOWNSIZING has occurred over the past 14 billion years as activity has shifted to smaller galaxies. In the first half of the universe's lifetime, giant galaxies gave birth to prodigious numbers of stars and supermassive black holes that powered brilliant quasars (*left*). In the second half, activity in the giant galaxies slowed, but star formation and black hole building continued in medium-size galaxies (*center*). In the future, the main sites of cosmic activity will be dwarf galaxies holding only a few million stars each (*right*).

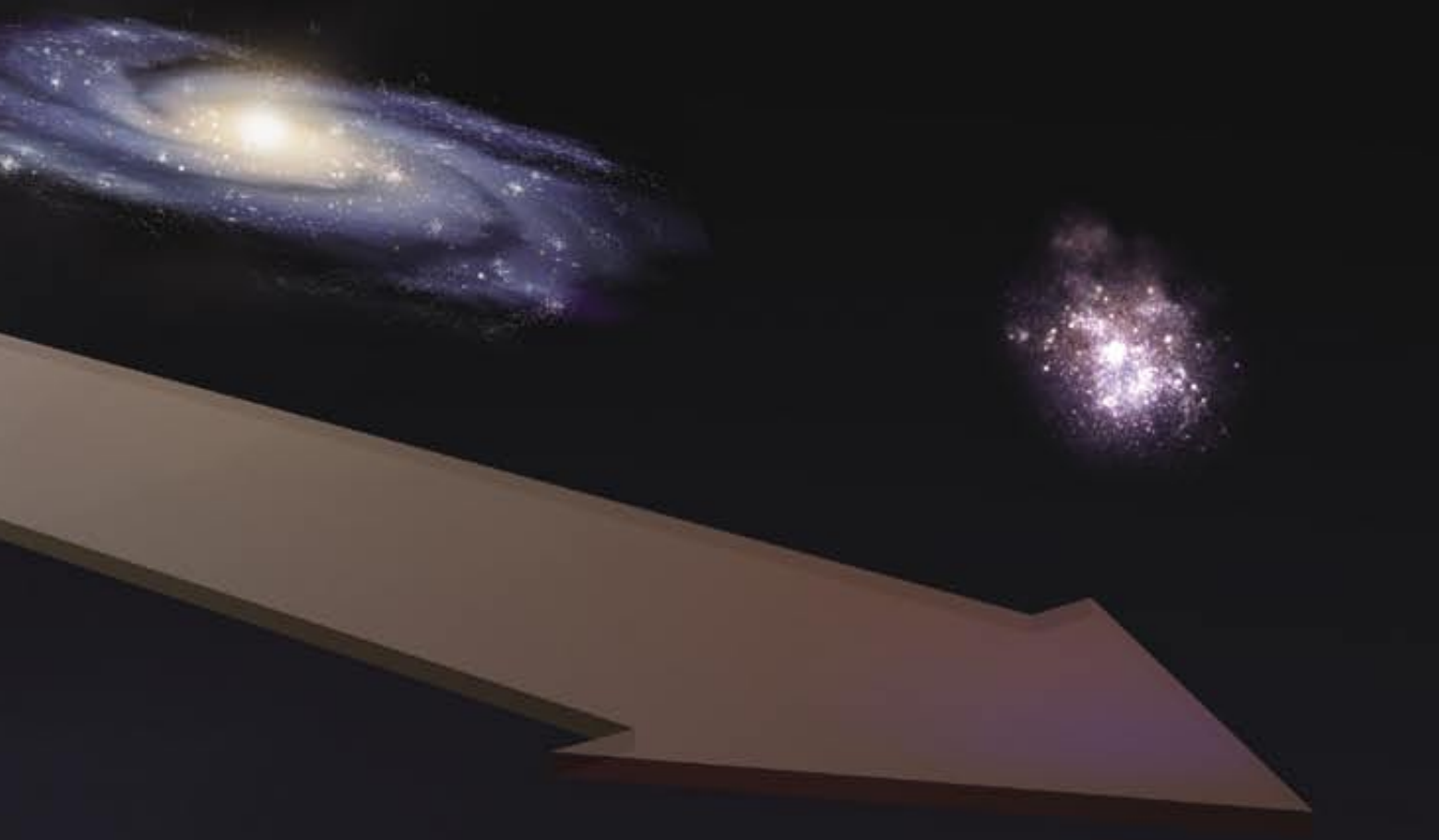
COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

the midlife CRISIS of the COSMOS

By Amy J. Barger

Although it is not as active as it used to be, the universe is still forming stars and building black holes at an impressive pace

originally published in January 2005



COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

Until recently, most astronomers believed that the universe had entered a very boring middle age. According to this paradigm, the early history of the universe—that is, until about six billion years after the big bang—was an era of cosmic fireworks: galaxies collided and merged, powerful black holes sucked in huge whirlpools of gas, and stars were born in unrivaled profusion. In the following eight billion years, in contrast, galactic mergers became much less common, the gargantuan black holes went dormant, and star formation slowed to a flicker. Many astronomers were convinced that they were witnessing the end of cosmic history and that the future held nothing but the relentless expansion of a becalmed and senescent universe.

In the past few years, however, new observations have made it clear that the reports of the universe's demise have been greatly exaggerated. With the advent of new space observatories and new instruments on ground-based telescopes, astronomers have detected violent activity occurring in nearby galaxies during the recent past. (The light from more distant galaxies takes longer to reach us, so we observe these structures in an earlier stage of development.) By examining the x-rays emitted by the cores of these relatively close galaxies, researchers have discovered many tremendously massive black holes still devouring the surrounding gas and dust. Furthermore, a more thorough study of the light emitted by galaxies of different ages has shown that the star formation rate has not declined as steeply as once believed.

The emerging consensus is that the early universe was dominated by a small number of giant galaxies containing colossal black holes and prodigious bursts of star formation, whereas the present universe has a more dispersed nature—the creation of stars and the accretion of material into black holes are now occurring in a large number of medium-size and small galaxies. Essentially, we are in the midst of a vast downsizing that is redistributing cosmic activity.

Overview/*Middle-Aged Cosmos*

- The early history of the universe was a turbulent era marked by galactic collisions, huge bursts of star formation and the creation of extremely massive black holes. The falloff in cosmic activity since then has led many astronomers to believe that the glory days of the universe are long gone.
- In recent years, though, researchers have found powerful black holes still actively consuming gas in many nearby galaxies. New observations also suggest that star formation has not dropped as steeply as once believed.
- The results point to a cosmic downsizing: whereas the early universe was dominated by a relatively small number of giant galaxies, activity in the current universe is dispersed among a large number of smaller galaxies.

Deep-Field Images

TO PIECE TOGETHER the history of the cosmos, astronomers must first make sense of the astounding multitude of objects they observe. Our most sensitive optical views of the universe come from the Hubble Space Telescope. In the Hubble Deep Field studies—10-day exposures of two tiny regions of the sky observed through four different wavelength filters—researchers have found thousands of distant galaxies, with the oldest dating back to about one billion years after the big bang. A more recent study, called the Hubble Ultra Deep Field, has revealed even older galaxies. Obtaining these deep-field images is only the beginning, however. Astronomers want to understand how the oldest and most distant objects evolved into present-day galaxies. It is somewhat like learning how a human baby grows to be an adult. Connecting the present with the past has become one of the dominant themes of modern astronomy.

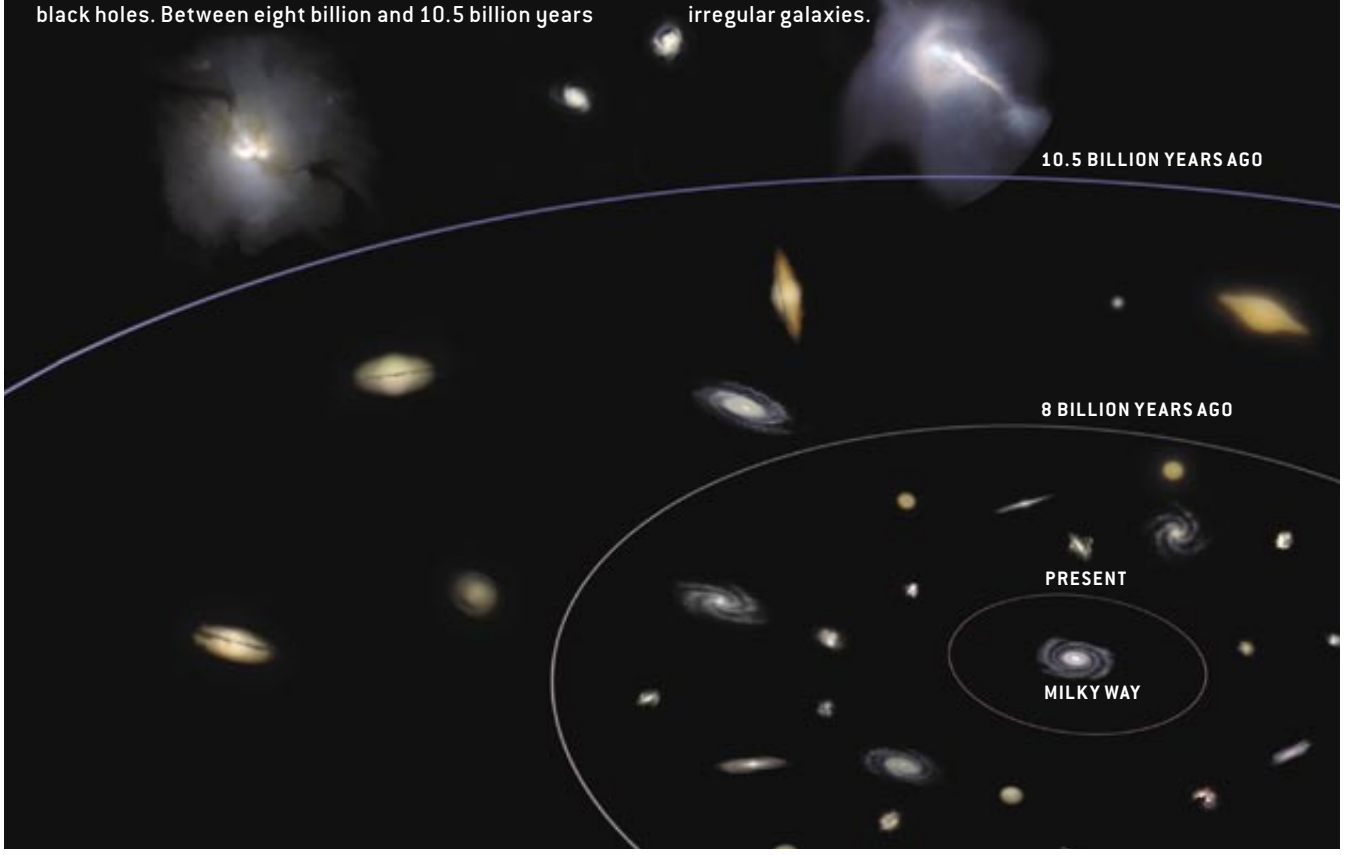
A major step in this direction is to determine the cosmic stratigraphy—which objects are in front and which are

ALFRED T. KAMAJIAN (*preceding pages*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

EVOLUTION OF THE UNIVERSE

As astronomers peer into the depths of space, they also look back in time, because the light from distant objects takes longer to reach us. More than 10.5 billion years ago, tremendous galaxies collided and merged, triggering bursts of star formation and the accretion of gas into supermassive black holes. Between eight billion and 10.5 billion years ago, stars continued to form at a high rate, and black holes continued to grow inside the galactic cores. In more recent times, star formation and black hole activity began to die down in the bigger galaxies; in the present-day universe, most of the star formation takes place in smaller spiral and irregular galaxies.



more distant—among the thousands of galaxies in a typical deep-field image. The standard way to perform this task is to obtain a spectrum of each galaxy in the image and measure its redshift. Because of the universe's expansion, the light from distant sources has been stretched, shifting its wavelength toward the red end of the spectrum. The more the light is shifted to the red, the farther away the source is and thus the older it is. For example, a redshift of one means that the wavelength has been stretched by 100 percent, that is, to twice its original size. Light from an object with this redshift was emitted about six billion years after the big bang, which is less than half the current age of the universe. In fact, astronomers usually talk in terms of redshift rather than years, because redshift is what we measure directly.

Obtaining redshifts is a practically foolproof technique for reconstructing cosmic history, but in the deepest of the deep-field images it is almost impossible to measure redshifts for all the galaxies. One reason is the sheer number of galaxies in the image, but a more fundamental problem is the intrinsic faintness of some of the galaxies. The light from these dim objects arrives at a trickle of only one photon per minute in each square centimeter. And when observers take a spectrum of the galaxy, the diffraction grating of the spectrograph disperses the light over a large area on the detector, rendering the signal even fainter at each wavelength.

In the late 1980s a team led by Lennox L. Cowie of the University of Hawaii Institute for Astronomy and Simon J. Lilly, now at the Swiss Federal Institute of Technology in Zurich, developed a novel approach to avoid the need for

THE AUTHOR

AMY J. BARGER studies the evolution of the universe by observing some of its oldest objects. She is an associate professor of astronomy at the University of Wisconsin–Madison and also holds an affiliate graduate faculty appointment at the University of Hawaii at Manoa. Barger earned her Ph.D. in astronomy in 1997 at the University of Cambridge, then did postdoctoral research at the University of Hawaii Institute for Astronomy. An observational cosmologist, she has explored the high-redshift universe using the Chandra X-ray Observatory, the Hubble Space Telescope, and the telescopes on Kitt Peak in Arizona and on Mauna Kea in Hawaii.

ALFRED T. KAMAJIAN

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

laborious redshift observations. The researchers observed regions of the sky with filters that selected narrow wavebands in the ultraviolet, green and red parts of the spectrum and then measured how bright the galaxies were in each of the wavebands [*see box on page 15*]. A nearby star-forming galaxy is equally bright in all three wavebands. The intrinsic light from a star-forming galaxy has a sharp cutoff just beyond the ultraviolet waveband, at a wavelength of about 912 angstroms. (The cutoff appears because the neutral hydrogen gas in and around the galaxy absorbs radiation with shorter wavelengths.) Because the light from distant galaxies is shifted to the red, the cutoff moves to longer wavelengths; if the redshift is great enough, the galaxy's light will not appear in the ultraviolet waveband, and if the redshift is greater still, the galaxy will not be visible in the green waveband either.

Thus, Cowie and Lilly could separate star-forming galaxies into broad redshift intervals that roughly indicated their ages. In 1996 Charles C. Steidel of the California Institute of Technology and his collaborators used this technique to isolate hundreds of ancient star-forming galaxies with redshifts of about three, dating from about two billion years after the big bang. The researchers confirmed many of the estimated redshifts by obtaining very deep spectra of the galaxies with the powerful 10-meter Keck telescope on Mauna Kea in Hawaii.

Once the redshifts of the galaxies have been measured, we can begin to reconstruct the history of star formation. We know from observations of nearby galaxies that a small number of high-mass stars and a larger number of low-mass stars usually form at the same time. For every 20 sunlike stars that are born, only one 10-solar-mass star (that is, a star with a mass 10 times as great as the sun's) is created. High-mass stars emit ultraviolet and blue light, whereas low-mass stars emit yellow and red light. If the redshift of a distant galaxy is known, astronomers can determine the galaxy's intrinsic spectrum (also called the rest-frame spectrum). Then, by measuring the total amount of rest-frame ultraviolet light, researchers can estimate the number of high-mass stars in the galaxy.

Because high-mass stars live for only a few tens of millions of years—a short time by galactic standards—their number closely tracks variations in the galaxy's overall star formation rate. As the pace of star creation slows, the number of high-mass stars declines soon afterward because they die so quickly after they are born. In our own Milky Way, which is quite typical of nearby, massive spiral galaxies, the number of observed high-mass stars indicates that stars are forming at a rate of a few solar masses a year. In high-redshift galaxies, however, the rate of star formation is 10 times as great.

When Cowie and Lilly calculated the star formation rates in all the galaxies they observed, they came to the remarkable conclusion that the universe underwent a veritable baby boom at a redshift of about one. In 1996 Piero Madau, now at the University of California at Santa Cruz, put the technique to work on the Hubble Deep Field North data, which were ideal for this approach because of the very precise intensity measurements in four wavebands. Madau combined his results with those from existing lower-redshift optical observations to refine the estimates of the star formation history of the universe. He inferred that the rate of star formation must have peaked when the universe was about four billion to six billion years old. This result led many astronomers to conclude that the universe's best days were far behind it.

An Absorbing Tale

ALTHOUGH MADAU'S ANALYSIS of star formation history was an important milestone, it was only a small part of the story. Galaxy surveys using optical telescopes cannot detect every source in the early universe. The more distant a galaxy is, the more it suffers from cosmological redshifting, and at high enough redshifts, the galaxy's rest-frame ultraviolet and optical emissions will be stretched into the infrared part of the spectrum. Furthermore, stars tend to reside in very dusty environments because of the detritus from supernova explosions and other processes. The starlight heats up the dust grains, which then reradiate this energy at far-infrared wavelengths. For very distant sources, the light that is absorbed by dust and reradiated into the far-infrared is shifted by the expansion of the universe to submillimeter wavelengths. Therefore, a bright source of submillimeter light is often a sign of intense star formation.

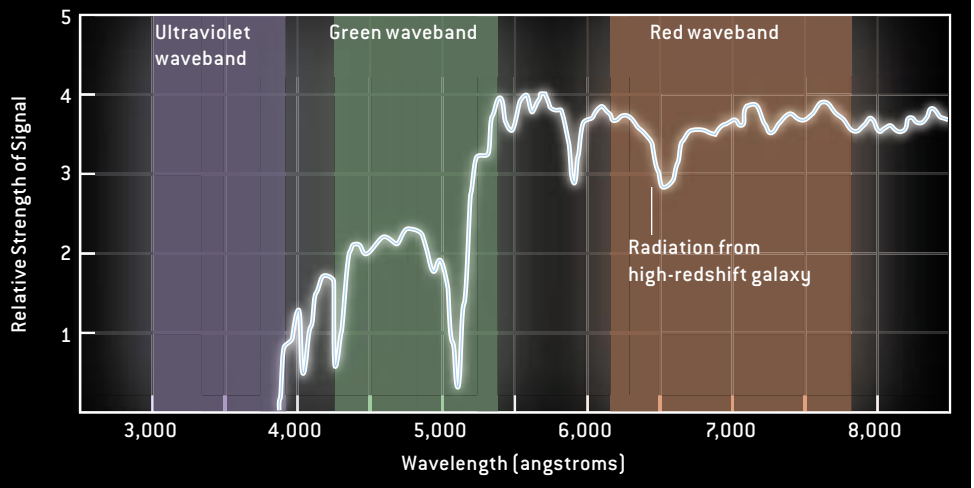
Until recently, astronomers found it difficult to make submillimeter observations with ground-based telescopes, partly because water vapor in the atmosphere absorbs signals of that wavelength. But those difficulties were eased with the introduction of the Submilli-

New observations make it clear that reports of the UNIVERSE'S DEMISE have been greatly exaggerated.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

FINDING ANCIENT GALAXIES

To efficiently detect the oldest galaxies in a survey field, astronomers have developed a technique employing filters that select wavebands in the ultraviolet, green and red parts of the spectrum. Because of the expansion of the universe, the light from the oldest galaxies has been shifted toward the red end; the graph shows how a relatively high redshift (about three) can push the radiation from a distant galaxy out of the ultraviolet waveband. As a result, the ancient galaxies appear in images made with the red and green filters but not in images made with the ultraviolet filter.



meter Common-User Bolometer Array (SCUBA), a camera that was installed on the James Clerk Maxwell Telescope on Mauna Kea in 1997. (Located at a height of four kilometers above sea level, the observatory is above 97 percent of the water in the atmosphere.) Several teams of researchers, one of which I led, used SCUBA to directly image regions of the sky with sufficient sensitivity and area coverage to discover distant, exceptionally luminous dust-obscured sources. Because the resolution is fairly coarse, the galaxies have a bloblike appearance. They are also relatively rare—even after many hours of exposure, few sources appeared on each SCUBA image—but they are among the most luminous galaxies in the universe. It is sobering to realize that before SCUBA became available, we did not even know that these powerful, distant systems existed! Their star formation rates are hundreds of times greater than those of present-day galaxies, another indication that the universe used to be much more exciting than it is now.

Finding all this previously hidden star formation was revolutionary, but might the universe be covering up other violent activity? For example, gas and dust within galaxies could also be obscuring the radiation emitted by the disks of material whirling around supermassive black holes (those weighing as much as billions of suns). These disks are believed to be the power sources of quasars, the prodigiously luminous objects found at high redshifts, as well as the active nuclei at the centers of many nearby galaxies. Optical studies in the 1980s suggested that there were far more quasars several billion years after the big bang than there are active galactic nuclei in the present-day universe. Because the supermassive black holes that powered the distant quasar activity cannot be destroyed, astronomers presumed that many nearby galaxies must contain dead quasars—black holes that have exhausted their fuel supply.

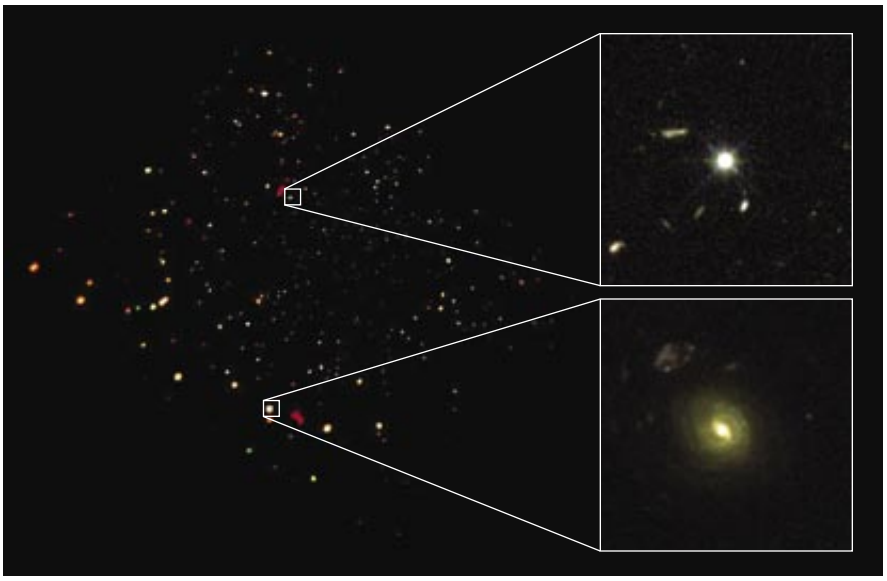
These dormant supermassive black holes have indeed been detected through their gravitational influence. Stars and gas continue to orbit around the holes even though little material is swirling into them. In fact, a nearly dormant black hole resides at the center of the Milky Way. Together these results led scientists to develop a scenario: most supermassive black holes formed during the quasar era, consumed all the material surrounding them in a violent fit of growth and then disappeared from optical observations once their fuel supply ran out. In short, quasar activity, like star formation, was more vigorous in the distant past, a third sign that we live in relatively boring times.

This scenario, however, is incomplete. By combining x-ray and visible-light observations, astronomers are now revisiting the conclusion that the vast majority of quasars died out long ago. X-rays are important because, unlike visible light, they can pass through the gas and dust surrounding hidden black holes. But x-rays are blocked by the earth's atmosphere, so researchers must rely on space telescopes such as the Chandra and XMM/Newton X-ray observatories to detect black hole activity [see "The Cosmic Reality Check," by Günther Hasinger and Roberto Gilli; SCIENTIFIC AMERICAN, March 2002]. In 2000 a team consisting of Cowie, Richard F. Mushotzky of the NASA Goddard Space Flight Center, Eric A. Richards, then at Arizona State University, and I used the Subaru telescope at Mauna Kea to identify optical counterparts to 20 x-ray sources found by Chandra in a survey field. We then employed the 10-meter Keck telescope to obtain the spectra of these objects.

Our result was quite unexpected:

TOMMY MOORMAN

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



X-RAY VISION can be used to find hidden black holes. The Chandra X-ray Observatory detected many black holes in its Deep Field North survey (*left*). Some were ancient, powering brilliant quasars that flourished just a few billion years after the big bang (*top right*). But others lurked in the centers of relatively nearby galaxies, still generating x-rays in the modern era (*bottom right*).

many of the active supermassive black holes detected by Chandra reside in relatively nearby, luminous galaxies. Modelers of the cosmic x-ray background had predicted the existence of a large population of obscured supermassive black holes, but they had not expected them to be so close at hand! Moreover, the optical spectra of many of these galaxies showed absolutely no evidence of black hole activity; without the x-ray observations, astronomers could never have discovered the supermassive black holes lurking in their cores.

This research suggests that not all supermassive black holes were formed in the quasar era. These mighty objects have apparently been assembling from the earliest times until the present. The supermassive black holes that are still active, however, do not exhibit the same behavioral patterns as the distant quasars. Quasars are voracious consumers, greedily gobbling up the material around them at an enormous rate. In contrast, most of the nearby sources that Chandra detected are more moderate eaters and thus radiate less intensely. Scientists have not yet determined what mechanism is responsible for this vastly different behavior. One possibility is that the present-day black holes have less gas to consume. Nearby galaxies undergo fewer collisions than the distant, ancient galaxies did, and such collisions could drive material into the supermassive black holes at the galactic centers.

Chandra had yet another secret to reveal: although the moderate x-ray sources were much less luminous than the quasars—generating as little as 1 percent of the radiation emitted by their older counterparts—when we added up the light produced by all the moderate sources in recent times, we found the amount to be about one tenth of that produced by the quasars in early times. The only way this result could arise is if there are many more moderate black holes active now than there were quasars active in the past. In other words, the contents of the universe have transitioned from a small number of bright objects to a large number of dimmer ones. Even though supermassive black holes are now being built smaller and cheaper, their combined effect is still potent.

Star-forming galaxies have also undergone a cosmic downsizing. Although some nearby galaxies are just as extravagant in their star-forming habits as the extremely luminous, dust-obscured galaxies found in the SCUBA images, the density of ultraluminous galaxies in the present-day universe is more than 400 times lower than their density in the distant universe. Again, however, smaller galaxies have taken up some of the slack. A team consisting of Cowie, Gillian Wilson, now at NASA's Infrared Processing and Analysis Center, Doug J. Burke, now at the Harvard-Smithsonian Center for Astrophysics, and I has refined the estimates of the universe's luminosity density by studying high-quality images produced with a wide range of filters and performing a complete spectroscopic follow-up. We found that the luminosity density of optical and ultraviolet light has not changed all that much with cosmic time. Although the overall star formation rate has dropped in the second half of the universe's lifetime because the monstrous dusty galaxies are no longer bursting with stars, the population of small, nearby star-forming galaxies is so numerous that the density of optical and ultraviolet light is declining rather gradually. This result gives us a much more optimistic outlook on the continuing health of the universe.

Middle-Aged Vigor

THE EMERGING PICTURE of continued vigor fits well with cosmological theory. New computer simulations suggest that the shift from a universe dominated by a few large and powerful galaxies to a universe filled with many smaller and meeker galaxies may be a direct consequence of cosmic expansion. As the universe expands, galaxies become more separated and mergers become rarer. Furthermore, as the gas surrounding galaxies grows more diffuse, it becomes easier to heat. Because hot gas is more energetic than cold gas, it does not gravitationally collapse as readily into the galaxy's potential well. Fabrizio Nicastro of the Harvard-Smithsonian Center for Astrophysics and his co-workers have recently detected a warm intergalactic fog through its

D. M. ALEXANDER, F. E. BAUER, W. N. BRANDT, G. P. GARMIRE ET AL. *NASA/Pennsylvania State University (Chandra image);*
A. J. BARGER AND M. GIAVALISCO *NASA/Great Observatories Origins Deep Survey (optical images)*

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

absorption of ultraviolet light and x-rays from distant quasars and active galactic nuclei. This warm fog surrounds our galaxy in every direction and is part of the Local Group of galaxies, which includes the Milky Way, Andromeda and 30 smaller galaxies. Most likely this gaseous material was left over from the galaxy formation process but is too warm to permit further galaxy formation to take place.

Small galaxies may lie in cooler environments because they may not have heated their surrounding regions of gas to the same extent that the big galaxies did through supernova explosions and quasar energy. Also, the small galaxies may have consumed less of their surrounding material, allowing them to continue their more modest lifestyles to the present day. In contrast, the larger and more profligate galaxies have exhausted their resources and are no longer able to collect more from their environments. Ongoing observational studies of the gaseous properties of small, nearby galaxies may reveal how they interact with their environments and thus provide a key to understanding galactic evolution.

But a crucial part of the puzzle remains unsolved: How did the universe form monster quasars so early in its history? The Sloan Digital Sky Survey, a major astronomical project to map one quarter of the entire sky and measure distances to more than a million remote objects, has discovered quasars that existed when the universe was only one sixteenth of its present age, about 800 million years after the big bang. In 2003 Fabian Walter, then at the National Radio Astronomy Observatory, and his collaborators detected the presence of carbon monoxide in the emission from one of these quasars; because carbon and oxygen could have been created only from the thermonuclear reactions in stars, this discovery suggests that a significant amount of star formation occurred in the universe's first several hundred million years. Recent results from the Wilkinson Microwave Anisotropy Probe, a satellite that studies the cosmic background radiation, also indicate that star formation began just 200 million years after the big bang.

Furthermore, computer simulations have shown that the first stars were most likely hundreds of times as massive as the sun. Such stars would have burned so brightly that they would have run out of fuel in just a few tens of millions of years; then the heaviest stars would have collapsed to black holes, which could have formed the seeds of the supermassive black holes that powered the first quasars. This explanation for the early appearance of quasars may be bolstered by the further study of gamma-ray bursts, which are believed to result from the collapse of very massive stars into black holes. Because gamma-ray bursts are the most powerful explosions in the universe since the big bang, astronomers can detect them at very great distances. This past November, NASA was expected to launch the Swift Gamma-Ray Burst Mission, a $250-million satellite with three telescopes designed to observe the explosions in the gamma-ray, x-ray, ultraviolet and optical wavelengths. By measuring the spectra of the gamma-ray bursts and their afterglows, the Swift satellite could provide scientists with a much better understanding of how collapsing stars could have started the growth of supermassive black holes in the early universe.

In comic books, Superman looked through walls with his x-ray vision. Astronomers have now acquired a similar ability with the Chandra and XMM/Newton observatories and are making good use of it to peer deep into the dust-enshrouded regions of the universe. What is being revealed is a dramatic transition from the mighty to the meek. The giant star-forming galaxies and voracious black holes of the universe's past are now moribund. A few billion years from now, the smaller galaxies that are active today will have consumed much of their fuel, and the total cosmic output of radiation will decline dramatically. Even our own Milky Way will someday face this same fate. As the cosmic downsizing continues, the dwarf galaxies—which hold only a few million stars each but are the most numerous type of galaxy in the universe—will become the primary hot spots of star formation. Inevitably, though, the universe will darken, and its only contents will be the fossils of galaxies from its glorious past. Old galaxies never die, they just fade away. SA

What is being revealed is a dramatic transition from the MIGHTY TO THE MEEK. Dwarf galaxies will become the PRIMARY HOT SPOTS of star formation.

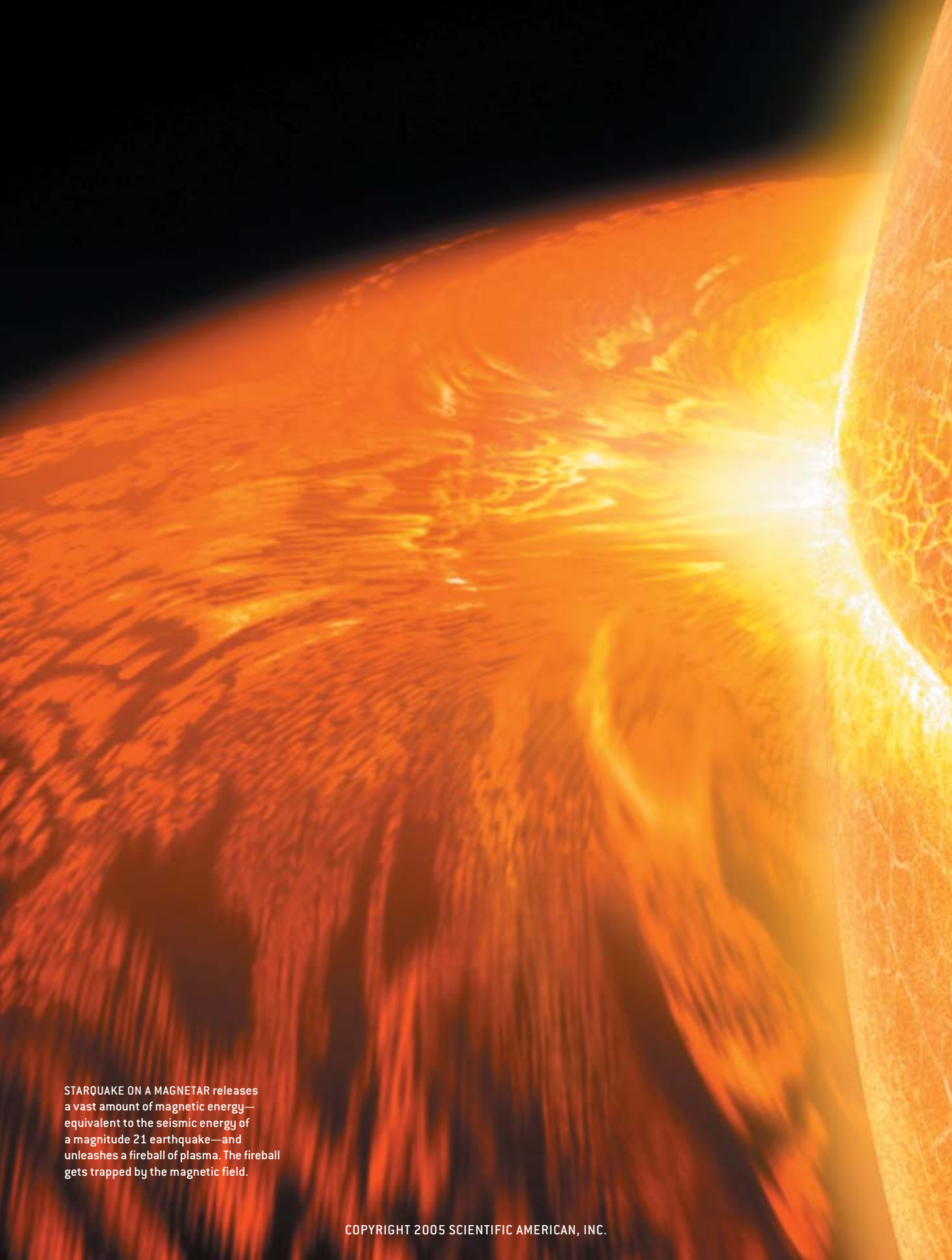
MORE TO EXPLORE

Star Formation History since z = 1 as Inferred from Rest-Frame Ultraviolet Luminosity Density Evolution. Gillian Wilson et al. in *Astronomical Journal,* Vol. 124, pages 1258–1265; September 2002. Available online at **www.arxiv.org/abs/astro-ph/0203168**

The Cosmic Evolution of Hard X-ray Selected Active Galactic Nuclei. Amy J. Barger et al. in *Astronomical Journal* (in press). Available online at **www.arxiv.org/abs/astro-ph/0410527**

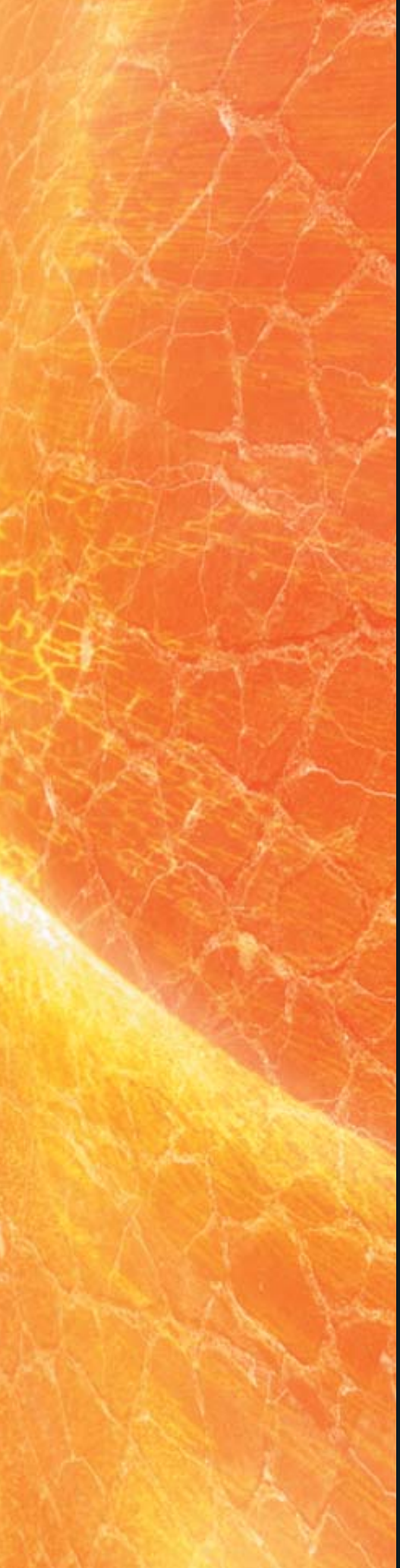
Supermassive Black Holes in the Distant Universe. Edited by Amy J. Barger. Astrophysics and Space Science Library, Vol. 308. Springer, 2004.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



STARQUAKE ON A MAGNETAR releases a vast amount of magnetic energy—equivalent to the seismic energy of a magnitude 21 earthquake—and unleashes a fireball of plasma. The fireball gets trapped by the magnetic field.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



MAGNETARS

Some stars are magnetized so intensely that they emit huge bursts of magnetic energy and alter the very nature of the quantum vacuum

By Chryssa Kouveliotou, Robert C. Duncan and Christopher Thompson

originally published in February 2003

On March 5, 1979, several months after dropping probes into the toxic atmosphere of Venus, two Soviet spacecraft, Venera 11 and 12, were drifting through the inner solar system on an elliptical orbit. It had been an uneventful cruise. The radiation readings on board both probes hovered around a nominal 100 counts per second. But at 10:51 A.M. EST, a pulse of gamma radiation hit them. Within a fraction of a millisecond, the radiation level shot above 200,000 counts per second and quickly went off scale.

Eleven seconds later gamma rays swamped the NASA space probe Helios 2, also orbiting the sun. A plane wave front of high-energy radiation was evidently sweeping through the solar system. It soon reached Venus and saturated the Pioneer Venus Orbiter's detector. Within seconds

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

the gamma rays reached Earth. They flooded detectors on three U.S. Department of Defense Vela satellites, the Soviet Prognoz 7 satellite, and the Einstein Observatory. Finally, on its way out of the solar system, the wave also blitzed the International Sun-Earth Explorer.

The pulse of highly energetic, or "hard," gamma rays was 100 times as intense as any previous burst of gamma rays detected from beyond the solar system, and it lasted just two tenths of a second. At the time, nobody noticed; life continued calmly beneath our planet's protective atmosphere. Fortunately, all 10 spacecraft survived the trauma without permanent damage. The hard pulse was followed by a fainter glow of lower-energy, or "soft," gamma rays, as well as x-rays, which steadily faded over the subsequent three minutes. As it faded away, the signal oscillated gently, with a period of eight seconds. Fourteen and a half hours later, at 1:17 A.M. on March 6, another, fainter burst of x-rays came from the same spot on the sky. Over the ensuing four years, Evgeny P. Mazets of the Ioffe Institute in St. Petersburg, Russia, and his collaborators detected 16 bursts coming from the same direction. They varied in intensity, but all were fainter and shorter than the March 5 burst.

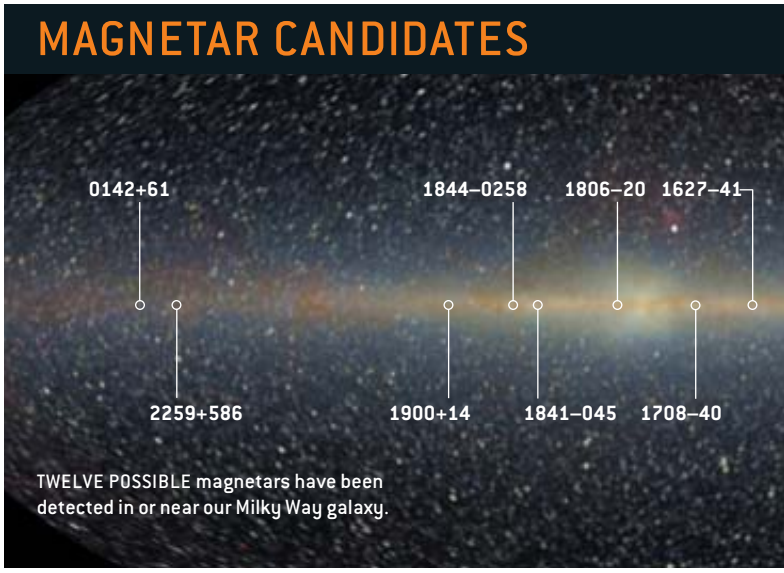
Astronomers had never seen anything like this. For want of a better idea, they initially listed these bursts in catalogues alongside the better-known gamma-ray bursts (GRBs), even though they clearly differed in several ways. In the mid-1980s Kevin C. Hurley of the University of California at Berkeley realized that similar outbursts were coming from two other areas of the sky. Evidently these sources were all repeating—unlike GRBs, which are one-shot events [see "The Brightest Explosions in the Universe," by Neil Gehrels, Luigi Piro and Peter J. T. Leonard; SCIENTIFIC AMERICAN, December 2002]. At a July 1986 meeting in Toulouse, France, astronomers agreed on the approximate locations of the three sources and dubbed them "soft gamma repeaters" (SGRs). The alphabet soup of astronomy had gained a new ingredient.

Another seven years passed before two of us (Duncan and Thompson) devised an explanation for these strange objects, and only in 1998 did one of us (Kouveliotou) and her team find compelling evidence for that explanation. Recent observations connect our theory to yet another class of celestial enigmas, known as anomalous x-ray pulsars (AXPs). These developments have led to a breakthrough in our understanding of one of the most exotic members of the celestial bestiary, the neutron star.

Neutron stars are the densest material objects known, packing slightly more than the sun's mass inside a ball just 20 kilometers across. Based on the study of SGRs, it seems that some neutron stars have magnetic fields so intense that they radically alter the material within them and the state of the quantum vacuum surrounding them, leading to physical effects observed nowhere else in the universe.

Overview/*Ultramagnetic Stars*

- Astronomers have seen a handful of stars that put out flares of gamma and x-radiation, which can be millions of times as bright as any other repeating outburst known. The enormous energies and pulsing signals implicate the second most extreme type of body in the universe (after the black hole): the neutron star.
- These neutron stars have the strongest magnetic fields ever measured—hence their name, magnetars. Magnetic instabilities analogous to earthquakes can account for the flares.
- Magnetars remain active for only about 10,000 years, implying that millions of them are drifting undetected through our galaxy.



Not Supposed to Do That

BECAUSE THE MARCH 1979 BURST was so bright, theorists at the time reckoned that its source was in our galactic neighborhood, hundreds of light-years from Earth at most. If that had been true, the intensity of the x-rays and gamma rays would have been just below the theoretical maximum steady luminosity that can be emitted by a star. That maximum, first derived in 1926 by English astrophysicist Arthur Eddington, is set by the force of radiation flowing through the hot outer layers of a star. If the radiation is any more intense, it will overpower gravity, blow away ionized matter and destabilize the star. Emission below the Eddington limit would have been fairly straightforward to explain. For example, various theorists proposed that the outburst was triggered by the impact of a chunk of matter, such as an asteroid or a comet, onto a nearby neutron star.

But observations soon confounded that hypothesis. Each spacecraft had recorded the time of arrival of the hard initial pulse. These data allowed astronomers, led by Thomas Lytton Cline of the NASA Goddard Space Flight Center, to triangulate the burst source. The researchers found that the position coincided with the Large Magellanic Cloud, a small galaxy about 170,000 light-years away. More specifically, the event's position matched that of a young supernova remnant, the glowing

DON DIXON (*preceding pages*); E. L. WRIGHT *University of California at Los Angeles*, THE COBE PROJECT, DIRBE AND NASA (*above*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

remains of a star that exploded 5,000 years ago. Unless this overlap was pure coincidence, it put the source 1,000 times as far away as theorists had thought—and thus made it a million times brighter than the Eddington limit. In 0.2 second the March 1979 event released as much energy as the sun radiates in roughly 10,000 years, and it concentrated that energy in gamma rays rather than spreading it across the electromagnetic spectrum.

No ordinary star could account for such energy, so the source was almost certainly something out of the ordinary—either a black hole or a neutron star. The former was ruled out by the eight-second modulation: a black hole is a featureless object, lacking the structure needed to produce regular pulses. The association with the supernova remnant further strengthened the case for a neutron star. Neutron stars are widely believed to form when the core of a massive but otherwise ordinary star exhausts its nuclear fuel and abruptly collapses under its own weight, thereby triggering a supernova explosion.

Identifying the source as a neutron star did not solve the puzzle; on the contrary, it merely heightened the mystery. Astronomers knew several examples of neutron stars that lie within supernova remnants. These stars were radio pulsars, objects that are observed to blink on and off in radio waves. Yet the March 1979 burster, with an apparent rotation period of eight seconds, was spinning much more slowly than any radio pulsar then known. Even when not bursting, the object emitted a steady glow of x-rays with more radiant power than could be supplied by the rotation of a neutron star. Oddly, the star was significantly displaced from the center of the supernova remnant. If it was born at the center, as is likely, then it must have recoiled with a velocity of about 1,000 kilometers per second at birth. Such high speed was considered unusual for a neutron star.

Finally, the outbursts themselves seemed inexplicable. X-ray flashes had previously been detected from some neutron stars, but they never exceeded the Eddington limit by very much. Astronomers ascribed them to thermonuclear fusion of hydrogen or helium or to the sudden accretion of matter onto the star. But the brightness of the SGR bursts was unprecedented, so a new physical mechanism seemed to be required.

Spin Forever Down

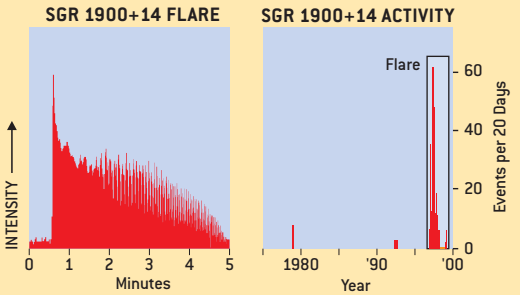
THE FINAL BURST FROM the March 1979 source was detected in May 1983; none has been seen in the 19 years since. Two other SGRs, both within our Milky Way galaxy, went off in 1979 and have remained active, emitting hundreds of bursts in the years since. A fourth SGR was located in 1998. Three of these four objects have possible, but unproved, associations with young supernova remnants. Two also lie near very dense clusters of massive young stars, intimating that SGRs tend to form from such stars. A fifth candidate SGR has gone off only twice; its precise location is still unknown.

As Los Alamos National Laboratory scientists Baolian L. Cheng, Richard I. Epstein, Robert A. Guyer and C. Alex Young pointed out in 1996, SGR bursts are statistically similar to earthquakes. The energies have very similar mathematical distributions, with less energetic events being more common. Our graduate student Ersin Gögüs of the University of Alabama at Huntsville verified this behavior for a large sample of bursts from various sources. This and other statistical properties are a hallmark of self-organized criticality, whereby a composite system attains a critical state in which a small perturbation can trigger a chain reaction. Such behavior occurs in systems as diverse as avalanches on sandpiles and magnetic flares on the sun.

But why would a neutron star behave like this? The solution emerged from an entirely separate line of work, on radio pulsars. Pulsars are widely thought to be rapidly rotating, magnetized neutron stars. The magnetic field, which is supported by electric currents flowing deep inside the star, rotates with the star. Beams of radio waves shine outward from the star's magnetic poles and sweep through space as it rotates, like lighthouse beacons—hence the observed pulsing. The pulsar also blows out a wind of charged particles and low-frequency electromag-

NAME	YEAR OF DISCOVERY	ROTATION PERIOD (seconds)
SGR 0526–66	1979	8.0
SGR 1900+14	1979	5.16
SGR 1806–20	1979	7.47
SGR 1801–23*	1997	?
SGR 1627–41	1998	?
AXP 1E 2259+586	1981	6.98
AXP 1E 1048–59†	1985	6.45
AXP 4U 0142+61	1993	8.69
AXP 1RXS 1708–40†	1997	11.0
AXP 1E 1841–045	1997	11.8
AXP AXJ1844–0258	1998	6.97
AXP CXJ0110–7211†	2002	5.44

* Not shown on map; location not known precisely
† Abbreviated name



GIANT X-RAY FLARE in August 1998 confirmed the existence of magnetars. It started with a spike of radiation lasting less than a second (*left*). Then came an extended train of pulses with a period of 5.16 seconds. This event was the most powerful outburst to come from the object, designated SGR 1900+14, since its discovery in 1979 (*right*).

SOURCE FOR TABLE: CHRYSSA KOUVELIOTOU, ROBERT C. DUNCAN AND CHRISTOPHER THOMPSON

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

netic waves, which carry away energy and angular momentum, causing its rate of spin to decrease gradually.

Perhaps the most famous pulsar lies within the Crab Nebula, the remnant of a supernova explosion that was observed in 1054. The pulsar rotates once every 33 milliseconds and is currently slowing at a rate of about 1.3 millisecond every century. Extrapolating backward, it was born rotating once every 20 milliseconds. Astronomers expect it to continue to spin down, eventually reaching a point when its rotation will be too slow to power the radio pulses. The spin-down rate has been measured for almost every radio pulsar, and theory indicates that it depends on the strength of the star's magnetic field. From this, most young radio pulsars are inferred to have magnetic fields between 10^{12} and 10^{13} gauss. For comparison, a refrigerator magnet has a strength of about 100 gauss.

The Ultimate Convection Oven

THIS PICTURE LEAVES a basic question unanswered: Where did the magnetic field come from in the first place? The traditional assumption was: it is as it is, because it was as it was. That is, most astronomers supposed that the magnetic field is a relic of the time before the star went supernova. All stars have weak magnetic fields, and those fields can be amplified simply by the act of compression. According to Maxwell's equations of electromagnetism, as a magnetized object shrinks by a factor of two, its magnetic field strengthens by a factor of four. The core of a massive star collapses by a factor of 10^5 from its birth through neutron star formation, so its magnetic field should become 10^{10} times stronger.

If the core magnetic field started with sufficient strength, this compression could explain pulsar magnetism. Unfortunately, the magnetic field deep inside a star cannot be measured, so this simple hypothesis cannot be tested. There are also good reasons to believe that compression is only part of the story.

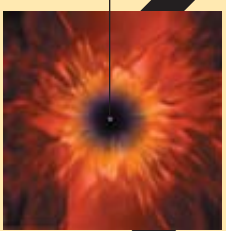
Within a star, gas can circulate by convection. Warm parcels of ionized gas rise, and cold ones sink. Because ionized gas conducts electricity well, any magnetic field lines threading the gas are dragged with it as it moves. The field can thus be reworked and sometimes amplified. This phenomenon, known as dynamo action, is thought to generate the magnetic fields of stars and planets. A dynamo might operate during each phase of the life of a massive star, as long as the turbulent core is rotating rapidly enough. Moreover, during a brief period after the core of the star turns into a neutron star, convection is especially violent.

This was first shown in computer simulations in 1986 by Adam Burrows of the University of Arizona and James M. Lattimer of the State University of New York at Stony Brook. They found that temperatures in a newborn neutron star exceed 30

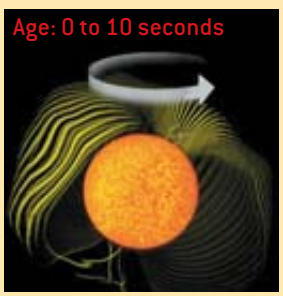
TWO TYPES OF NEUTRON STARS

1 Most neutron stars are thought to begin as massive but otherwise ordinary stars, between eight and 20 times as heavy as the sun.

2 Massive stars die in a type II supernova explosion, as the stellar core implodes into a dense ball of subatomic particles.



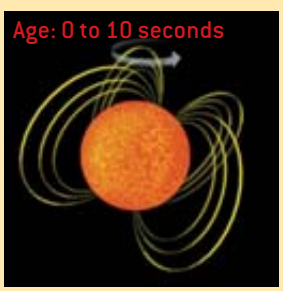
3A: If the newborn neutron star spins fast enough, it generates an intense magnetic field. Field lines inside the star get twisted.



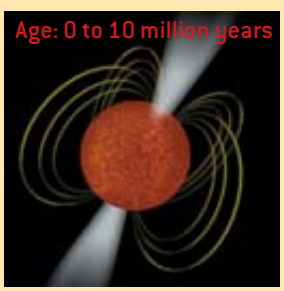
4A: The magnetar settles into neat layers, with twisted field lines inside and smooth lines outside. It might emit a narrow radio beam.

5A: The old magnetar has cooled off, and much of its magnetism has decayed away. It emits very little energy.

3B: If the newborn neutron star spins slowly, its magnetic field, though strong by everyday standards, does not reach magnetar levels.



4B: The mature pulsar is cooler than a magnetar of equal age. It emits a broad radio beam, which radio telescopes can readily detect.



5B: The old pulsar has cooled off and no longer emits a radio beam.

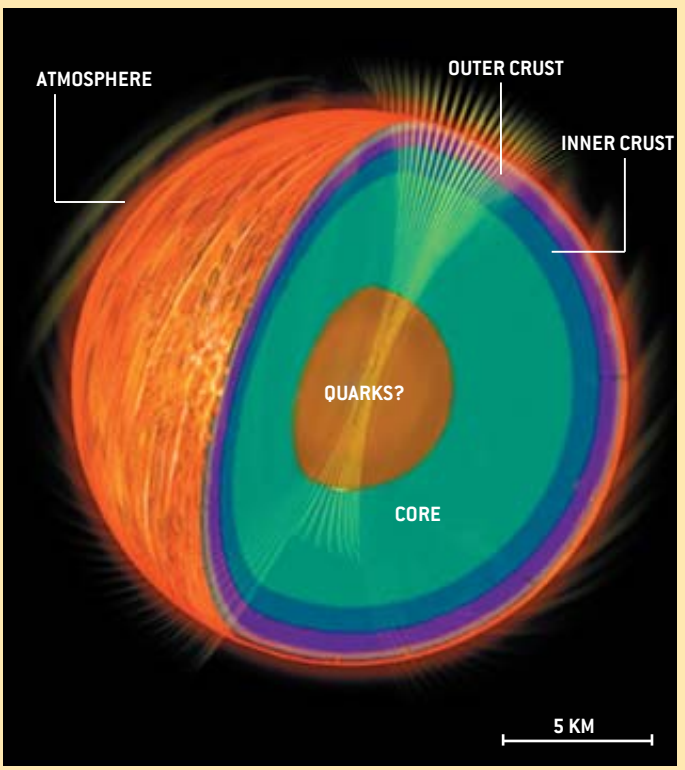
COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

billion kelvins. Hot nuclear fluid circulates in 10 milliseconds or less, carrying enormous kinetic energy. After about 10 seconds, the convection ceases.

Not long after Burrows and Lattimer conducted their first simulations, Duncan and Thompson, then at Princeton University, estimated what this furious convection means for neutron-star magnetism. The sun, which undergoes a sedate version of the same process, can be used as a reference point. As solar fluid circulates, it drags along magnetic field lines and gives up about 10 percent of its kinetic energy to the field. If the moving fluid in a newborn neutron star also transfers a tenth of its kinetic energy to the magnetic field, then the field would grow stronger than 10^{15} gauss, which is more than 1,000 times as strong as the fields of most radio pulsars.

Whether the dynamo operates globally (rather than in limited regions) would depend on whether the star's rate of rotation was comparable to its rate of convection. Deep inside the sun, these two rates are similar, and the magnetic field is able to organize itself on large scales. By analogy, a neutron star born rotating as fast as or faster than the convective period of 10 milliseconds could develop a widespread, ultrastrong magnetic field. In 1992 we named these hypothetical neutron stars "magnetars."

An upper limit to neutron-star magnetism is about 10^{17} gauss; beyond this limit, the fluid inside the star would tend to mix and the field would dissipate. No known objects in the universe can generate and maintain fields stronger than this level. One ramification of our calculations is that radio pulsars are neutron stars in which the large-scale dynamo has *failed* to operate. In the case of the Crab pulsar, the newborn neutron star rotated once every 20 milliseconds, much slower than the rate of convection, so the dynamo never got going.



STRUCTURE OF A NEUTRON STAR can be inferred from theories of nuclear matter. Starquakes can occur in the crust, a lattice of atomic nuclei and electrons. The core consists mainly of neutrons and perhaps quarks. An atmosphere of hot plasma might extend a grand total of a few centimeters.

Crinkle Twinkle Little Magnetar

ALTHOUGH WE DID NOT develop the magnetar concept to explain SGRs, its implications soon became apparent to us. The magnetic field should act as a strong brake on a magnetar's rotation. Within 5,000 years a field of 10^{15} gauss would slow the spin rate to once every eight seconds—neatly explaining the oscillations observed during the March 1979 outburst.

As the field evolves, it changes shape, driving electric currents along the field lines outside the star. These currents, in turn, generate x-rays. Meanwhile, as the magnetic field moves through the solid crust of a magnetar, it bends and stretches the crust. This process heats the interior of the star and occasionally breaks the crust in a powerful "starquake." The accompanying release of magnetic energy creates a dense cloud of electrons and positrons, as well as a sudden burst of soft gamma rays—accounting for the fainter bursts that give SGRs their name.

More infrequently, the magnetic field becomes unstable and undergoes a large-scale rearrangement. Similar (but smaller) upheavals sometimes happen on the sun, leading to solar flares. A magnetar easily has enough energy to power a giant flare such as the March 1979 event. Theory indicates that the first half-second of that tremendous outburst came from an expanding fireball. In 1995 we suggested that part of the fireball was trapped by the magnetic field lines and held close to the star. This trapped fireball gradually shrank and then evaporated, emitting x-rays all the while. Based on the amount of energy released, we calculated the strength of the magnetic field needed to confine the enormous fireball pressure: greater than 10^{14} gauss, which agrees with the field strength inferred from the spin-down rate.

A separate estimate of the field had been given in 1992 by Bohdan Paczyński of Princeton. He noted that x-rays can slip

THE AUTHORS

CHRYSSA KOUVELIOTOU, ROBERT C. DUNCAN and *CHRISTOPHER THOMPSON* have studied magnetars for a collective 40 years and have collaborated for the past five. Kouveliotou, an observer, works at the National Space Science and Technology Center in Huntsville, Ala. Besides soft-gamma repeaters, her pets include gamma-ray bursts, x-ray binaries and her cat, Felix; her interests range from jazz to archaeology to linguistics. Duncan and Thompson are theorists, the former at the University of Texas at Austin, the latter at the Canadian Institute for Theoretical Astrophysics in Toronto. Duncan has studied supernovae, quark matter and intergalactic gas clouds. In his younger days he ran a 2:19 marathon in the 1980 U.S. Olympic trials. Thompson has worked on topics from cosmic strings to giant impacts in the early solar system. He, too, is an avid runner as well as a backpacker.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

through a cloud of electrons more easily if the charged particles are immersed in a very intense magnetic field. For the x-rays during the burst to have been so bright, the magnetic field must have been stronger than 10^{14} gauss.

What makes the theory so tricky is that the fields are stronger than the quantum electrodynamic threshold of 4×10^{13} gauss. In such strong fields, bizarre things happen. X-ray photons readily split in two or merge together. The vacuum itself is polarized, becoming strongly birefringent, like a calcite crystal. Atoms are deformed into long cylinders thinner than the quantum-relativistic wavelength of an electron [*see box on following page*]. All these strange phenomena have observable effects on magnetars. Because this physics was so exotic, the theory attracted few researchers at the time.

Zapped Again

AS THESE THEORETICAL developments were slowly unfolding, observers were still struggling to see the objects that were the sources of the bursts. The first opportunity came when NASA's orbiting Compton Gamma Ray Observatory recorded a burst of gamma rays late one evening in October 1993. This was the break Kouveliotou had been looking for when she joined the Compton team in Huntsville. The instrument that registered the burst could determine its position only to within a fairly broad swath of sky. Kouveliotou turned for help to the Japanese ASCA satellite. Toshio Murakami of the Institute of Space and Astronautical Science in Japan and his collaborators soon found an x-ray source from the same swath of sky. The source held steady, then gave off another burst—proving beyond all doubt that it was an SGR. The same object had first been seen in 1979 and, based on its approximate celestial coordinates, was identified as SGR 1806–20. Now its position was fixed much more precisely, and it could be monitored across the electromagnetic spectrum.

The next leap forward came in 1995, when NASA launched the Rossi X-ray Timing Explorer (RXTE), a satellite designed to be highly sensitive to variations in x-ray intensity. Using this instrument, Kouveliotou found that the emission from SGR 1806–20 was oscillating with a period of 7.47 seconds—amazingly close to the 8.0-second periodicity observed in the March 1979 burst (from SGR 0526–66). Over the course of five years, the SGR slowed by nearly two parts in 1,000. Although the slowdown may seem small, it is faster than that of any radio pulsar known, and it implies a magnetic field approaching 10^{15} gauss.

More thorough tests of the magnetar model would require a second giant flare. Luckily, the heavens soon complied. In the early morning of August 27, 1998, some 19 years after the giant flare that began SGR astronomy was observed, an even more intense wave of gamma rays and x-rays reached Earth from the depths of space. It drove detectors on seven scientific spacecraft to their maximum or off scale. One interplanetary probe, NASA's Comet Rendezvous Asteroid Flyby, was forced into a protective shutdown mode. The gamma rays hit Earth on its nightside, with the source in the zenith over the mid-Pacific Ocean.

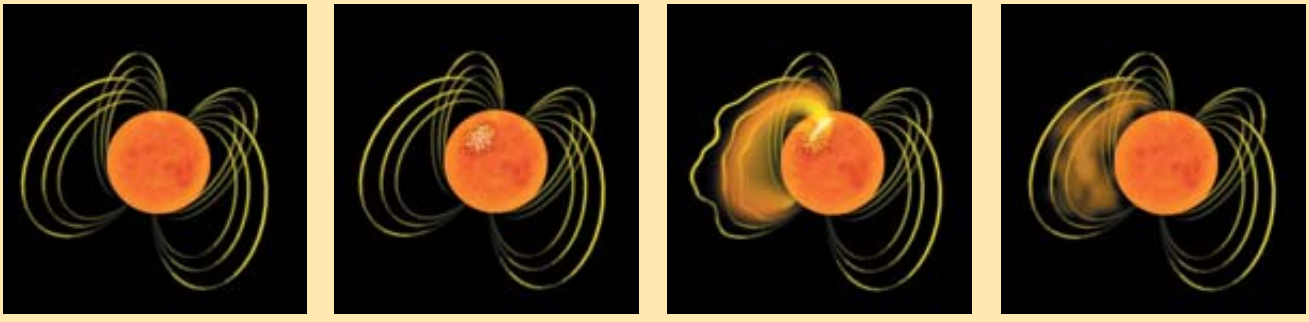
Fortuitously, in those early morning hours electrical engineer Umran S. Inan and his colleagues from Stanford University were gathering data on the propagation of very low frequency radio waves around Earth. At 3:22 A.M. PDT, they noticed an abrupt change in the ionized upper atmosphere. The inner edge of the ionosphere plunged down from 85 to 60 kilometers for five minutes. It was astonishing. This effect on our planet was caused by a neutron star far across the galaxy, 20,000 light-years away.

Another Magneto Marvel

THE AUGUST 27 FLARE was almost a carbon copy of the March 1979 event. Intrinsically, it was only one tenth as powerful, but because the source was closer to Earth it remains the most intense burst of gamma rays from beyond our solar system ever detected. The last few hundred seconds of the flare showed conspicuous pulsations, with a 5.16-second period. Kouveliotou

HOW MAGNETAR BURSTS HAPPEN

THE MAGNETIC FIELD OF THE STAR is so strong that the rigid crust sometimes breaks and crumbles, releasing a huge surge of energy.



1 Most of the time the magnetar is quiet. But magnetic stresses are slowly building up.

2 At some point the solid crust is stressed beyond its limit. It fractures, probably into many small pieces.

3 This "starquake" creates a surging electric current, which decays and leaves behind a hot fireball.

4 The fireball cools by releasing x-rays from its surface. It evaporates in minutes or less.

DON DIXON

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

and her team measured the spin-down rate of the star with RXTE; sure enough, it was slowing down at a rate comparable to that of SGR 1806–20, implying a similarly strong magnetic field. Another SGR was placed into the magnetar hall of fame.

The precise localizations of SGRs in x-rays have allowed them to be studied using radio and infrared telescopes (though not in visible light, which is blocked by interstellar dust). This work has been pioneered by many astronomers, notably Dale Frail of the National Radio Astronomy Observatory and Shri Kulkarni of the California Institute of Technology. Other observations have shown that all four confirmed SGRs continue to release energy, albeit faintly, even between outbursts. "Faintly" is a relative term: this x-ray glow represents 10 to 100 times as much power as the sun radiates in visible light.

By now one can say that magnetar magnetic fields are better measured than pulsar magnetic fields. In isolated pulsars, almost the only evidence for magnetic fields as strong as 10^{12} gauss comes from their measured spin-down. In contrast, the combination of rapid spin-down and bright x-ray flares provides several independent arguments for 10^{14}- to 10^{15}-gauss fields in magnetars. As this article goes to press, Alaa Ibrahim of the NASA Goddard Space Flight Center and his collaborators have reported yet another line of evidence for strong magnetic fields in magnetars: x-ray spectral lines that seem to be generated by protons gyrating in a 10^{15}-gauss field.

One intriguing question is whether magnetars are related to cosmic phenomena besides SGRs. The shortest-duration gamma-ray bursts, for example, have yet to be convincingly explained, and at least a handful of them could be flares from magnetars in other galaxies. If seen from a great distance, even a giant flare would be near the limit of telescope sensitivity. Only the brief, hard, intense pulse of gamma rays at the onset of the flare would be detected, so telescopes would register it as a GRB.

Thompson and Duncan suggested in the mid-1990s that magnetars might also explain anomalous x-ray pulsars, a class of objects that resemble SGRs in many ways. The one difficulty with this idea was that AXPs had not been observed to burst. Recently, however, Victoria M. Kaspi and Fotis P. Gavriil of McGill University and Peter M. Woods of the National Space and Technology Center in Huntsville detected bursts from two of the seven known AXPs. One of these objects is associated with a young supernova remnant in the constellation Cassiopeia.

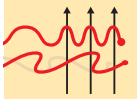
Another AXP in Cassiopeia is the first magnetar candidate to have been detected in visible light. Ferdi Hulleman and Marten van Kerkwijk of Utrecht University in the Netherlands, working with Kulkarni, spotted it three years ago, and Brian Kern and Christopher Martin of Caltech have since monitored its brightness in visible light. Though exceedingly faint, the AXP fades in and out with the x-ray period of the neutron star. These observations lend support to the idea that it is indeed a magnetar. The main alternative—that AXPs are ordinary neutron stars surrounded by disks of matter—predicts too much visible and infrared emission with too little pulsation.

In view of these recent discoveries, and the apparent silence of the Large Magellanic Cloud burster for nearly 20 years, it appears that magnetars can change their clothes. They can remain quiescent for years, even decades, before undergoing sudden periods of extreme activity. Some astronomers argue that AXPs are younger on average than SGRs, but this is still a matter of debate. If both SGRs and AXPs are magnetars, then magnetars plausibly constitute a substantial fraction of all neutron stars.

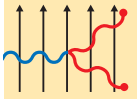
The story of magnetars is a sobering reminder of how much we have yet to understand about our universe. Thus far, we have discerned at most a dozen magnetars among the countless stars. They reveal themselves for a split second, in light that only the most sophisticated telescopes can detect. Within 10,000 years, their magnetic fields freeze and they stop emitting bright x-rays. So those dozen magnetars betray the presence of more than a million, and perhaps as many as 100 million, other objects—old magnetars that long ago went dark. Dim and dead, these strange worlds wander through interstellar space. What other phenomena, so rare and fleeting that we have not recognized them, lurk out there? SA

EXTREME MAGNETISM

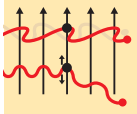
MAGNETAR FIELDS wreak havoc with radiation and matter.



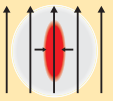
VACUUM BIREFRINGENCE
Polarized light waves (*orange*) change speed and hence wavelength when they enter a very strong magnetic field (*black lines*).



PHOTON SPLITTING
In a related effect, x-rays freely split in two or merge together. This process is important in fields stronger than 10^{14} gauss.



SCATTERING SUPPRESSION
A light wave can glide past an electron (*black circle*) with little hindrance if the field prevents the electron from vibrating with the wave.



DISTORTION OF ATOMS
Fields above 10^9 gauss squeeze electron orbitals into cigar shapes. In a 10^{14}-gauss field, a hydrogen atom becomes 200 times narrower.

MORE TO EXPLORE

Formation of Very Strongly Magnetized Neutron Stars: Implications for Gamma-Ray Bursts. Robert C. Duncan and Christopher Thompson in *Astronomical Journal,* Vol. 392, No. 1, pages L9–L13; June 10, 1992. Available at **makeashorterlink.com/?B16A425A2**

An X-ray Pulsar with a Superstrong Magnetic Field in the Soft Gamma-Ray Repeater SGR1806–20. C. Kouveliotou, S. Dieters, T. Strohmayer, J. Von Paradijs, G. J. Fishman, C. A. Meegan, K. Hurley, J. Kommers, I. Smith, D. Frail and T. Murakami in *Nature,* Vol. 393, pages 235–237; May 21, 1998.

The Life of a Neutron Star. Joshua N. Winn in *Sky & Telescope,* Vol. 98, No. 1, pages 30–38; July 1999.

Physics in Ultra-strong Magnetic Fields. Robert C. Duncan. Fifth Huntsville Gamma-Ray Burst Symposium, February 23, 2002. Available at **arXiv.org/abs/astro-ph/0002442**

Flash! The Hunt for the Biggest Explosions in the Universe. Govert Schilling. Cambridge University Press, 2002.

More information can be found at Robert C. Duncan's Web site: **solomon.as.utexas.edu/magnetar.html**

BRYAN CHRISTIE DESIGN

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

By Max Tegmark

Parallel Universes

Not just a staple of science fiction, other universes are a direct implication of cosmological observations

originally published in May 2003

Is there a copy of you reading this article? A person who is not you but who lives on a planet called Earth, with misty mountains, fertile fields and sprawling cities, in a solar system with eight other planets? The life of this person has been identical to yours in every respect. But perhaps he or she now decides to put down this article without finishing it, while you read on.

The idea of such an alter ego seems strange and implausible, but it looks as if we will just have to live with it, because it is supported by astronomical observations. The simplest and most popular cosmological model today predicts that you have a twin in a galaxy about 10 to the 10^{28} meters from here. This distance is so large that it is beyond astronomical, but that does not make your doppelgänger any less real. The estimate is derived from elementary probability and does not even assume speculative modern physics, merely that space is infinite (or at least sufficiently large) in size and almost uniformly filled with matter, as observations indicate. In infinite space, even the most unlikely events must take place somewhere. There are infinitely many other inhabited planets, including not just one but infinitely many that have people with the same appearance, name and memories as you, who play out every possible permutation of your life choices.

You will probably never see your other selves. The farthest you can observe is the distance that light has been able to travel during the 14 billion years since the big bang expansion began. The most distant visible objects are now about 4×10^{26} meters away—a distance that defines our observable universe, also called our Hubble volume, our horizon volume or simply our universe. Likewise, the universes of your other selves are spheres of the same size centered on their planets. They are the most straightforward example of parallel universes. Each universe is merely a small part of a larger "multiverse."

By this very definition of "universe," one might expect the notion of a multiverse to be forever in the domain of metaphysics. Yet the borderline between physics and metaphysics is defined by whether a theory is experimentally testable, not by whether it is weird or involves unobservable entities. The frontiers of physics have gradually expanded to incorporate ever more abstract (and once metaphysical) concepts such as a round Earth, invisible electromagnetic fields, time slowdown at high speeds, quantum superpositions, curved space, and black holes. Over the past several years the concept of a multiverse has joined this list. It is grounded in well-tested theories such as relativity and quantum mechanics, and it fulfills both of the basic criteria

ALFRED T. KAMAJIAN

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

of an empirical science: it makes predictions, and it can be falsified. Scientists have discussed as many as four distinct types of parallel universes. The key question is not whether the multiverse exists but rather how many levels it has.

Level I: Beyond Our Cosmic Horizon

THE PARALLEL UNIVERSES of your alter egos constitute the Level I multiverse. It is the least controversial type. We all accept the existence of things that we cannot see but could see if we moved to a different vantage point or merely waited, like people watching for ships to come over the horizon. Objects beyond the cosmic horizon have a similar status. The observable universe grows by a light-year every year as light from farther away has time to reach us. An infinity lies out there, waiting to be seen. You will probably die long before your alter egos come into view, but in principle, and if cosmic expansion cooperates, your descendants could observe them through a sufficiently powerful telescope.

If anything, the Level I multiverse sounds trivially obvious. How could space *not* be infinite? Is there a sign somewhere saying "Space Ends Here—Mind the Gap"? If so, what lies beyond it? In fact, Einstein's theory of gravity calls this intuition into question. Space could be finite if it has a convex curvature or an unusual topology (that is, interconnectedness). A spherical, doughnut-shaped or pretzel-shaped universe would have a limited volume and no edges. The cosmic microwave background radiation allows sensitive tests of such scenarios [see "Is Space Finite?" by Jean-Pierre Luminet, Glenn D. Starkman and Jeffrey R. Weeks; SCIENTIFIC AMERICAN, April 1999]. So far, however, the evidence is against them. Infinite models fit the data, and strong limits have been placed on the alternatives.

Another possibility is that space is infinite but matter is confined to a finite region around us—the historically popular "island universe" model. In a variant on this model, matter thins out on large scales in a fractal pattern. In both cases, almost all universes in the Level I multiverse would be empty and dead. But recent observations of the three-dimensional galaxy distribution and the microwave background have shown that the arrangement of matter gives way to dull uniformity on large scales, with no coherent structures larger than about 10^{24} meters. Assuming that this pattern continues, space beyond our observable universe teems with galaxies, stars and planets.

Observers living in Level I parallel universes experience the same laws of physics as we do but with different initial conditions. According to current theories, processes early in the big bang spread matter around with a degree of randomness, generating all possible arrangements with nonzero probability. Cosmologists assume that our universe, with an almost uniform distribution of matter and initial density fluctuations of one part in 100,000, is a fairly typical one (at least among those that contain observers). That assumption underlies the estimate that your closest identical copy is 10 to the 10^{28} meters away. About 10 to the 10^{92} meters away, there should be a sphere of radius 100 light-years identical to the one centered here, so all perceptions that we have during the next century will be identical to those of our counterparts over there. About 10 to the 10^{118} meters away should be an entire Hubble volume identical to ours.

These are extremely conservative estimates, derived simply by counting all possible quantum states that a Hubble volume can have if it is no hotter than 10^8 kelvins. One way to do the calculation is to ask how many protons could be packed into a Hubble volume at that temperature. The answer is 10^{118} protons. Each of those particles may or may not, in fact, be present, which makes for 2 to the 10^{118} possible arrangements of protons. A box containing that many Hubble volumes exhausts all the possibilities. If you round off the numbers, such a box is about 10 to the 10^{118} meters across. Beyond that box, universes—including ours—must repeat. Roughly the same number could be derived by using thermodynamic or quantum-gravitational estimates of the total information content of the universe.

Your nearest doppelgänger is most likely to be much closer than these numbers suggest, given the processes of planet formation and biological evolution that tip the odds in your favor. Astronomers suspect that our Hubble volume has at least 10^{20} habitable planets; some might well look like Earth.

The Level I multiverse framework is used routinely to evaluate theories in modern cosmology, although this procedure is rarely spelled out explicitly. For instance, consider how cosmologists used the microwave background to rule out a finite spherical geometry. Hot and cold spots in microwave background maps have a characteristic size that depends on the curvature of space, and the observed spots appear too small to be consistent with a spherical shape. But it is important to be statistically rigorous. The average spot size varies randomly from one Hubble volume to another, so it is possible that our universe is fooling us—it could be spherical but happen to have abnormally small spots. When cosmologists say they have ruled out the spherical model with 99.9 percent confidence, they really mean that if this model were true, fewer than one in 1,000 Hubble volumes would show spots as small as those we observe.

Overview/*Multiverses*

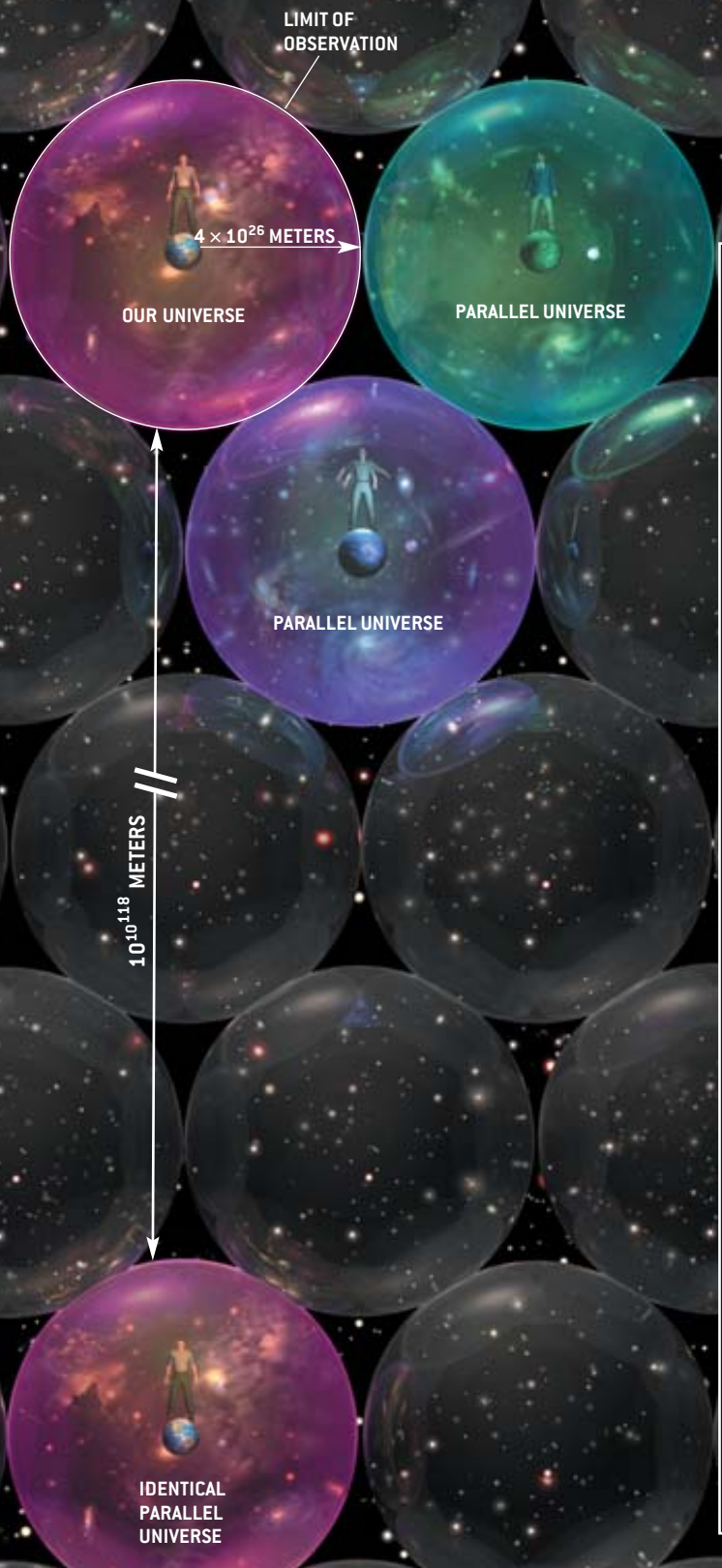
- One of the many implications of recent cosmological observations is that the concept of parallel universes is no mere metaphor. Space appears to be infinite in size. If so, then somewhere out there, everything that is possible becomes real, no matter how improbable it is. Beyond the range of our telescopes are other regions of space that are identical to ours. Those regions are a type of parallel universe. Scientists can even calculate how distant these universes are, on average.
- And that is fairly solid physics. When cosmologists consider theories that are less well established, they conclude that other universes can have entirely different properties and laws of physics. The presence of those universes would explain various strange aspects of our own. It could even answer fundamental questions about the nature of time and the comprehensibility of the physical world.

ALFRED T. KAMAJIAN (*background*); SARA CHEN (*inset*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

LEVEL I MULTIVERSE

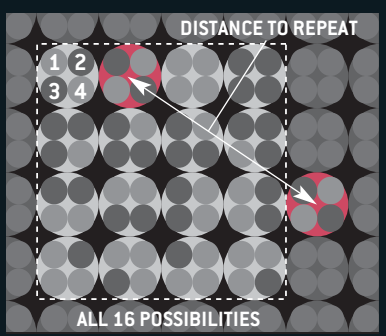
THE SIMPLEST TYPE of parallel universe is simply a region of space that is too far away for us to have seen yet. The farthest that we can observe is currently about 4×10^{26} meters, or 42 billion light-years—the distance that light has been able to travel since the big bang began. (The distance is greater than 14 billion light-years because cosmic expansion has lengthened distances.) Each of the Level I parallel universes is basically the same as ours. All the differences stem from variations in the initial arrangement of matter.



How Far Away Is a Duplicate Universe?

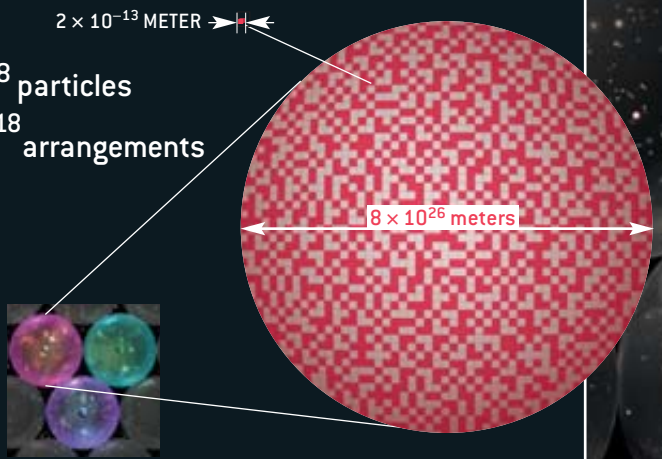
EXAMPLE UNIVERSE

Imagine a two-dimensional universe with space for four particles. Such a universe has 2^4, or 16, possible arrangements of matter. If more than 16 of these universes exist, they must begin to repeat. In this example, the distance to the nearest duplicate is roughly four times the diameter of each universe.

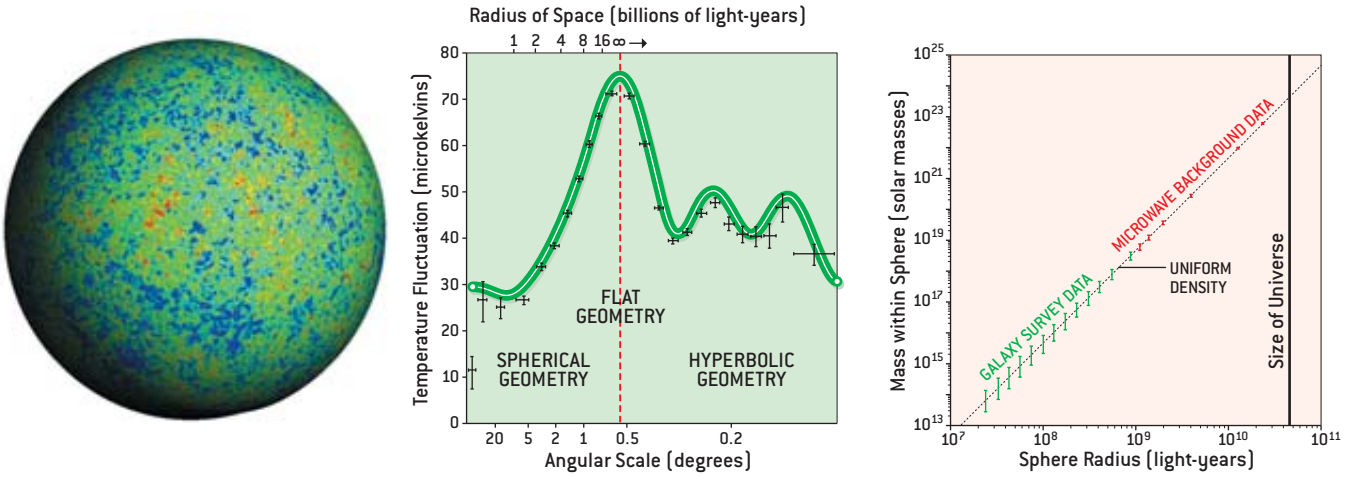


OUR UNIVERSE

The same argument applies to our universe, which has space for about 10^{118} subatomic particles. The number of possible arrangements is therefore 2 to the 10^{118}, or approximately 10 to the 10^{118}. Multiplying by the diameter of the universe gives an average distance to the nearest duplicate of 10 to the 10^{118} meters.



COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



COSMOLOGICAL DATA support the idea that space continues beyond the confines of our observable universe. The WMAP satellite recently measured the fluctuations in the microwave background (*left*). The strongest fluctuations are just over half a degree across, which indicates—after applying the rules of geometry—that space is very large or infinite (*center*). (One caveat: some cosmologists speculate that the discrepant point on the left of the graph is evidence for a finite volume.) In addition, WMAP and the 2dF Galaxy Redshift Survey have found that space on large scales is filled with matter uniformly (*right*), meaning that other universes should look basically like ours.

The lesson is that the multiverse theory can be tested and falsified even though we cannot see the other universes. The key is to predict what the ensemble of parallel universes is and to specify a probability distribution, or what mathematicians call a "measure," over that ensemble. Our universe should emerge as one of the most probable. If not—if, according to the multiverse theory, we live in an improbable universe—then the theory is in trouble. As I will discuss later, this measure problem can become quite challenging.

Level II: Other Postinflation Bubbles

IF THE LEVEL I MULTIVERSE was hard to stomach, try imagining an infinite set of distinct Level I multiverses, some perhaps with different spacetime dimensionality and different physical constants. Those other multiverses—which constitute a Level II multiverse—are predicted by the currently popular theory of chaotic eternal inflation.

Inflation is an extension of the big bang theory and ties up many of the loose ends of that theory, such as why the universe is so big, so uniform and so flat. A rapid stretching of space long ago can explain all these and other attributes in one fell swoop [see "The Inflationary Universe," by Alan H. Guth and Paul J. Steinhard; SCIENTIFIC AMERICAN, May 1984; and "The Self-Reproducing Inflationary Universe," by Andrei Linde, November 1994]. Such stretching is predicted by a wide class of theories of elementary particles, and all available evidence bears it out. The phrase "chaotic eternal" refers to what happens on the very largest scales. Space as a whole is stretching and will continue doing so forever, but some regions of space stop stretching and form distinct bubbles, like gas pockets in a loaf of rising bread. Infinitely many such bubbles emerge. Each is an embryonic Level I multiverse: infinite in size and filled with matter deposited by the energy field that drove inflation.

Those bubbles are more than infinitely far away from Earth, in the sense that you would never get there even if you traveled at the speed of light forever. The reason is that the space between our bubble and its neighbors is expanding faster than you could travel through it. Your descendants will never see their doppelgängers elsewhere in Level II. For the same reason, if cosmic expansion is accelerating, as observations now suggest, they might not see their alter egos even in Level I.

The Level II multiverse is far more diverse than the Level I multiverse. The bubbles vary not only in their initial conditions but also in seemingly immutable aspects of nature. The prevailing view in physics today is that the dimensionality of spacetime, the qualities of elementary particles and many of the so-called physical constants are not built into physical laws but are the outcome of processes known as symmetry breaking. For instance, theorists think that the space in our universe once had nine dimensions, all on an equal footing. Early in cosmic history, three of them partook in the cosmic expansion and became the three dimensions we now observe. The other six are now unobservable, either because they have stayed microscopic with a doughnutlike topology or because all matter is confined to a three-dimensional surface (a membrane, or simply "brane") in the nine-dimensional space.

Thus, the original symmetry among the dimensions broke. The quantum fluctuations that drive chaotic inflation could cause different symmetry breaking in different bubbles. Some might become four-dimensional, others could contain only two rather than three generations of quarks, and still others might have a stronger cosmological constant than our universe does.

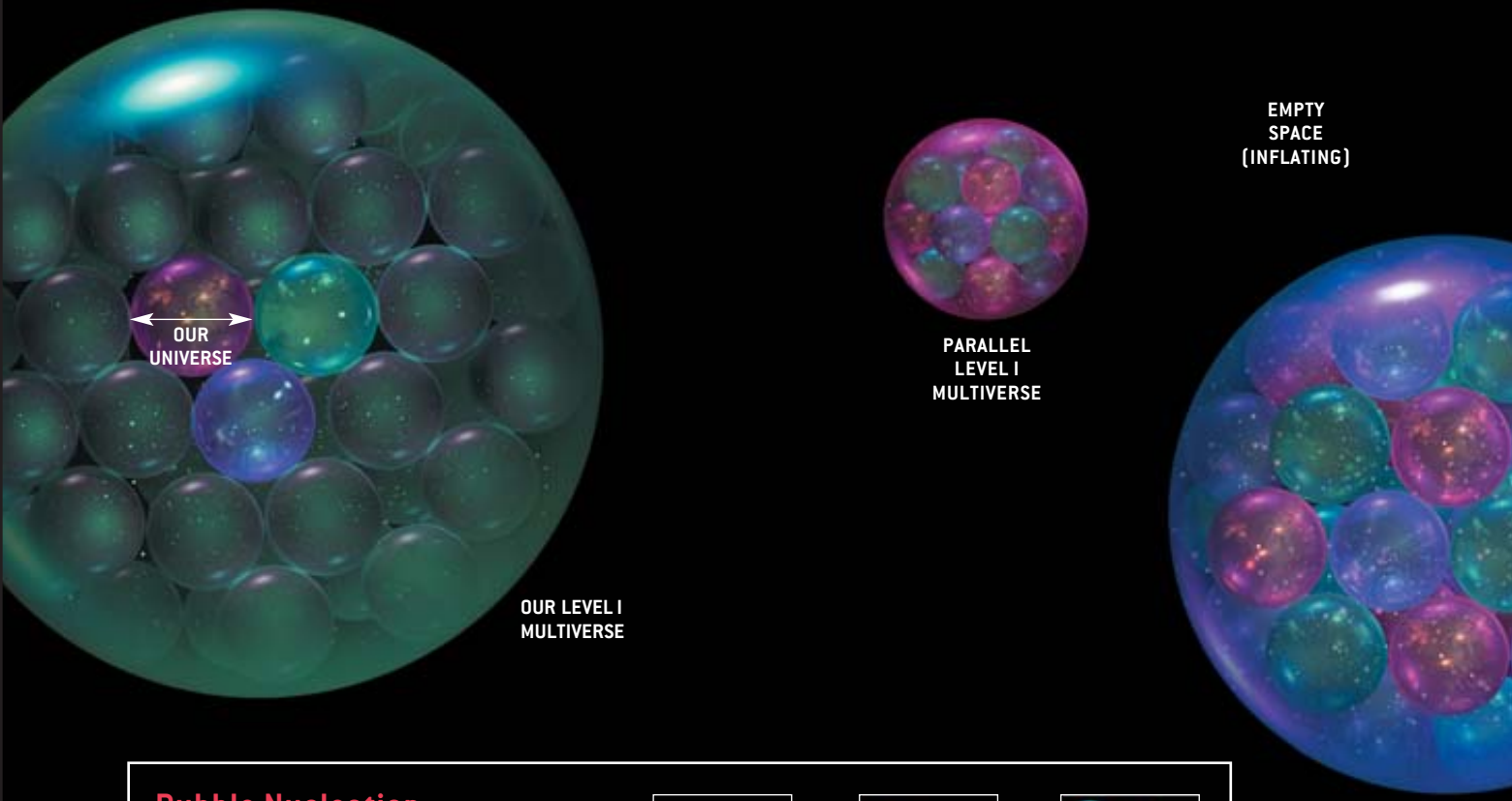
Another way to produce a Level II multiverse might be through a cycle of birth and destruction of universes. In a scientific context, this idea was introduced by physicist Richard C. Tolman in the 1930s and recently elaborated on by Paul J. Steinhardt of Princeton University and Neil Turok of the University of Cambridge. The Steinhardt and Turok proposal and related models involve a second three-dimensional brane that is quite literally parallel to ours, merely offset in a higher dimension [see "Been There, Done That," by George Musser; News Scan, SCIENTIFIC AMERICAN, March 2002]. This parallel universe is not

MAX TEGMARK (*sphere*); SARA CHEN (*graphs*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

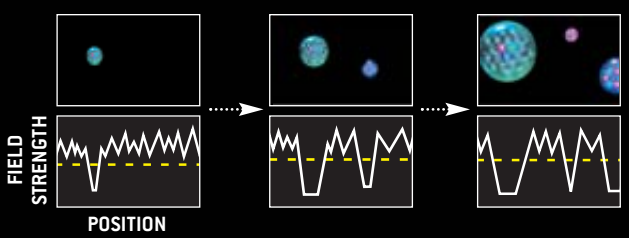
LEVEL II MULTIVERSE

A SOMEWHAT MORE ELABORATE type of parallel universe emerges from the theory of cosmological inflation. The idea is that our Level I multiverse—namely, our universe and contiguous regions of space—is a bubble embedded in an even vaster but mostly empty volume. Other bubbles exist out there, disconnected from ours. They nucleate like raindrops in a cloud. During nucleation, variations in quantum fields endow each bubble with properties that distinguish it from other bubbles.



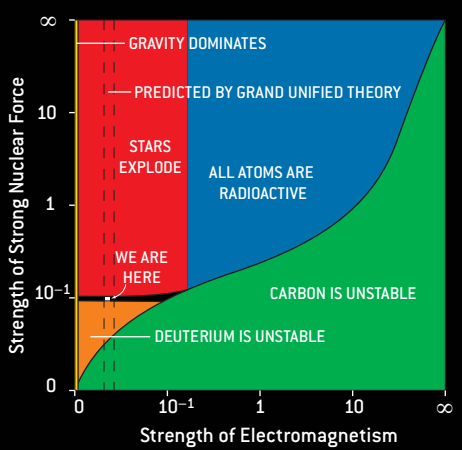
Bubble Nucleation

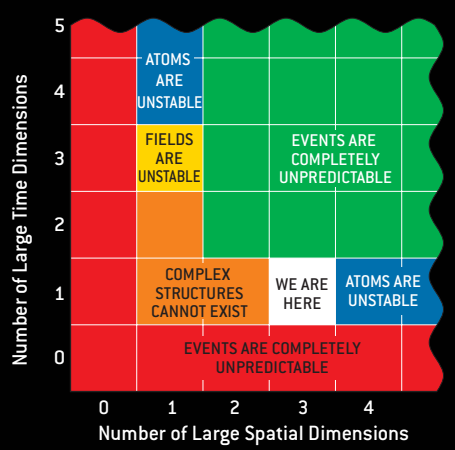
A QUANTUM FIELD known as the inflaton causes space to expand rapidly. In the bulk of space, random fluctuations prevent the field from decaying away. But in certain regions, the field loses its strength and the expansion slows down. Those regions become bubbles.



Evidence

COSMOLOGISTS INFER the presence of Level II parallel universes by scrutinizing the properties of our universe. These properties, including the strength of the forces of nature (*right*) and the number of observable space and time dimensions (*far right*), were established by random processes during the birth of our universe. Yet they have exactly the values that sustain life. That suggests the existence of other universes with other values.





ALFRED T. KAMAJIAN (*background*); CORNELIA BLIK (*top inset*); SARA CHEN (*bottom inset*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

really a separate universe, because it interacts with ours. But the ensemble of universes—past, present and future—that these branes create would form a multiverse, arguably with a diversity similar to that produced by chaotic inflation. An idea proposed by physicist Lee Smolin of the Perimeter Institute in Waterloo, Ontario, involves yet another multiverse comparable in diversity to that of Level II but mutating and sprouting new universes through black holes rather than through brane physics.

Although we cannot interact with other Level II parallel universes, cosmologists can infer their presence indirectly, because their existence can account for unexplained coincidences in our universe. To give an analogy, suppose you check into a hotel, are assigned room 1967 and note that this is the year you were born. What a coincidence, you say. After a moment of reflection, however, you conclude that this is not so surprising after all. The hotel has hundreds of rooms, and you would not have been having these thoughts in the first place if you had been assigned one with a number that meant nothing to you. The lesson is that even if you knew nothing about hotels, you could infer the existence of other hotel rooms to explain the coincidence.

As a more pertinent example, consider the mass of the sun. The mass of a star determines its luminosity, and using basic physics, one can compute that life as we know it on Earth is possible only if the sun's mass falls into the narrow range between 1.6×10^{30} and 2.4×10^{30} kilograms. Otherwise Earth's climate would be colder than that of present-day Mars or hotter than that of present-day Venus. The measured solar mass is 2.0×10^{30} kilograms. At first glance, this apparent coincidence of the habitable and observed mass values appears to be a wild stroke of luck. Stellar masses run from 10^{29} to 10^{32} kilograms, so if the sun acquired its mass at random, it had only a small chance of falling into the habitable range. But just as in the hotel example, one can explain this apparent coincidence by postulating an ensemble (in this case, a number of planetary systems) and a selection effect (the fact that we must find ourselves living on a habitable planet). Such observer-related selection effects are referred to as "anthropic," and although the "A-word" is notorious for triggering controversy, physicists broadly agree that these selection effects cannot be neglected when testing fundamental theories.

What applies to hotel rooms and planetary systems applies to parallel universes. Most, if not all, of the attributes set by symmetry breaking appear to be fine-tuned. Changing their values by modest amounts would have resulted in a qualitatively different universe—one in which we probably would not exist. If protons were 0.2 percent heavier, they could decay into neutrons, destabilizing atoms. If the electromagnetic force were 4 percent weaker, there would be no hydrogen and no normal stars. If the weak interaction were much weaker, hydrogen would not exist; if it were much stronger, supernovae would fail to seed interstellar space with heavy elements. If the cosmological constant were much larger, the universe would have blown itself apart before galaxies could form.

Although the degree of fine-tuning is still debated, these examples suggest the existence of parallel universes with other values of the physical constants [see "Exploring Our Universe and Others," by Martin Rees; Scientific American, December 1999]. The Level II multiverse theory predicts that physicists will never be able to determine the values of these constants from first principles. They will merely compute probability distributions for what they should expect to find, taking selection effects into account. The result should be as generic as is consistent with our existence.

Level III: Quantum Many Worlds

THE LEVEL I AND LEVEL II multiverses involve parallel worlds that are far away, beyond the domain even of astronomers. But the next level of multiverse is right around you. It arises from the famous, and famously controversial, many-worlds interpretation of quantum mechanics—the idea that random quantum processes cause the universe to branch into multiple copies, one for each possible outcome.

In the early 20th century the theory of quantum mechanics revolutionized physics by explaining the atomic realm, which does not abide by the classical rules of Newtonian mechanics. Despite the obvious successes of the theory, a heated debate rages about what it really means. The theory specifies the state of the universe not in classical terms, such as the positions and velocities of all particles, but in terms of a mathematical object called a wave function. According to the Schrödinger equation, this state evolves over time in a fashion that mathematicians term "unitary," meaning that the wave function rotates in an abstract infinite-dimensional space called Hilbert space. Although quantum mechanics is often described as inherently random and uncertain, the wave function evolves in a deterministic way. There is nothing random or uncertain about it.

The sticky part is how to connect this wave function with what we observe. Many legitimate wave functions correspond to counterintuitive situations, such as a cat being dead and alive at the same time in a so-called superposition. In the 1920s physicists explained away this weirdness by postulating that the wave function "collapsed" into some definite classical outcome whenever someone made an observation. This add-on had the virtue of explaining observations, but it turned an elegant, unitary theory into a kludgy, nonunitary one. The intrinsic randomness commonly ascribed to quantum mechanics is the result of this postulate.

Over the years many physicists have abandoned this view in favor of one developed in 1957 by Princeton graduate student Hugh Everett III. He showed that the collapse postulate is unnecessary. Unadulterated quantum theory does not, in fact, pose any contradictions. Although it predicts that one classical reality gradually splits into superpositions of many such realities, observers subjectively experience this splitting merely as a slight randomness, with probabilities in exact agreement with those from the old collapse postulate. This superposition of classical worlds is the Level III multiverse.

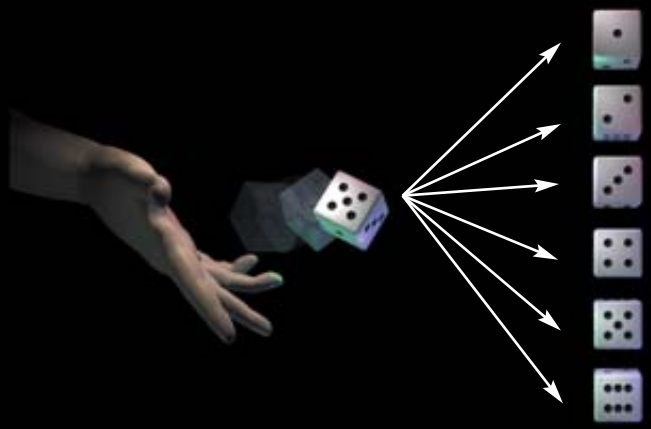
Everett's many-worlds interpretation has been boggling minds inside and outside physics for more than four decades. But the theory becomes easier to grasp when one distinguishes

ALFRED T. KAMAJIAN

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

LEVEL III MULTIVERSE

QUANTUM MECHANICS PREDICTS a vast number of parallel universes by broadening the concept of "elsewhere." These universes are located elsewhere, not in ordinary space but in an abstract realm of all possible states. Every conceivable way that the world could be (within the scope of quantum mechanics) corresponds to a different universe. The parallel universes make their presence felt in laboratory experiments, such as wave interference and quantum computation.

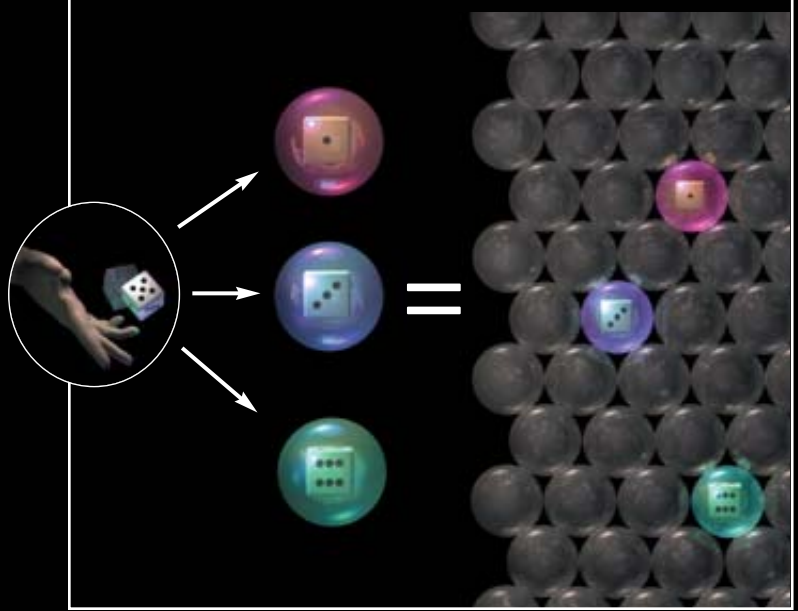


Quantum Dice

IMAGINE AN IDEAL DIE whose randomness is purely quantum. When you roll it, the die appears to land on a certain value at random. Quantum mechanics, however, predicts that it lands on all values at once. One way to reconcile these contradictory views is to conclude that the die lands on different values in different universes. In one sixth of the universes, it lands on 1; in one sixth, on 2, and so on. Trapped within one universe, we can perceive only a fraction of the full quantum reality.

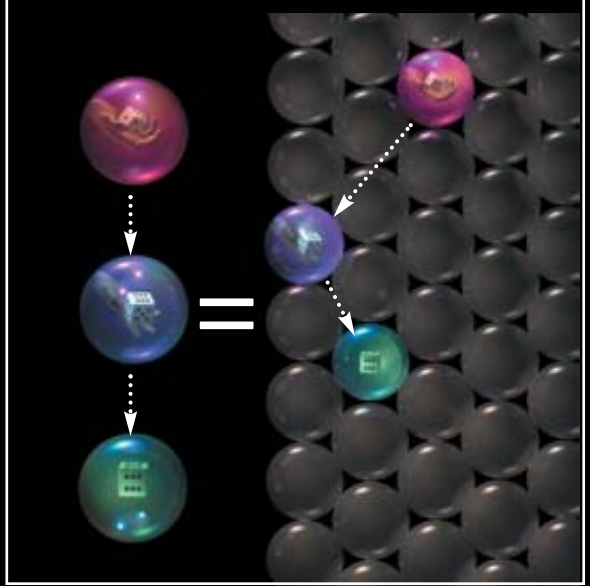
Ergodicity

ACCORDING TO THE PRINCIPLE of ergodicity, quantum parallel universes are equivalent to more prosaic types of parallel universes. A quantum universe splits over time into multiple universes (*left*). Yet those new universes are no different from parallel universes that already exist somewhere else in space—in, for example, other Level I universes (*right*). The key idea is that parallel universes, of whatever type, embody different ways that events could have unfolded.



The Nature of Time

MOST PEOPLE THINK of time as a way to describe change. At one moment, matter has a certain arrangement; a moment later, it has another (*left*). The concept of multiverses suggests an alternative view. If parallel universes contain all possible arrangements of matter (*right*), then time is simply a way to put those universes into a sequence. The universes themselves are static; change is an illusion, albeit an interesting one.



COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

between two ways of viewing a physical theory: the outside view of a physicist studying its mathematical equations, like a bird surveying a landscape from high above it, and the inside view of an observer living in the world described by the equations, like a frog living in the landscape surveyed by the bird.

From the bird perspective, the Level III multiverse is simple. There is only one wave function. It evolves smoothly and deterministically over time without any kind of splitting or parallelism. The abstract quantum world described by this evolving wave function contains within it a vast number of parallel classical story lines, continuously splitting and merging, as well as a number of quantum phenomena that lack a classical description. From their frog perspective, observers perceive only a tiny fraction of this full reality. They can view their own Level I universe, but a process called decoherence—which mimics wave function collapse while preserving unitarity—prevents them from seeing Level III parallel copies of themselves.

Whenever observers are asked a question, make a snap decision and give an answer, quantum effects in their brains lead to a superposition of outcomes, such as "Continue reading the article" and "Put down the article." From the bird perspective, the act of making a decision causes a person to split into multiple copies: one who keeps on reading and one who doesn't. From their frog perspective, however, each of these alter egos is unaware of the others and notices the branching merely as a slight randomness: a certain probability of continuing to read or not.

As strange as this may sound, the exact same situation occurs even in the Level I multiverse. You have evidently decided to keep on reading the article, but one of your alter egos in a distant galaxy put down the magazine after the first paragraph. The only difference between Level I and Level III is where your doppelgängers reside. In Level I they live elsewhere in good old three-dimensional space. In Level III they live on another quantum branch in infinite-dimensional Hilbert space.

The existence of Level III depends on one crucial assumption: that the time evolution of the wave function is unitary. So far experimenters have encountered no departures from unitarity. In the past few decades they have confirmed unitarity for ever larger systems, including carbon 60 buckyball molecules and kilometer-long optical fibers. On the theoretical side, the case for unitarity has been bolstered by the discovery of decoherence [see "100 Years of Quantum Mysteries," by Max Tegmark and John Archibald Wheeler; Scientific American, February 2001]. Some theorists who work on quantum gravity have questioned unitarity; one concern is that evaporating black holes might destroy information, which would be a nonunitary process. But a recent breakthrough in string theory known as AdS/CFT correspondence suggests that even quantum gravity is unitary. If so, black holes do not destroy information but merely transmit it elsewhere. [*Editors' note: An upcoming article will discuss this correspondence in greater detail.*]

If physics is unitary, then the standard picture of how quantum fluctuations operated early in the big bang must change. These fluctuations did not generate initial conditions at random. Rather they generated a quantum superposition of all possible initial conditions, which coexisted simultaneously. Decoherence then caused these initial conditions to behave classically in separate quantum branches. Here is the crucial point: the distribution of outcomes on different quantum branches in a given Hubble volume (Level III) is identical to the distribution of outcomes in different Hubble volumes within a single quantum branch (Level I). This property of the quantum fluctuations is known in statistical mechanics as ergodicity.

The same reasoning applies to Level II. The process of symmetry breaking did not produce a unique outcome but rather a superposition of all outcomes, which rapidly went their separate ways. So if physical constants, spacetime dimensionality and so on can vary among parallel quantum branches at Level III, then they will also vary among parallel universes at Level II.

In other words, the Level III multiverse adds nothing new beyond Level I and Level II, just more indistinguishable copies of the same universes—the same old story lines playing out again and again in other quantum branches. The passionate debate about Everett's theory therefore seems to be ending in a grand anticlimax, with the discovery of less controversial multiverses (Levels I and II) that are equally large.

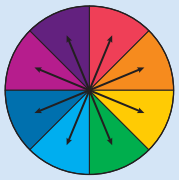
Needless to say, the implications are profound, and physicists are only beginning to explore them. For instance, consider the ramifications of the answer to a long-standing question: Does the number of universes exponentially increase over time? The surprising answer is no. From the bird perspective, there is of course only one quantum universe. From the frog perspective, what matters is the number of universes that are distinguishable at a given instant—that is, the number of noticeably different Hubble volumes. Imagine moving planets to random new locations, imagine having married someone else, and so on. At the quantum level, there are 10 to the 10^{118} universes with temperatures below 10^8 kelvins. That is a vast number, but a finite one.

From the frog perspective, the evolution of the wave function corresponds to a never-ending sliding from one of these 10 to the 10^{118} states to another. Now you are in universe A, the one in which you are reading this sentence. Now you are in universe B, the one in which you are reading this other sentence. Put differently, universe B has an observer identical to one in universe A, except with an extra instant of memories. All possible states exist at every instant, so the passage of time may be in the eye of the beholder—an idea explored in Greg Egan's

THE AUTHOR

MAX TEGMARK wrote a four-dimensional version of the computer game Tetris while in college. In another universe, he went on to become a highly paid software developer. In our universe, however, he wound up as professor of physics and astronomy at the University of Pennsylvania. Tegmark is an expert in analyzing the cosmic microwave background and galaxy clustering. Much of his work bears on the concept of parallel universes: evaluating evidence for infinite space and cosmological inflation; developing insights into quantum decoherence; and studying the possibility that the amplitude of microwave background fluctuations, the dimensionality of spacetime and the fundamental laws of physics can vary from place to place.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



The Mystery of Probability:
What Are the Odds?

AS MULTIVERSE THEORIES gain credence, the sticky issue of how to compute probabilities in physics is growing from a minor nuisance into a major embarrassment. If there are indeed many identical copies of you, the traditional notion of determinism evaporates. You could not compute your own future even if you had complete knowledge of the entire state of the multiverse, because there is no way for you to determine which of these copies is you (they all feel they are). All you can predict, therefore, are probabilities for what you would observe. If an outcome has a probability of, say, 50 percent, it means that half the observers observe that outcome.

Unfortunately, it is not an easy task to compute what fraction of the infinitely many observers perceive what. The answer depends on the order in which you count them. By analogy, the fraction of the integers that are even is 50 percent if you order them numerically (1, 2, 3, 4, ...) but approaches 100 percent if you sort them digit by digit, the way your word processor would (1, 10, 100, 1,000, ...). When observers reside in disconnected universes, there is no obviously natural way in which to order them. Instead one must sample from the different universes with some statistical weights referred to by mathematicians as a "measure."

This problem crops up in a mild and treatable manner at Level I, becomes severe at Level II, has caused much debate at Level III, and is horrendous at Level IV. At Level II, for instance, Alexander Vilenkin of Tufts University and others have published predictions for the probability distributions of various cosmological parameters. They have argued that different parallel universes that have inflated by different amounts should be given statistical weights proportional to their volume. On the other hand, any mathematician will tell you that $2 \times \infty = \infty$, so there is no objective sense in which an infinite universe that has expanded by a factor of two has gotten larger. Moreover, a finite universe with the topology of a torus is equivalent to a perfectly periodic universe with infinite volume, both from the mathematical bird perspective and from the frog perspective of an observer within it. So why should its infinitely smaller volume give it zero statistical weight? After all, even in the Level I multiverse, Hubble volumes start repeating (albeit in a random order, not periodically) after about 10 to the 10^{118} meters.

If you think that is bad, consider the problem of assigning statistical weights to different mathematical structures at Level IV. The fact that our universe seems relatively simple has led many people to suggest that the correct measure somehow involves complexity. —*M.T.*

1994 science-fiction novel *Permutation City* and developed by physicist David Deutsch of the University of Oxford, independent physicist Julian Barbour, and others. The multiverse framework may thus prove essential to understanding the nature of time.

Level IV: Other Mathematical Structures

THE INITIAL CONDITIONS and physical constants in the Level I, Level II and Level III multiverses can vary, but the fundamental laws that govern nature remain the same. Why stop there? Why not allow the laws themselves to vary? How about a universe that obeys the laws of classical physics, with no quantum effects? How about time that comes in discrete steps, as for computers, instead of being continuous? How about a universe that is simply an empty dodecahedron? In the Level IV multiverse, all these alternative realities actually exist.

A hint that such a multiverse might not be just some beer-fueled speculation is the tight correspondence between the worlds of abstract reasoning and of observed reality. Equations and, more generally, mathematical structures such as numbers, vectors and geometric objects describe the world with remarkable verisimilitude. In a famous 1959 lecture, physicist Eugene P. Wigner argued that "the enormous usefulness of mathematics in the natural sciences is something bordering on the mysterious." Conversely, mathematical structures have an eerily real feel to them. They satisfy a central criterion of objective existence: they are the same no matter who studies them. A theorem is true regardless of whether it is proved by a human, a computer or an intelligent dolphin. Contemplative alien civilizations would find the same mathematical structures as we have. Accordingly, mathematicians commonly say that they discover mathematical structures rather than create them.

There are two tenable but diametrically opposed paradigms for understanding the correspondence between mathematics and physics, a dichotomy that arguably goes as far back as Plato and Aristotle. According to the Aristotelian paradigm, physical reality is fundamental and mathematical language is merely a useful approximation. According to the Platonic paradigm, the mathematical structure is the true reality and observers perceive it imperfectly. In other words, the two paradigms disagree on which is more basic, the frog perspective of the observer or the bird perspective of the physical laws. The Aristotelian paradigm prefers the frog perspective, whereas the Platonic paradigm prefers the bird perspective.

As children, long before we had even heard of mathematics, we were all indoctrinated with the Aristotelian paradigm. The Platonic view is an acquired taste. Modern theoretical physicists tend to be Platonists, suspecting that mathematics describes the universe so well because the universe is inherently mathematical. Then all of physics is ultimately a mathematics problem: a mathematician with unlimited intelligence and resources could in principle compute the frog perspective—that is, compute what self-aware observers the universe contains, what they perceive, and what languages they invent to describe their perceptions to one another.

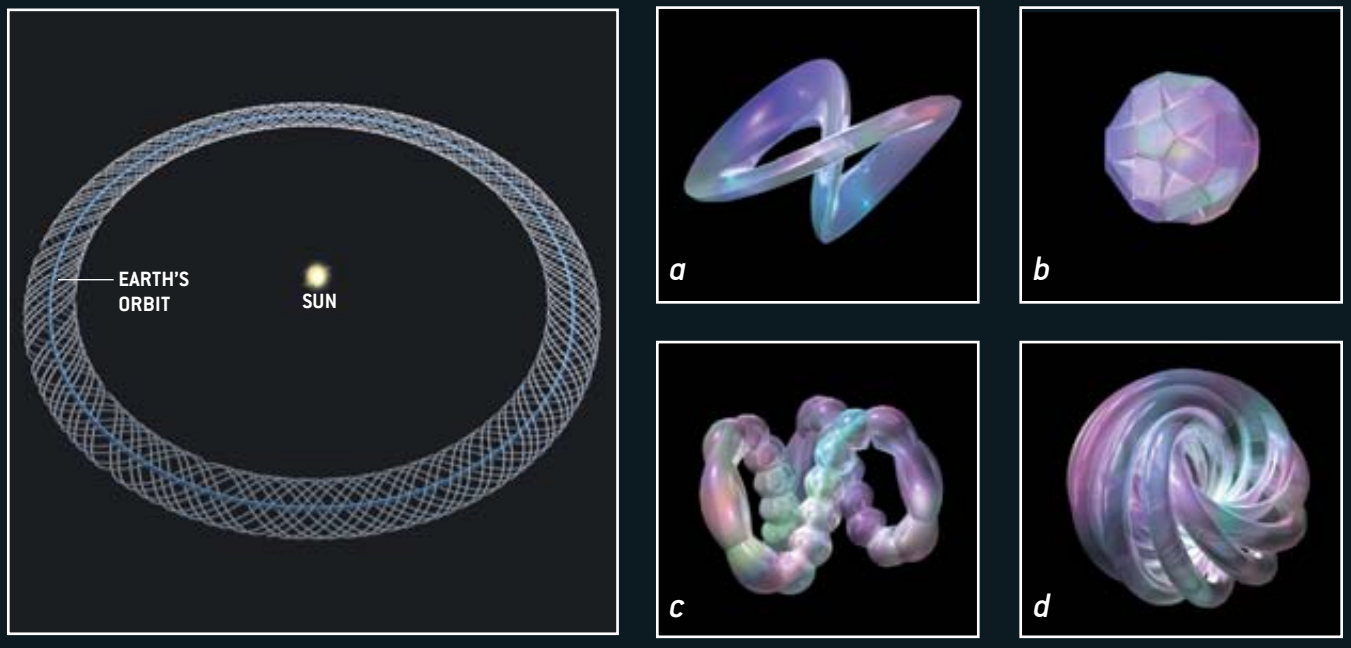
A mathematical structure is an abstract, immutable entity existing outside of space and time. If history were a movie, the structure would correspond not to a single frame of it but to the entire videotape. Consider, for example, a world made up of pointlike particles moving around in three-dimensional space.

SARA CHEN

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

LEVEL IV MULTIVERSE

THE ULTIMATE TYPE of parallel universe opens up the full realm of possibility. Universes can differ not just in location, cosmological properties or quantum state but also in the laws of physics. Existing outside of space and time, they are almost impossible to visualize; the best one can do is to think of them abstractly, as static sculptures that represent the mathematical structure of the physical laws that govern them. For example, consider a simple universe: Earth, moon and sun, obeying Newton's laws. To an objective observer, this universe looks like a circular ring (Earth's orbit smeared out in time) wrapped in a braid (the moon's orbit around Earth). Other shapes embody other laws of physics (*a, b, c, d*). This paradigm solves various problems concerning the foundations of physics.



In four-dimensional spacetime—the bird perspective—these particle trajectories resemble a tangle of spaghetti. If the frog sees a particle moving with constant velocity, the bird sees a straight strand of uncooked spaghetti. If the frog sees a pair of orbiting particles, the bird sees two spaghetti strands intertwined like a double helix. To the frog, the world is described by Newton's laws of motion and gravitation. To the bird, it is described by the geometry of the pasta—a mathematical structure. The frog itself is merely a thick bundle of pasta, whose highly complex intertwining corresponds to a cluster of particles that store and process information. Our universe is far more complicated than this example, and scientists do not yet know to what, if any, mathematical structure it corresponds.

The Platonic paradigm raises the question of why the universe is the way it is. To an Aristotelian, this is a meaningless question: the universe just is. But a Platonist cannot help but wonder why it could not have been different. If the universe is inherently mathematical, then why was only one of the many mathematical structures singled out to describe a universe? A fundamental asymmetry appears to be built into the very heart of reality.

As a way out of this conundrum, I have suggested that complete mathematical symmetry holds: that all mathematical structures exist physically as well. Every mathematical structure corresponds to a parallel universe. The elements of this multiverse do not reside in the same space but exist outside of space and time. Most of them are probably devoid of observers. This hypothesis can be viewed as a form of radical Platonism, asserting that the mathematical structures in Plato's realm of ideas or the "mindscape" of mathematician Rudy Rucker of San Jose State University exist in a physical sense. It is akin to what cosmologist John D. Barrow of the University of Cambridge refers to as "π in the sky," what the late Harvard University philosopher Robert Nozick called the principle of fecundity and what the late Princeton philosopher David K. Lewis called modal realism. Level IV brings closure to the hierarchy of multiverses, because any self-consistent fundamental physical theory can be phrased as some kind of mathematical structure.

The Level IV multiverse hypothesis makes testable predictions. As with Level II, it involves an ensemble (in this case, the full range of mathematical structures) and selection effects. As mathematicians continue to categorize mathematical structures, they should find that the structure describing our world is the most generic one consistent with our observations. Similarly, our future observations should be the most generic ones that are consistent with our past observations, and our past observations should be the most generic ones that are consistent with our existence.

Quantifying what "generic" means is a severe problem, and this investigation is only now beginning. But one striking and

BRYAN CHRISTIE DESIGN (*left*); ALFRED T. KAMAJIAN (*a–d*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

encouraging feature of mathematical structures is that the symmetry and invariance properties that are responsible for the simplicity and orderliness of our universe tend to be generic, more the rule than the exception. Mathematical structures tend to have them by default, and complicated additional axioms must be added to make them go away.

What Says Occam?

THE SCIENTIFIC THEORIES of parallel universes, therefore, form a four-level hierarchy, in which universes become progressively more different from ours. They might have different initial conditions (Level I); different physical constants and particles (Level II); or different physical laws (Level IV). It is ironic that Level III is the one that has drawn the most fire in the past decades, because it is the only one that adds no qualitatively new types of universes.

In the coming decade, dramatically improved cosmological measurements of the microwave background and the large-scale matter distribution will support or refute Level I by further pinning down the curvature and topology of space. These measurements will also probe Level II by testing the theory of chaotic eternal inflation. Progress in both astrophysics and high-energy physics should also clarify the extent to which physical constants are fine-tuned, thereby weakening or strengthening the case for Level II.

If current efforts to build quantum computers succeed, they will provide further evidence for Level III, as they would, in essence, be exploiting the parallelism of the Level III multiverse for parallel computation. Experimenters are also looking for evidence of unitarity violation, which would rule out Level III. Finally, success or failure in the grand challenge of modern physics—unifying general relativity and quantum field theory—will sway opinions on Level IV. Either we will find a mathematical structure that exactly matches our universe, or we will bump up against a limit to the unreasonable effectiveness of mathematics and have to abandon that level.

So should you believe in parallel universes? The principal arguments against them are that they are wasteful and that they are weird. The first argument is that multiverse theories are vulnerable to Occam's razor because they postulate the existence of other worlds that we can never observe. Why should nature be so wasteful and indulge in such opulence as an infinity of different worlds? Yet this argument can be turned around to argue *for* a multiverse. What precisely would nature be wasting? Certainly not space, mass or atoms—the uncontroversial Level I multiverse already contains an infinite amount of all three, so who cares if nature wastes some more? The real issue here is the apparent reduction in simplicity. A skeptic worries about all the information necessary to specify all those unseen worlds.

But an entire ensemble is often much simpler than one of its members. This principle can be stated more formally using the notion of algorithmic information content. The algorithmic information content in a number is, roughly speaking, the length of the shortest computer program that will produce that number as output. For example, consider the set of all integers. Which is simpler, the whole set or just one number? Naively, you might think that a single number is simpler, but the entire set can be generated by quite a trivial computer program, whereas a single number can be hugely long. Therefore, the whole set is actually simpler.

Similarly, the set of all solutions to Einstein's field equations is simpler than a specific solution. The former is described by a few equations, whereas the latter requires the specification of vast amounts of initial data on some hypersurface. The lesson is that complexity increases when we restrict our attention to one particular element in an ensemble, thereby losing the symmetry and simplicity that were inherent in the totality of all the elements taken together.

In this sense, the higher-level multiverses are simpler. Going from our universe to the Level I multiverse eliminates the need to specify initial conditions, upgrading to Level II eliminates the need to specify physical constants, and the Level IV multiverse eliminates the need to specify anything at all. The opulence of complexity is all in the subjective perceptions of observers—the frog perspective. From the bird perspective, the multiverse could hardly be any simpler.

The complaint about weirdness is aesthetic rather than scientific, and it really makes sense only in the Aristotelian worldview. Yet what did we expect? When we ask a profound question about the nature of reality, do we not expect an answer that sounds strange? Evolution provided us with intuition for the everyday physics that had survival value for our distant ancestors, so whenever we venture beyond the everyday world, we should expect it to seem bizarre.

A common feature of all four multiverse levels is that the simplest and arguably most elegant theory involves parallel universes by default. To deny the existence of those universes, one needs to complicate the theory by adding experimentally unsupported processes and ad hoc postulates: finite space, wave function collapse and ontological asymmetry. Our judgment therefore comes down to which we find more wasteful and inelegant: many worlds or many words. Perhaps we will gradually get used to the weird ways of our cosmos and find its strangeness to be part of its charm. SA

MORE TO EXPLORE

Why Is the CMB Fluctuation Level 10^{-5}? Max Tegmark and Martin Rees in *Astrophysical Journal,* Vol. 499, No. 2, pages 526–532; June 1, 1998. Available online at **arXiv.org/abs/astro-ph/9709058**

Is "The Theory of Everything" Merely the Ultimate Ensemble Theory? Max Tegmark in *Annals of Physics,* Vol. 270, No.1, pages 1–51; November 20, 1998. Available online at **arXiv.org/abs/gr-qc/9704009**

Many Worlds in One. Jaume Garriga and Alexander Vilenkin in *Physical Review,* Vol. D64, No. 043511; July 26, 2001. Available online at **arXiv.org/abs/gr-qc/0102010**

Our Cosmic Habitat. Martin Rees. Princeton University Press, 2001.

Inflation, Quantum Cosmology and the Anthropic Principle. Andrei Linde in *Science and Ultimate Reality: From Quantum to Cosmos*. Edited by J. D. Barrow, P.C.W. Davies and C. L. Harper. Cambridge University Press, 2003. Available online at **arXiv.org/abs/hep-th/0211048**

The author's Web site has more information at **www.hep.upenn.edu/~max/multiverse.html**

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

Information in the HOLOGRAPHIC UNIVERSE

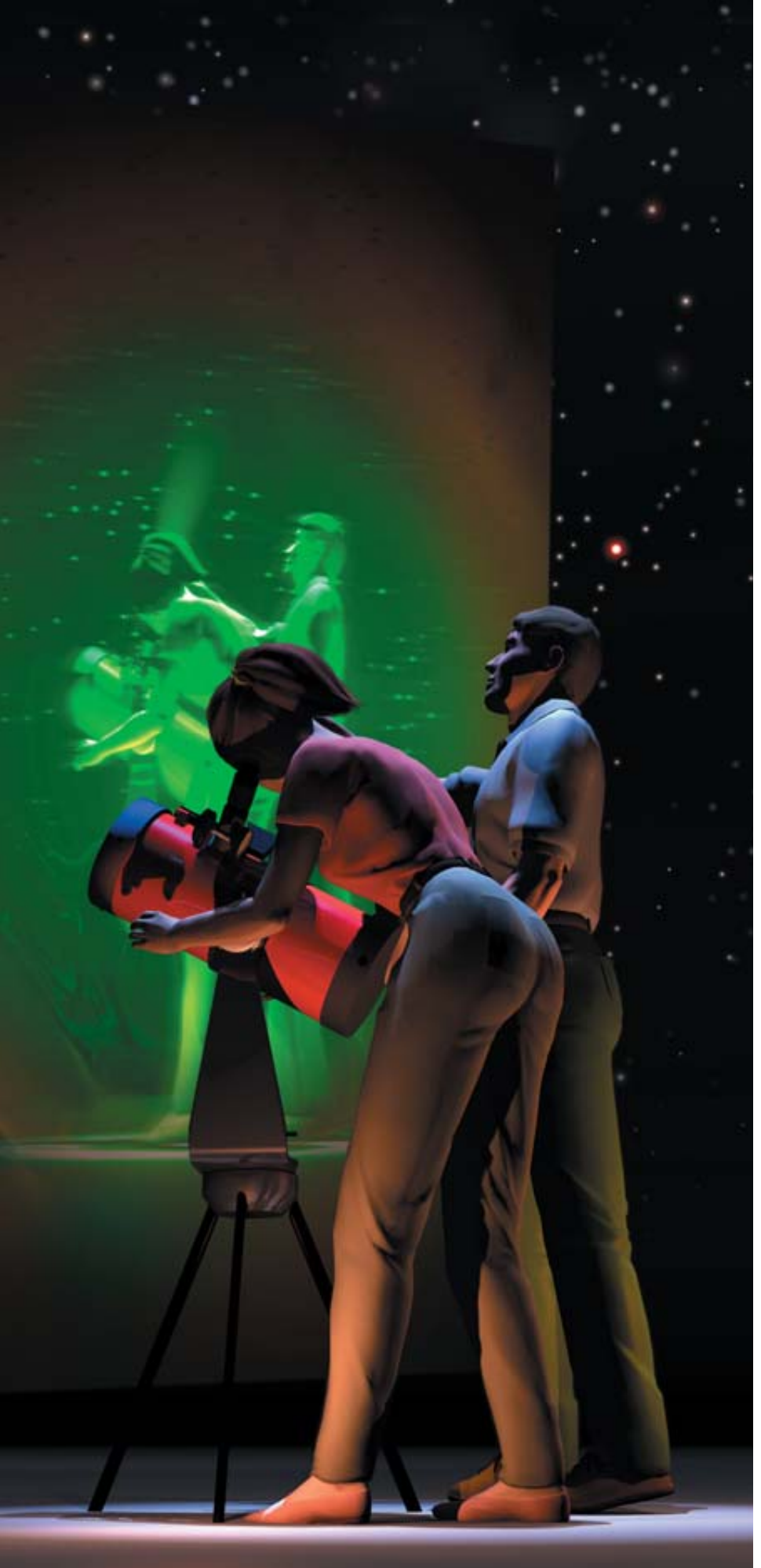
Theoretical results about black holes suggest that the universe could be like a gigantic hologram

By Jacob D. Bekenstein

originally published in August 2003

Illustrations by Alfred T. Kamajian

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



Ask anybody what the physical world is made of, and you are likely to be told "matter and energy."

Yet if we have learned anything from engineering, biology and physics, information is just as crucial an ingredient. The robot at the automobile factory is supplied with metal and plastic but can make nothing useful without copious instructions telling it which part to weld to what and so on. A ribosome in a cell in your body is supplied with amino acid building blocks and is powered by energy released by the conversion of ATP to ADP, but it can synthesize no proteins without the information brought to it from the DNA in the cell's nucleus. Likewise, a century of developments in physics has taught us that information is a crucial player in physical systems and processes. Indeed, a current trend, initiated by John A. Wheeler of Princeton University, is to regard the physical world as made of information, with energy and matter as incidentals.

This viewpoint invites a new look at venerable questions. The information storage capacity of devices such as hard disk drives has been increasing by leaps and bounds. When will such progress halt? What is the ultimate information capacity of a device that weighs, say, less than a gram and can fit inside a cubic centimeter (roughly the size of a computer chip)? How much information

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

does it take to describe a whole universe? Could that description fit in a computer's memory? Could we, as William Blake memorably penned, "see the world in a grain of sand," or is that idea no more than poetic license?

Remarkably, recent developments in theoretical physics answer some of these questions, and the answers might be important clues to the ultimate theory of reality. By studying the mysterious properties of black holes, physicists have deduced absolute limits on how much information a region of space or a quantity of matter and energy can hold. Related results suggest that our universe, which we perceive to have three spatial dimensions, might instead be "written" on a two-dimensional surface, like a hologram. Our everyday perceptions of the world as three-dimensional would then be either a profound illusion or merely one of two alternative ways of viewing reality. A grain of sand may not encompass our world, but a flat screen might.

A Tale of Two Entropies

FORMAL INFORMATION theory originated in seminal 1948 papers by American applied mathematician Claude E. Shannon, who introduced today's most widely used measure of information content: entropy. Entropy had long been a central concept of thermodynamics, the branch of physics dealing with heat. Thermodynamic entropy is popularly described as the disorder in a physical system. In 1877 Austrian physicist Ludwig Boltzmann characterized it more precisely in terms of the number of distinct microscopic states that the particles composing a chunk of matter could be in while still looking like the same macroscopic chunk of matter. For example, for the air in the room around you, one would count all the ways that the individual gas molecules could be distributed in the room and all the ways they could be moving.

When Shannon cast about for a way to quantify the information contained in, say, a message, he was led by logic to a formula with the same form as Boltzmann's. The Shannon entropy of a message is the number of binary digits, or bits, needed to encode it. Shannon's entropy does not enlighten us about the value of information, which is highly dependent on context. Yet as an objective measure of quantity of information, it has been enormously useful in science and technology. For instance, the design of every modern communications device—from cellular phones to modems to compact-disc players—relies on Shannon entropy.

Thermodynamic entropy and Shannon entropy are conceptually equivalent: the number of arrangements that are counted by Boltzmann entropy reflects the amount of Shannon information one would need to implement any particular arrangement. The two entropies have two salient differences, though. First, the thermodynamic entropy used by a chemist or a refrigeration engineer is expressed in units of energy divided by temperature, whereas the Shannon entropy used by a communications engineer is in bits, essentially dimensionless. That difference is merely a matter of convention.

Even when reduced to common units, however, typical values of the two entropies differ vastly in magnitude. A silicon microchip carrying a gigabyte of data, for instance, has a Shannon entropy of about 10^{10} bits (one byte is eight bits), tremendously smaller than the chip's thermodynamic entropy, which is about 10^{23} bits at room temperature. This discrepancy occurs because the entropies are computed for different degrees of freedom. A degree of freedom is any quantity that can vary, such as a coordinate specifying a particle's location or one component of its velocity. The Shannon entropy of the chip cares only about the overall state of each tiny transistor etched in the silicon crystal—the transistor is on or off; it is a 0 or a 1—a single binary degree of freedom. Thermodynamic entropy, in contrast, depends on the states of all the billions of atoms (and their roaming electrons) that make up each transistor. As miniaturization brings closer the day when each atom will store one bit of information for us, the useful Shannon entropy of the state-of-the-art microchip will edge closer in magnitude to its material's thermodynamic entropy. When the two entropies are calculated for the same degrees of freedom, they are equal.

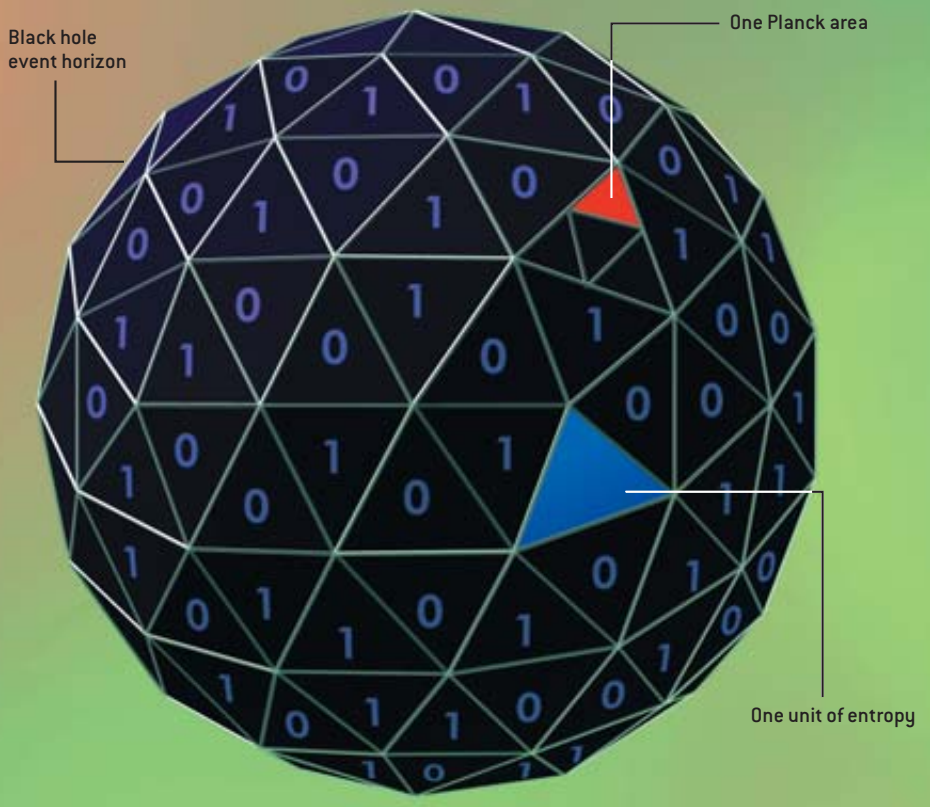
What are the ultimate degrees of freedom? Atoms, after all, are made of electrons and nuclei, nuclei are agglomerations of protons and neutrons, and those in turn are composed of quarks. Many physicists today consider electrons and quarks to be excitations of superstrings, which they hypothesize to be the most fundamental entities. But the vicissitudes of a century of revelations in physics warn us not to be dogmatic. There could be more levels of structure in our universe than are dreamt of in today's physics.

One cannot calculate the ultimate information capacity of a chunk of matter or, equivalently, its true thermodynamic entropy, without knowing the nature of the ultimate constituents of matter or of the deepest level of structure, which I shall refer to as level X. (This ambiguity causes no problems in analyzing practical thermodynamics, such as that of car

Overview/*The World as a Hologram*

- An astonishing theory called the holographic principle holds that the universe is like a hologram: just as a trick of light allows a fully three-dimensional image to be recorded on a flat piece of film, our seemingly three-dimensional universe could be completely equivalent to alternative quantum fields and physical laws "painted" on a distant, vast surface.
- The physics of black holes—immensely dense concentrations of mass—provides a hint that the principle might be true. Studies of black holes show that, although it defies common sense, the maximum entropy or information content of any region of space is defined not by its volume but by its surface area.
- Physicists hope that this surprising finding is a clue to the ultimate theory of reality.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



THE ENTROPY OF A BLACK HOLE is proportional to the area of its event horizon, the surface within which even light cannot escape the gravity of the hole. Specifically, a hole with a horizon spanning *A* Planck areas has $A/4$ units of entropy. (The Planck area, approximately 10^{-66} square centimeter, is the fundamental quantum unit of area determined by the strength of gravity, the speed of light and the size of quanta.) Considered as information, it is as if the entropy were written on the event horizon, with each bit (each digital 1 or 0) corresponding to four Planck areas.

engines, for example, because the quarks within the atoms can be ignored—they do not change their states under the relatively benign conditions in the engine.) Given the dizzying progress in miniaturization, one can playfully contemplate a day when quarks will serve to store information, one bit apiece perhaps. How much information would then fit into our one-centimeter cube? And how much if we harness superstrings or even deeper, yet undreamt of levels? Surprisingly, developments in gravitation physics in the past three decades have supplied some clear answers to what seem to be elusive questions.

Black Hole Thermodynamics

A CENTRAL PLAYER in these developments is the black hole. Black holes are a consequence of general relativity, Albert Einstein's 1915 geometric theory of gravitation. In this theory, gravitation arises from the curvature of spacetime, which makes objects move as if they were pulled by a force. Conversely, the curvature is caused by the presence of matter and energy. According to Einstein's equations, a sufficiently dense concentration of matter or energy will curve spacetime so extremely that it rends, forming a black hole. The laws of relativity forbid anything that went into a black hole from coming out again, at least within the classical (nonquantum) description of the physics. The point of no return, called the event horizon of the black hole, is of crucial importance. In the simplest case, the horizon is a sphere, whose surface area is larger for more massive black holes.

It is impossible to determine what is inside a black hole. No detailed information can emerge across the horizon and escape into the outside world. In disappearing forever into a black hole, however, a piece of matter does leave some traces. Its energy (we count any mass as energy in accordance with Einstein's $E = mc^2$) is permanently reflected in an increment in the black hole's mass. If the matter is captured while circling the hole, its associated angular momentum is added to the black hole's angular momentum. Both the mass and angular momentum of a black hole are measurable from their effects on spacetime around the hole. In this way, the laws of conservation of energy and angular momentum are upheld by black holes. Another fundamental law, the second law of thermodynamics, appears to be violated.

The second law of thermodynamics summarizes the familiar observation that most processes in nature are irreversible: a teacup falls from the table and shatters, but no one has ever seen shards jump up of their own accord and assemble into a teacup. The second law of thermodynamics forbids such inverse processes. It states that the entropy of an isolated physical system can never decrease; at best, entropy remains constant, and usually it increases. This law is central to physical chemistry and engineering; it is arguably the physical law with the greatest impact outside physics.

As first emphasized by Wheeler, when matter disappears into a black hole, its entropy is gone for good, and the second law seems to be transcended, made irrelevant. A clue to resolving this puzzle came in 1970, when Demetrious Christodoulou, then a graduate student of Wheeler's at Princeton, and Stephen W. Hawking of the University of Cambridge independently proved that in various processes, such as black hole mergers, the total area of the event horizons never decreases. The analogy with the tendency of entropy to increase led me to propose in 1972 that a black hole has entropy proportional to

THE AUTHOR

JACOB D. BEKENSTEIN has contributed to the foundation of black hole thermodynamics and to other aspects of the connections between information and gravitation. He is Polak Professor of Theoretical Physics at the Hebrew University of Jerusalem, a member of the Israel Academy of Sciences and Humanities, and a recipient of the Rothschild Prize. Bekenstein dedicates this article to John Archibald Wheeler (his Ph.D. supervisor 30 years ago). Wheeler belongs to the third generation of Ludwig Boltzmann's students: Wheeler's Ph.D. adviser, Karl Herzfeld, was a student of Boltzmann's student Friedrich Hasenöhrl.

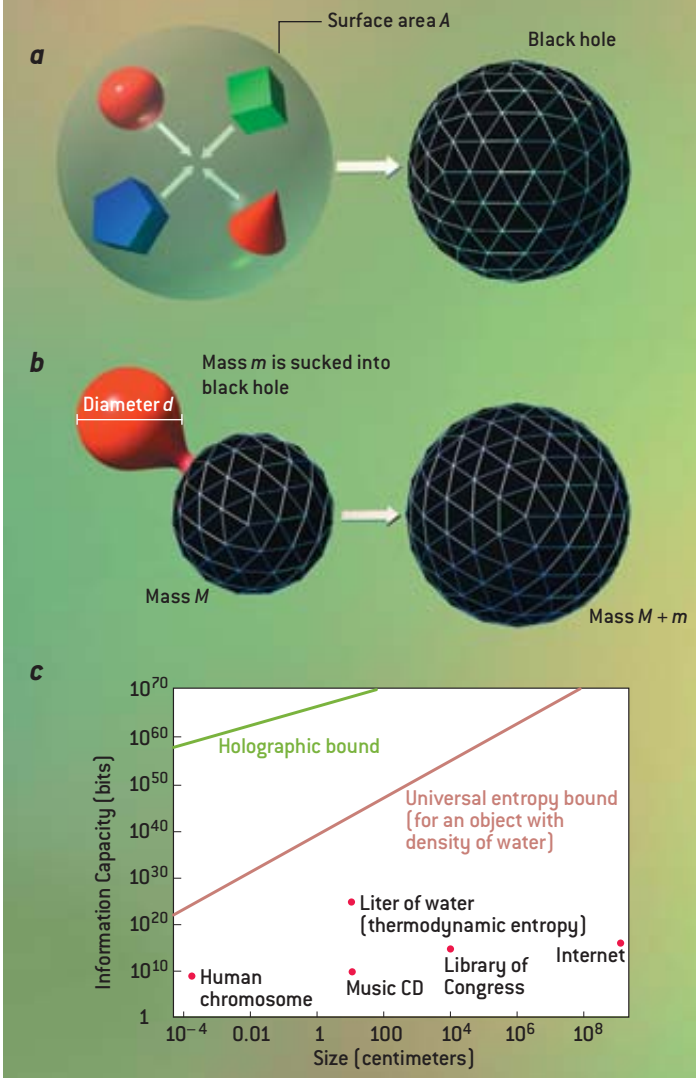
COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

the area of its horizon [*see illustration on preceding page*]. I conjectured that when matter falls into a black hole, the increase in black hole entropy always compensates or overcompensates for the "lost" entropy of the matter. More generally, the sum of black hole entropies and the ordinary entropy outside the black holes cannot decrease. This is the generalized second law—GSL for short.

The GSL has passed a large number of stringent, if purely theoretical, tests. When a star collapses to form a black hole, the black hole entropy greatly exceeds the star's entropy. In 1974 Hawking demonstrated that a black hole spontaneously emits thermal radiation, now known as Hawking radiation, by a quantum process [see "The Quantum Mechanics of Black Holes," by Stephen W. Hawking; SCIENTIFIC AMERICAN, January 1977]. The Christodoulou-Hawking theorem fails in the face of this phenomenon (the mass of the black hole, and therefore its horizon area, decreases), but the GSL copes with it: the entropy of the emergent radiation more than compensates for the decrement in black hole entropy, so the GSL is preserved. In 1986 Rafael D. Sorkin of Syracuse University exploited the horizon's role in barring information inside the black hole from influencing affairs outside to show that the GSL (or something very similar to it) must be valid for any conceivable process that black holes undergo. His deep argument makes it clear that the entropy entering the GSL is that calculated down to level X, whatever that level may be.

Hawking's radiation process allowed him to determine the proportionality constant between black hole entropy and horizon area: black hole entropy is precisely one quarter of the event horizon's area measured in Planck areas. (The Planck length, about 10^{-33} centimeter, is the fundamental length scale related to gravity and quantum mechanics. The Planck area is its square.) Even in thermodynamic terms, this is a vast quantity of entropy. The entropy of a black hole

LIMITS ON INFORMATION DENSITY



THE THERMODYNAMICS OF BLACK HOLES allows one to deduce limits on the density of entropy or information in various circumstances.

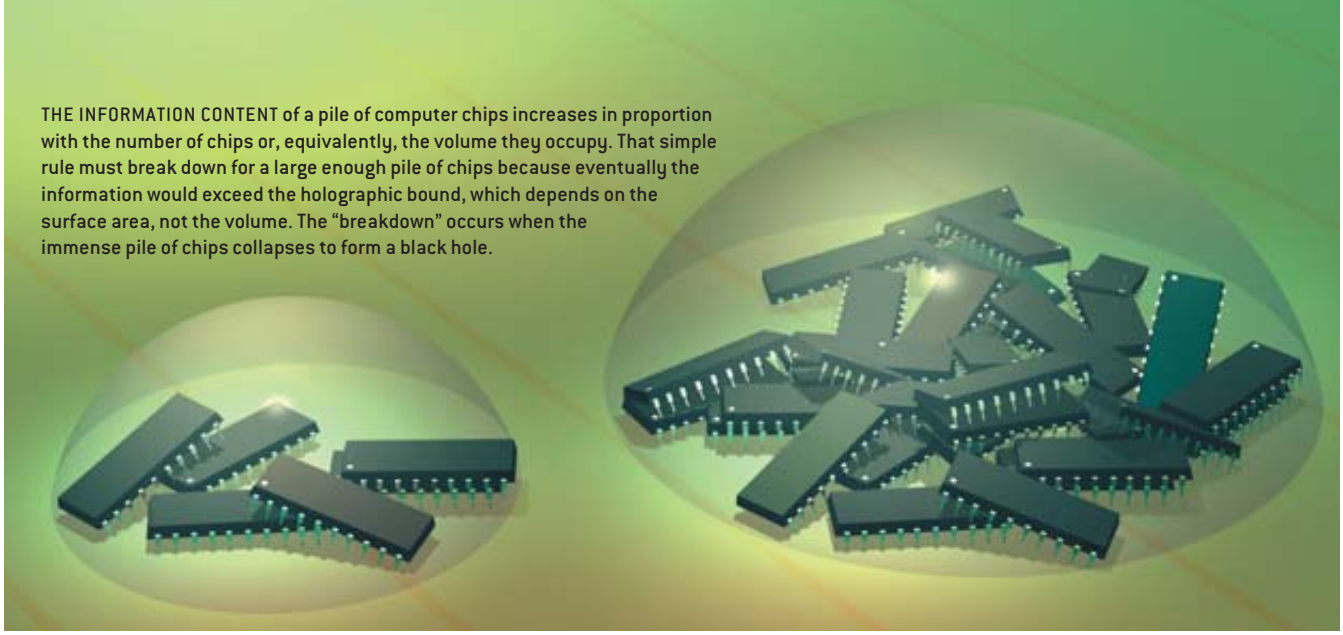
The holographic bound defines how much information can be contained in a specified region of space. It can be derived by considering a roughly spherical distribution of matter that is contained within a surface of area *A*. The matter is induced to collapse to form a black hole (*a*). The black hole's area must be smaller than *A*, so its entropy must be less than $A/4$ [*see illustration on preceding page*]. Because entropy cannot decrease, one infers that the original distribution of matter also must carry less than $A/4$ units of entropy or information. This result—that the maximum information content of a region of space is fixed by its area—defies the commonsense expectation that the capacity of a region should depend on its volume.

The universal entropy bound defines how much information can be carried by a mass *m* of diameter *d*. It is derived by imagining that a capsule of matter is engulfed by a black hole not much wider than it (*b*). The increase in the black hole's size places a limit on how much entropy the capsule could have contained. This limit is tighter than the holographic bound, except when the capsule is almost as dense as a black hole (in which case the two bounds are equivalent).

The holographic and universal information bounds are far beyond the data storage capacities of any current technology, and they greatly exceed the density of information on chromosomes and the thermodynamic entropy of water (*c*). —*J.D.B.*

LAURIE GRACE (*graph*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



THE INFORMATION CONTENT of a pile of computer chips increases in proportion with the number of chips or, equivalently, the volume they occupy. That simple rule must break down for a large enough pile of chips because eventually the information would exceed the holographic bound, which depends on the surface area, not the volume. The "breakdown" occurs when the immense pile of chips collapses to form a black hole.

one centimeter in diameter would be about 10^{66} bits, roughly equal to the thermodynamic entropy of a cube of water 10 billion kilometers on a side.

The World as a Hologram

THE GSL ALLOWS US to set bounds on the information capacity of any isolated physical system, limits that refer to the information at all levels of structure down to level X. In 1980 I began studying the first such bound, called the universal entropy bound, which limits how much entropy can be carried by a specified mass of a specified size [*see box on previous page*]. A related idea, the holographic bound, was devised in 1995 by Leonard Susskind of Stanford University. It limits how much entropy can be contained in matter and energy occupying a specified volume of space.

In his work on the holographic bound, Susskind considered any approximately spherical isolated mass that is not itself a black hole and that fits inside a closed surface of area A. If the mass can collapse to a black hole, that hole will end up with a horizon area smaller than A. The black hole entropy is therefore smaller than $A/4$. According to the GSL, the entropy of the system cannot decrease, so the mass's original entropy cannot have been bigger than $A/4$. It follows that the entropy of an isolated physical system with boundary area A is necessarily less than $A/4$. What if the mass does not spontaneously collapse? In 2000 I showed that a tiny black hole can be used to convert the system to a black hole not much different from the one in Susskind's argument. The bound is therefore independent of the constitution of the system or of the nature of level X. It just depends on the GSL.

We can now answer some of those elusive questions about the ultimate limits of information storage. A device measuring a centimeter across could in principle hold up to 10^{66} bits—a mind-boggling amount. The visible universe contains at least 10^{100} bits of entropy, which could in principle be packed inside a sphere a tenth of a light-year across. Estimating the entropy of the universe is a difficult problem, however, and much larger numbers, requiring a sphere almost as big as the universe itself, are entirely plausible.

But it is another aspect of the holographic bound that is truly astonishing. Namely, that the maximum possible entropy depends on the boundary area instead of the volume. Imagine that we are piling up computer memory chips in a big heap. The number of transistors—the total data storage capacity—increases with the volume of the heap. So, too, does the total thermodynamic entropy of all the chips. Remarkably, though, the theoretical ultimate information capacity of the space occupied by the heap increases only with the surface area. Because volume increases more rapidly than surface area, at some point the entropy of all the chips would exceed the holographic bound. It would seem that either the GSL or our commonsense ideas of entropy and information capacity must fail. In fact, what fails is the pile itself: it would collapse under its own gravity and form a black hole before that impasse was reached. Thereafter each additional memory chip would increase the mass and surface area of the black hole in a way that would continue to preserve the GSL.

This surprising result—that information capacity depends on surface area—has a natural explanation if the holographic *principle* (proposed in 1993 by Nobelist Gerard 't Hooft of the University of Utrecht in the Netherlands and elaborated by Susskind) is true. In the everyday world, a hologram is a special kind of photograph that generates a full three-dimensional image when it is illuminated in the right manner. All the information describing the 3-D scene is encoded into the pattern of light and dark areas on the two-dimensional piece of film, ready to be regenerated. The holographic principle contends that an analogue of this visual magic applies to the full physical description of any system occupying a 3-D region: it proposes that another physical theory defined only on the 2-D boundary of the region completely describes the 3-D physics. If a 3-D system can be fully described by a physical theory operating solely on its 2-D boundary, one would expect the information content of the system not to exceed that of the description on the boundary.

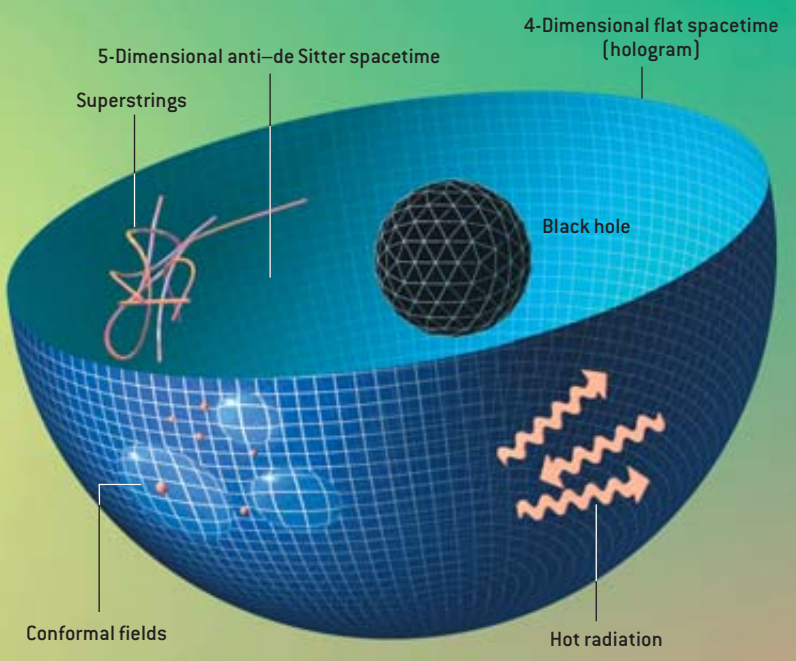
A Universe Painted on Its Boundary

CAN WE APPLY the holographic principle to the universe at large? The real universe is a 4-D system: it has volume and extends in time. If the physics of our universe is holographic, there would be an alternative set of physical laws, operating on a 3-D boundary of spacetime

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

A HOLOGRAPHIC SPACETIME

TWO UNIVERSES of different dimension and obeying disparate physical laws are rendered completely equivalent by the holographic principle. Theorists have demonstrated this principle mathematically for a specific type of five-dimensional spacetime ("anti–de Sitter") and its four-dimensional boundary. In effect, the 5-D universe is recorded like a hologram on the 4-D surface at its periphery. Superstring theory rules in the 5-D spacetime, but a so-called conformal field theory of point particles operates on the 4-D hologram. A black hole in the 5-D spacetime is equivalent to hot radiation on the hologram—for example, the hole and the radiation have the same entropy even though the physical origin of the entropy is completely different for each case. Although these two descriptions of the universe seem utterly unalike, no experiment could distinguish between them, even in principle. —*J.D.B.*



somewhere, that would be equivalent to our known 4-D physics. We do not yet know of any such 3-D theory that works in that way. Indeed, what surface should we use as the boundary of the universe? One step toward realizing these ideas is to study models that are simpler than our real universe.

A class of concrete examples of the holographic principle at work involves so-called anti–de Sitter spacetimes. The original de Sitter spacetime is a model universe first obtained by Dutch astronomer Willem de Sitter in 1917 as a solution of Einstein's equations, including the repulsive force known as the cosmological constant. De Sitter's spacetime is empty, expands at an accelerating rate and is very highly symmetrical. In 1997 astronomers studying distant supernova explosions concluded that our universe now expands in an accelerated fashion and will probably become increasingly like a de Sitter spacetime in the future. Now, if the repulsion in Einstein's equations is changed to attraction, de Sitter's solution turns into the anti–de Sitter spacetime, which has equally as much symmetry. More important for the holographic concept, it possesses a boundary, which is located "at infinity" and is a lot like our everyday spacetime.

Using anti–de Sitter spacetime, theorists have devised a concrete example of the holographic principle at work: a universe described by superstring theory functioning in an anti–de Sitter spacetime is completely equivalent to a quantum field theory operating on the boundary of that spacetime [*see box above*]. Thus, the full majesty of superstring theory in an anti–de Sitter universe is painted on the boundary of the universe. Juan Maldacena, then at Harvard University, first conjectured such a relation in 1997 for the 5-D anti–de Sitter case, and it was later confirmed for many situations by Edward Witten of the Institute for Advanced Study in Princeton, N.J., and Steven S. Gubser, Igor R. Klebanov and Alexander M. Polyakov of Princeton University. Examples of this holographic correspondence are now known for spacetimes with a variety of dimensions.

This result means that two ostensibly very different theories—not even acting in spaces of the same dimension—are equivalent. Creatures living in one of these universes would be incapable of determining if they inhabited a 5-D universe described by string theory or a 4-D one described by a quantum field theory of point particles. (Of course, the structures of their brains might give them an overwhelming "commonsense" prejudice in favor of one description or another, in just the way that our brains construct an innate perception that our universe has three spatial dimensions; see the illustration on the opposite page.)

The holographic equivalence can allow a difficult calculation in the 4-D boundary spacetime, such as the behavior of quarks and gluons, to be traded for another, easier calculation in the highly symmetric, 5-D anti–de Sitter spacetime. The correspondence works the other way, too. Witten has shown that a black hole in anti–de Sitter spacetime corresponds to hot radiation in the alternative physics operating on the bounding spacetime. The entropy of the hole—a deeply mysterious concept—equals the radiation's entropy, which is quite mundane.

The Expanding Universe

HIGHLY SYMMETRIC and empty, the 5-D anti–de Sitter universe is hardly like our universe existing in 4-D, filled with matter and radiation, and riddled with violent events. Even if we approximate our real universe with one that has matter and radiation spread uniformly throughout, we get not an anti–de Sitter universe but rather a "Friedmann-Robertson-Walker" universe. Most cosmologists today concur

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

that our universe resembles an FRW universe, one that is infinite, has no boundary and will go on expanding ad infinitum.

Does such a universe conform to the holographic principle or the holographic bound? Susskind's argument based on collapse to a black hole is of no help here. Indeed, the holographic bound deduced from black holes must break down in a uniform expanding universe. The entropy of a region uniformly filled with matter and radiation is truly proportional to its volume. A sufficiently large region will therefore violate the holographic bound.

In 1999 Raphael Bousso, then at Stanford, proposed a modified holographic bound, which has since been found to work even in situations where the bounds we discussed earlier cannot be applied. Bousso's formulation starts with any suitable 2-D surface; it may be closed like a sphere or open like a sheet of paper. One then imagines a brief burst of light issuing simultaneously and perpendicularly from all over one side of the surface. The only demand is that the imaginary light rays are converging to start with. Light emitted from the inner surface of a spherical shell, for instance, satisfies that requirement. One then considers the entropy of the matter and radiation that these imaginary rays traverse, up to the points where they start crossing. Bousso conjectured that this entropy cannot exceed the entropy represented by the initial surface—one quarter of its area, measured in Planck areas. This is a different way of tallying up the entropy than that used in the original holographic bound. Bousso's bound refers not to the entropy of a region at one time but rather to the sum of entropies of locales at a variety of times: those that are "illuminated" by the light burst from the surface.

Bousso's bound subsumes other entropy bounds while avoiding their limitations. Both the universal entropy bound and the 't Hooft-Susskind form of the holographic bound can be deduced from Bousso's for any isolated system that is not evolving rapidly and whose gravitational field is not strong. When these conditions are overstepped—as for a collapsing sphere of matter already inside a black hole—these bounds eventually fail, whereas Bousso's bound continues to hold. Bousso has also shown that his strategy can be used to locate the 2-D surfaces on which holograms of the world can be set up.

Augurs of a Revolution

RESEARCHERS HAVE proposed many other entropy bounds. The proliferation of variations on the holographic motif makes it clear that the subject has not yet reached the status of physical law. But although the holographic way of thinking is not yet fully understood, it seems to be here to stay. And with it comes a realization that the fundamental belief, prevalent for 50 years, that field theory is the ultimate language of physics must give way. Fields, such as the electromagnetic field, vary continuously from point to point, and they thereby describe an infinity of degrees of freedom. Superstring theory also embraces an infinite number of degrees of freedom. Holography restricts the number of degrees of freedom that can be present inside a bounding surface to a finite number; field theory with its infinity cannot be the final story. Furthermore, even if the infinity is tamed, the mysterious dependence of information on surface area must be somehow accommodated.

Holography may be a guide to a better theory. What is the fundamental theory like? The chain of reasoning involving holography suggests to some, notably Lee Smolin of the Perimeter Institute for Theoretical Physics in Waterloo, that such a final theory must be concerned not with fields, not even with spacetime, but rather with information exchange among physical processes. If so, the vision of information as the stuff the world is made of will have found a worthy embodiment. SA

OUR INNATE PERCEPTION that the world is three-dimensional could be an extraordinary illusion.

MORE TO EXPLORE

Black Hole Thermodynamics. Jacob D. Bekenstein in *Physics Today,* Vol. 33, No. 1, pages 24–31; January 1980.

Black Holes and Time Warps: Einstein's Outrageous Legacy. Kip S. Thorne. W. W. Norton, 1995.

Black Holes and the Information Paradox. Leonard Susskind in *Scientific American,* Vol. 276, No. 4, pages 52–57; April 1997.

The Universe in a Nutshell. Stephen Hawking. Bantam Books, 2001.

Three Roads to Quantum Gravity. Lee Smolin. Basic Books, 2002.

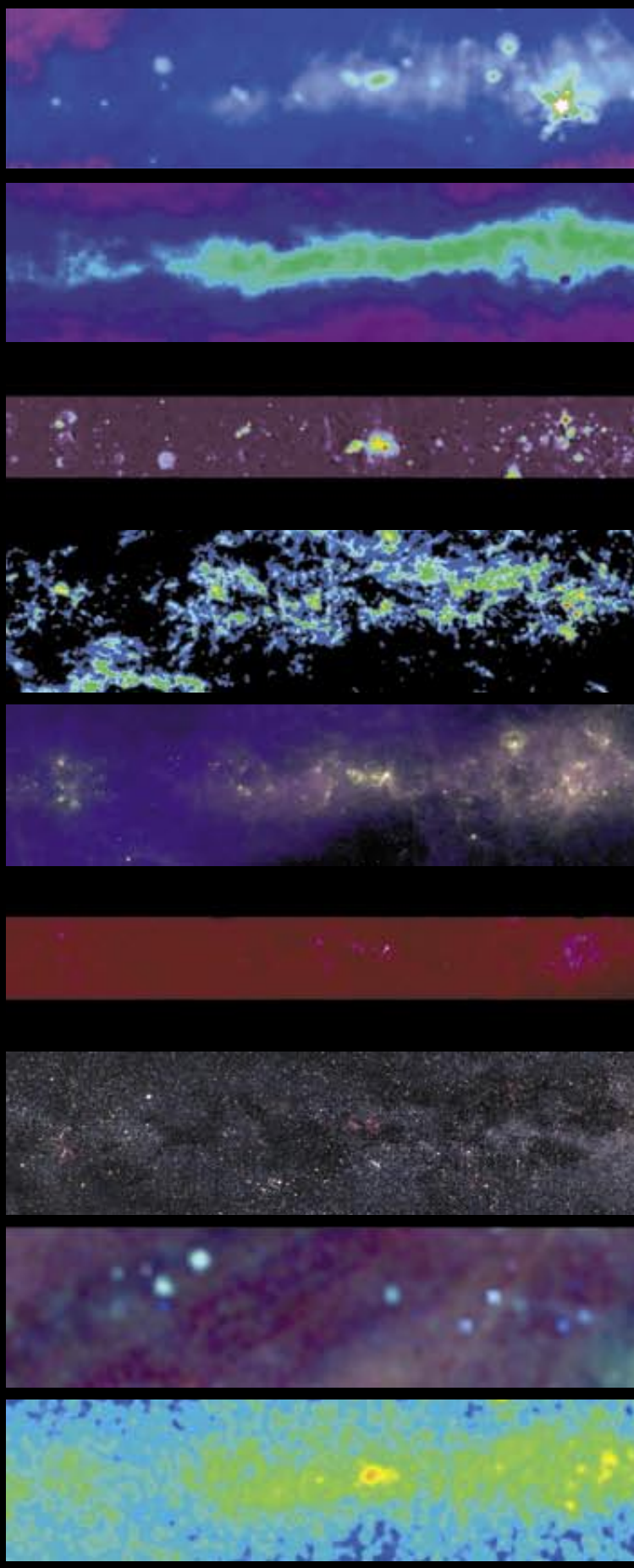
COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

The Gas between the Stars

by Ronald J. Reynolds

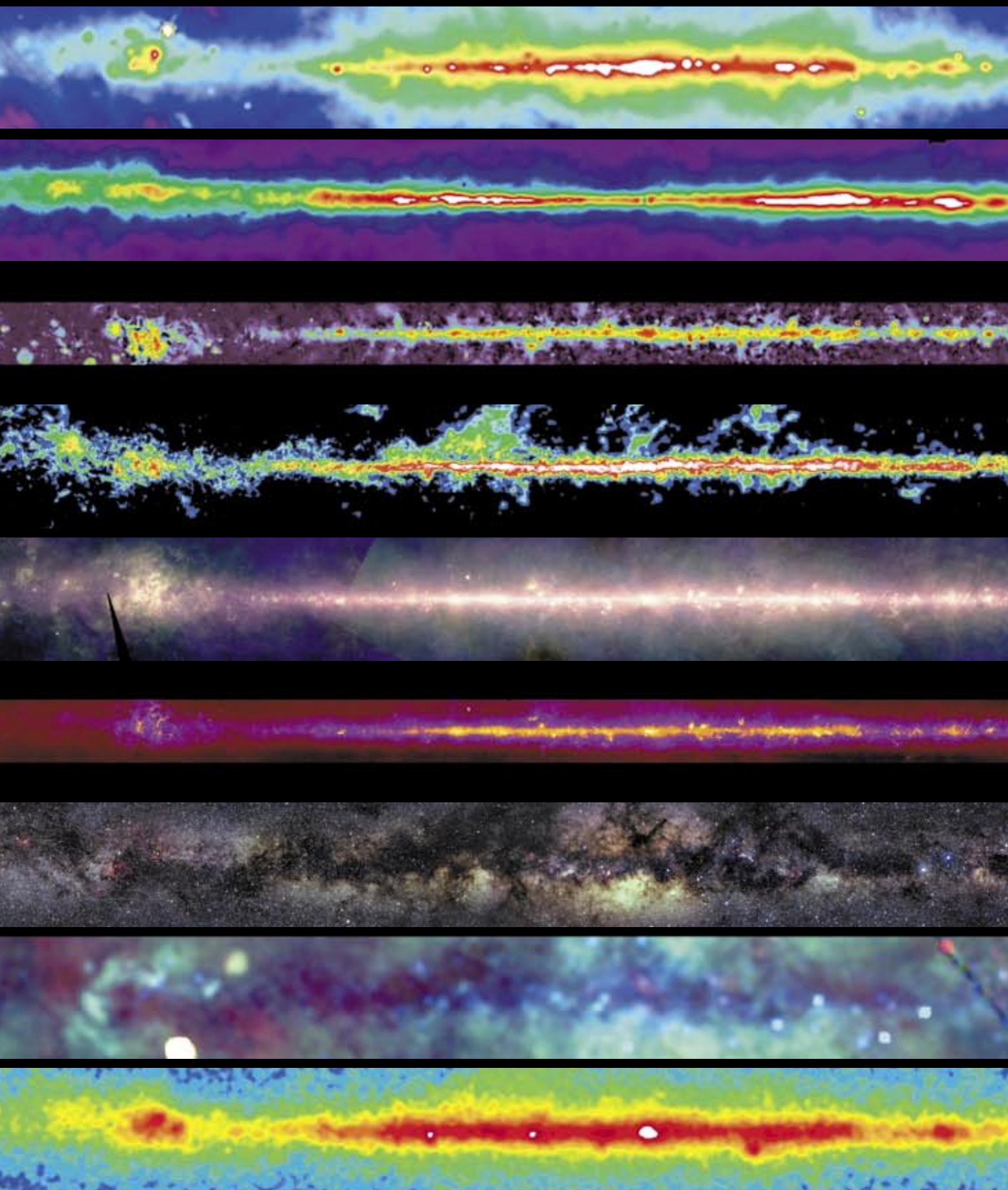
originally published in January 2002

Filled with colossal fountains of hot gas and vast bubbles blown by exploding stars, the interstellar medium is far more interesting than scientists once thought



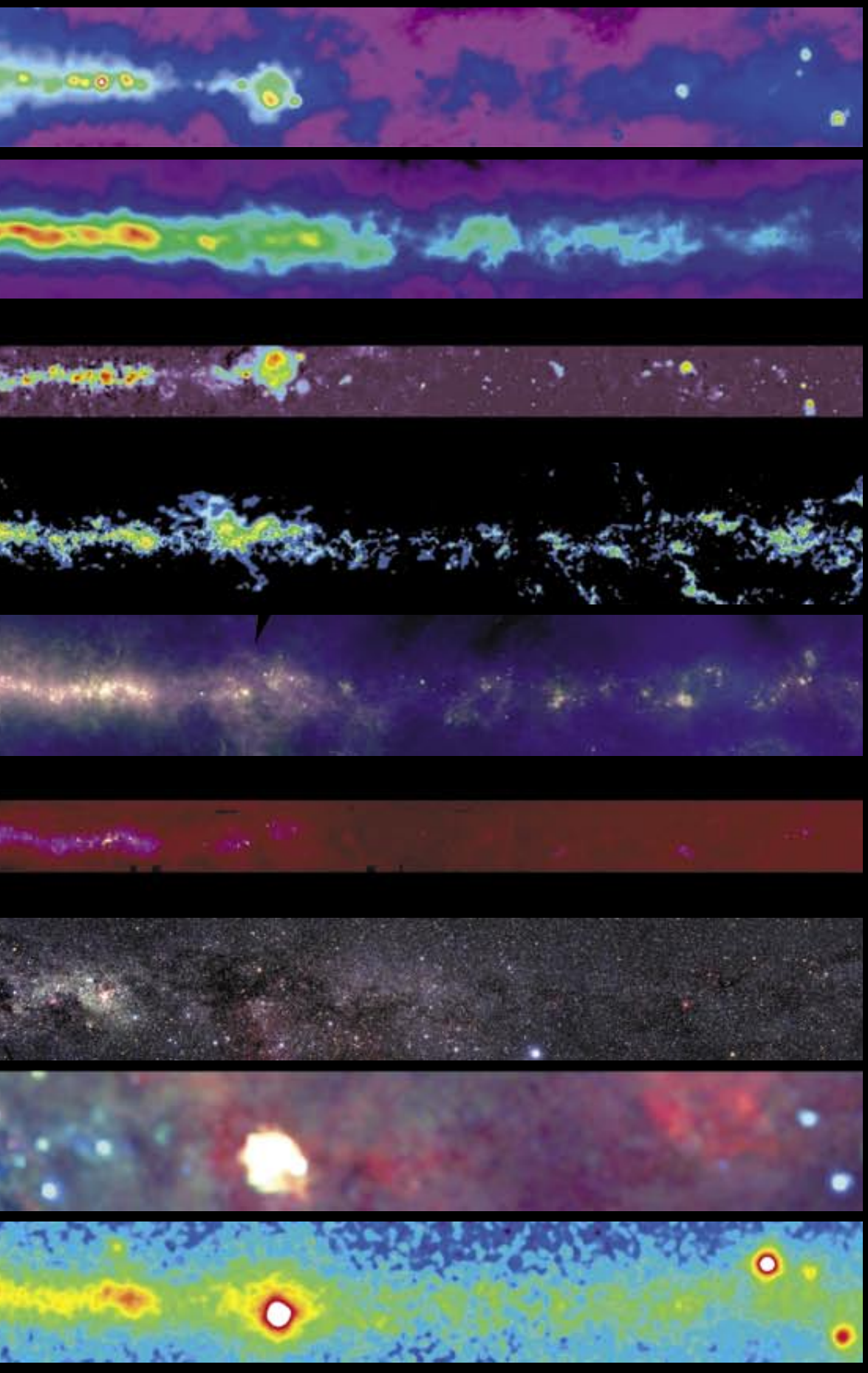
MILKY WAY GALAXY looks profoundly different depending on the frequency at which astronomers observe it. Fifty years ago, when astronomers were restricted to visible light, interstellar gas seemed like just a nuisance—blocking the real objects of interest, the stars. Today scientists think the gas may be as important to the evolution of the galaxy as are the stars. These panels appear on a poster prepared by the NASA Goddard Space Flight Center; for more information, visit http://nvo.gsfc.nasa.gov/mw/mmw_sci.html

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



NASA GSFC ASTROPHYSICS DATA FACILITY (*radio continuum [408 MHz], atomic hydrogen, far-infrared, x-ray and gamma ray*); ROY DUNCAN *Software Infrastructure Group* (*radio continuum (2.4–2.7 GHz)*); THOMAS DAME *Harvard-Smithsonian Center for Astrophysics* (*molecular hydrogen*); STEPHAN D. PRICE *Hanscom AFB* (*mid-infrared*); AXEL MELLINGER *University of Potsdam* (*visible light*)

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



RADIO CONTINUUM
(408 MHz)
Reveals fast-moving electrons, found especially at sites of past supernovae

ATOMIC HYDROGEN
(1420 MHz)
Reveals neutral atomic hydrogen in interstellar clouds and diffuse gas

RADIO CONTINUUM
(2.4–2.7 GHz)
Reveals warm, ionized gas and high-energy electrons

MOLECULAR HYDROGEN
(115 GHz)
Reveals molecular hydrogen (as traced by carbon monoxide) in cold clouds

FAR-INFRARED
(12–100 microns)
Reveals dust warmed by starlight, specially in star-forming regions

MID-INFRARED
(6.8–10.8 microns)
Reveals complex molecules in interstellar clouds, as well as reddish stars

VISIBLE LIGHT
(0.4–0.6 micron)
Reveals nearby stars and tenuous ionized gas; dark areas are cold and dense

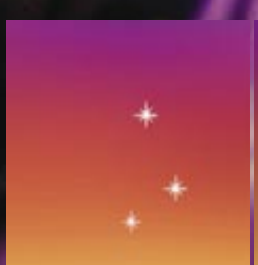
X-RAY
(0.25–1.5 kiloelectron-volt)
Reveals hot, shocked gas from supernovae

GAMMA RAY
(greater than 300 megaelectron-volts)
Reveals high-energy phenomena such as pulsars and cosmic-ray collisions

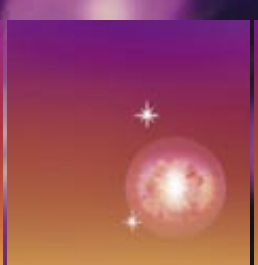
COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

The Galaxy's Dynamic Atmosphere

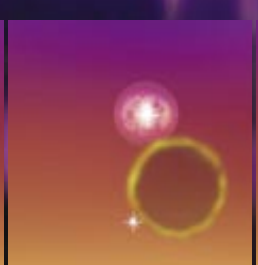
The views above and on the preceding page are cross sections through the Milky Way.



1 A superbubble originates with a cluster of massive stars.



2 One star goes supernova, forming a bubble of hot, low-density gas.



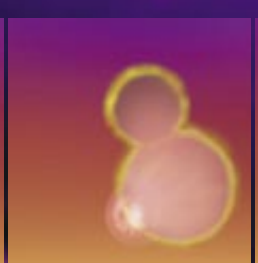
3 Because massive stars have similar lifespans, another one soon blows.

4 wi bu

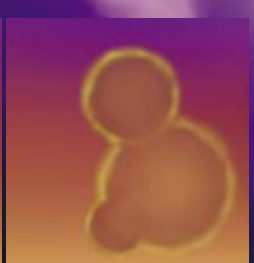
COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

The term "interstellar medium" once conjured up a picture like the one at right: frigid, inky clouds of gas and dust in repose near the galactic plane. Today astronomers recognize the medium as a protean atmosphere roiled by supernova explosions. Gas gushes through towering chimneys, then showers back down in mighty fountains.

The two bubbles link up. Stellar nds help energize the bbles.



5 A third explodes. The interstellar medium starts to look like Swiss cheese.



6 All three bubbles link up, forming a passage for hot gas and radiation.

Composition of the Galactic Atmosphere

	IN CLOUDS		BETWEEN CLOUDS		
Component	H_2	H I	WARM H I	WARM H II	HOT H II
Temperature (K)	15	120	8,000	8,000	$\sim 10^6$
Midplane Density (cm^{-3})	200	25	0.3	0.15	0.002
Thickness of Layer (parsecs)	150	200	1,000	2,000	6,000
Volume Fraction (%)	0.1	2	35	20	43
Mass Fraction (%)	18	30	30	20	2

Some of the interstellar medium takes the form of discrete clouds of atomic hydrogen (H I) or molecular hydrogen (H_2); most of the rest is in a pervasive ionized (H II) or atomic gas. Intermixed is a trace amount of other elements. The total mass is about one fifth of the galaxy's stars.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

We often think of the moon as a place, but in fact it is a hundred million places, an archipelago of solitude. You could go from

100 degrees below zero to 100 degrees above with a small step. You could yell in your friend's ear and he would never hear you. Without an atmosphere to transmit heat or sound, each patch of the moon is an island in an unnavigable sea.

The atmosphere of a planet is what binds its surface into a unified whole. It lets conditions such as temperature vary smoothly. More dramatically, events such as the impact of an asteroid, the eruption of a volcano and the emission of gas from a factory's chimney can have effects that reach far beyond the spots where they took place. Local phenomena can have global consequences. This characteristic of atmospheres has begun to capture the interest of astronomers who study the Milky Way galaxy.

For many years, we have known that an extremely thin atmosphere called the interstellar medium envelops our galaxy and threads the space between its billions of stars. Until fairly recently, the medium seemed a cold, static reservoir of gas quietly waiting to condense into stars. You barely even notice it when looking up into the starry sky. Now we recognize the medium as a tempestuous mixture with an extreme diversity of density, temperature and ionization. Supernova explosions blow giant bubbles; fountains and chimneys may arch above the spiral disk; and clouds could be falling in from beyond the disk. These and other processes interconnect far-flung reaches of our galaxy much as atmospheric phenomena convey disturbances from one side of Earth to the other.

In fact, telescopes on the ground and in space are showing the galaxy's atmosphere to be as complex as any planet's. Held by the combined gravitational pull of the stars and other matter, permeated by starlight, energetic particles and a magnetic field, the interstellar medium is continuously stirred, heated, recycled and transformed. Like any atmosphere, it has its highest density and pressure at the "bottom," in this case the plane that defines the middle of the galaxy, where the pressure must balance the weight of the medium from "above." Dense concentrations of gas—clouds—form near the midplane, and from the densest subcondensations, stars precipitate.

When stars exhaust their nuclear fuel and die, those that are at least as massive as the sun expel much of their matter back into the interstellar medium. Thus, as the galaxy ages, each generation of stars pollutes the medium with heavy elements. As in the water cycle on Earth, precipitation is followed by "evaporation," so that material can be recycled over and over again.

Up in the Air

THINKING OF THE INTERSTELLAR medium as a true atmosphere brings unity to some of the most pressing problems in astrophysics. First and foremost is star formation. Although astronomers have known the basic principles for decades, they still do not grasp exactly what determines when and at what rate stars precipitate from the interstellar medium. Theorists used to explain the creation of stars only in terms of the local conditions within an isolated gas cloud. Now they are considering conditions in the galaxy as a whole.

Not only do these conditions influence star formation, they are influenced by it. What one generation of stars does determines the environment in which subsequent generations are born, live and die. Understanding this feedback—the sway of stars, especially the hottest, rarest, most massive stars, over the large-scale properties of the interstellar medium—is another of the great challenges for researchers. Feedback can be both positive and negative. On the one hand, massive stars can heat and ionize the medium and cause it to bulge out from the midplane. This expansion increases the ambient pressure, compressing the clouds and perhaps triggering their collapse into a new generation of stars. On the other hand, the heating and ionization can also agitate clouds, inhibiting the birth of new stars. When the largest stars blow up, they can even destroy the clouds that gave them birth. In fact, negative feedback could explain why the gravitational collapse of clouds into stars is so inefficient. Typically only a few percent of a cloud's mass becomes stars.

A third conundrum is that star formation often occurs in sporadic but intense bursts. In the Milky Way the competing feedback effects almost balance out, so that stars form at an unhurried pace—just 10 per year on average. In some galaxies, however, such as the "exploding galaxy" M82, positive feedback has gained the upper hand. Starting 20 million to 50 million years ago, star formation in the central parts of M82 began running out of control, proceeding 10 times faster than before. Our galaxy, too, may have had sporadic bursts. How these starbursts occur and what turns them off must be tied to the complex relation between stars and the tenuous atmosphere from which they precipitate.

Finally, astronomers debate how quickly the atmospheric activity is petering out. The majority of stars—those less massive than the sun, which live tens or even hundreds of billions of years—do not contribute to the feedback loops. More and more of the interstellar gas is being locked up into very long lived stars. Eventually all the spare gas in our Milky Way may be exhausted, leaving only stellar dregs behind. How soon this will happen depends on whether the Milky Way is a closed box. Recent observations suggest that the galaxy is still an open system, both gaining and losing mass to its cosmic surroundings. High-velocity clouds of relatively unpolluted hydrogen appear to be raining down from intergalactic space, rejuvenating our

DON DIXON

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

galaxy. Meanwhile the galaxy may be shedding gas in the form of a high-speed wind from its outer atmosphere, much as the sun slowly sheds mass in the solar wind.

Hot and Cold Running Hydrogen

TO TACKLE THESE PROBLEMS, those of us who study the interstellar medium have first had to identify its diverse components. Astronomers carried out the initial step, an analysis of its elemental composition, in the 1950s and 1960s using the spectra of light emitted by bright nebulae, such as the Orion Nebula. In terms of the number of atomic nuclei, hydrogen constitutes 90 percent, helium about 10 percent, and everything else—from lithium to uranium—just a trace, about 0.1 percent.

Because hydrogen is so dominant, the structure of the galaxy's atmosphere depends mainly on what forms the hydrogen takes. Early observations were sensitive primarily to cooler, neutral components. The primary marker of interstellar material is the most famous spectral line of astronomy: the 1,420-megahertz (21-centimeter) line emitted by neutral hydrogen atoms, denoted by astronomers as H I. Beginning in the 1950s, radio astronomers mapped out the distribution of H I within the galaxy. It resides in lumps and filaments with densities of 10 to 100 atoms per cubic centimeter and temperatures near 100 kelvins, embedded in a more diffuse, thinner (roughly 0.1 atom per cubic centimeter) and warmer (a few thousand kelvins) phase. Most of the H I is close to the galactic midplane, forming a gaseous disk about 300 parsecs (1,000 light-years) thick, roughly half the thickness of the main stellar disk you see when you notice the Milky Way in the night sky.

Hydrogen can also come in a molecular form (H_2), which is extremely difficult to detect directly. Much of the information about it has been inferred from high-frequency radio observations of the trace molecule carbon monoxide. Where carbon monoxide exists, so should molecular hydrogen. The molecules appear to be confined to the densest and coldest clouds—the places where starlight, which breaks molecules into their constituent atoms, cannot penetrate. These dense clouds, which are active sites of star formation, are found in a thin layer (100 parsecs thick) at the very bottom of the galactic atmosphere.

Until very recently, hydrogen molecules were seen directly only in places where they were being destroyed—that is, converted to atomic hydrogen—by a nearby star's ultraviolet radiation or wind of outflowing particles. In these environments, H_2 glows at an infrared wavelength of about 2.2 microns. In the past few years, however, orbiting spectrographs, such as the shuttle-based platform called ORFEUS-SPAS and the new Far Ultraviolet Spectroscopic Explorer (FUSE) satellite, have sought molecular hydrogen at ultraviolet wavelengths near 0.1 micron. These instruments look for hydrogen that is backlit by distant stars and quasars: the H_2 leaves telltale absorption lines in the ultraviolet spectra of those objects. The advantage of this approach is that it can detect molecular hydrogen in quiescent regions of the galaxy, far from any star.

To general astonishment, two teams, led respectively by Philipp Richter of the University of Wisconsin and Wolfgang Gringel of the University of Tübingen in Germany, have discovered H_2 not just in the usual places—the high-density clouds located within the galactic disk—but also in low-density areas far outside the disk. This is a bit of a mystery, because high densities are needed to shield the molecules from the ravages of starlight. Perhaps a population of cool clouds extends much farther from the midplane than previously believed.

A third form of hydrogen is a plasma of hydrogen ions. Astronomers used to assume that ionized hydrogen was confined to a few small, isolated locations—the glowing nebulae near luminous stars and the wispy remnants left over from supernovae. Advances in detection technology and the advent of space astronomy have changed that. Two new components of our galaxy's atmosphere have come into view: hot (10^6 kelvins) and warm (10^4 kelvins) ionized hydrogen (H II).

Like the recently detected hydrogen molecules, these H II phases stretch far above the cold H I cloud layer, forming a thick gaseous "halo" around the entire galaxy. "Interstellar" no longer seems an appropriate description for these outermost parts of our galaxy's atmosphere. The hotter phase may extend thousands of parsecs from the midplane and thin out to a density near 10^{-3} ion per cubic centimeter. It is our galaxy's corona, analogous to the extended hot atmosphere of our sun. As in the case of the solar corona, the mere existence of the galactic corona implies an unconventional source of energy to maintain the high temperatures. Supernova shocks and fast stellar winds appear to do the trick. Coexisting with the hot plasma is the warm plasma, which is powered by extreme ultraviolet radiation. The weight of these extended layers increases the gas pressure at the midplane, with significant effects on star formation. Other galaxies appear to have coronas as well. The Chandra X-ray Observatory has recently seen one around the galaxy NGC 4631 [*see photo page 54*].

Blowing Bubbles

HAVING IDENTIFIED these new, more energetic phases of the medium, astronomers have turned to the question of how the diverse components behave and interrelate. Not only does the interstellar medium cycle through stars, it changes from H_2 to H I to H II and from cold to hot and back again. Massive stars are the only known source of energy powerful enough to account for all this activity. A study by Ralf-Jürgen Dettmar of the University of Bochum in Germany found that galaxies with a larger-than-average massive star population seem to have

THE AUTHOR

RONALD J. REYNOLDS bought a 4.25-inch reflecting telescope in sixth grade and used it to take pictures of the moon. But it wasn't until he started his Ph.D. in physics that he took his first astronomy course and began to consider a career in the subject. Today Reynolds is an astronomy professor at the University of Wisconsin–Madison. He has designed and built high-sensitivity spectrometers to study warm ionized gas in the Milky Way galaxy. He is principal investigator for the Wisconsin H-Alpha Mapper, which spent two years mapping hydrogen over the entire northern sky.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

atmospheres that are more extended or puffed up. How the stars wield power over an entire galaxy is somewhat unclear, but astronomers generally pin the blame on the creation of hot ionized gas.

This gas appears to be produced by the high-velocity (100 to 200 kilometers per second) shock waves that expand into the interstellar medium following a supernova. Depending on the density of the gas and strength of the magnetic field in the ambient medium, the spherically expanding shock may clear out a cavity 50 to 100 parsecs in radius—a giant bubble.

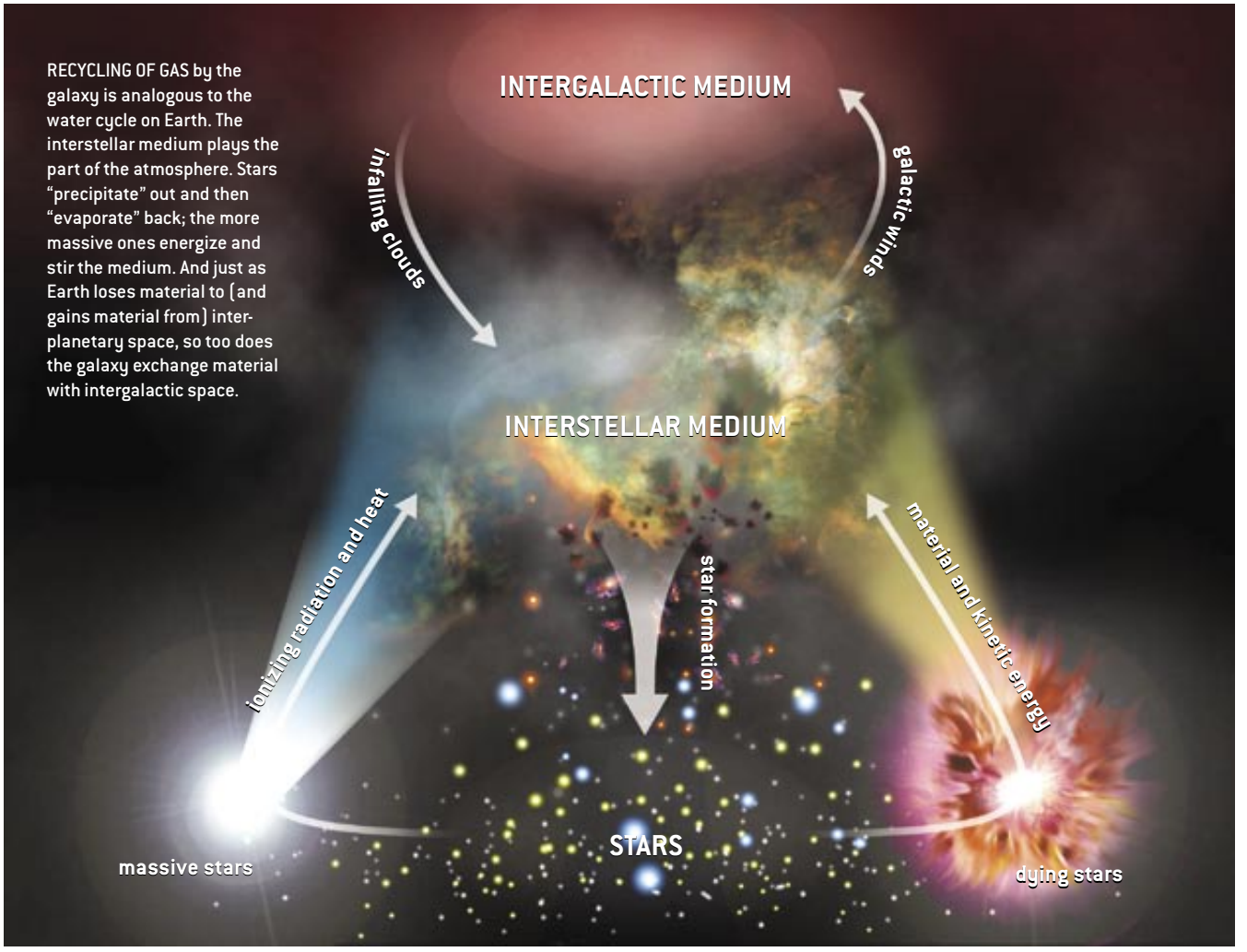
In doing so, the shock accelerates a small fraction of the ions and electrons to near light speed. Known as cosmic rays, these fleet-footed particles are one way that stellar death feeds back (both positively and negatively) into stellar birth. Cosmic rays raise the pressure of the interstellar medium; higher pressures, in turn, compress the dense molecular clouds and increase the chance that they will collapse into stars. By ionizing some of the hydrogen, the cosmic rays also drive chemical reactions that synthesize complex molecules, some of which are the building blocks of life as we know it. And because the ions attach themselves to magnetic field lines, they trap the field within the clouds, which slows the rate of cloud collapse into stars.

If hot bubbles are created frequently enough, they could interconnect in a vast froth. This idea was first advanced in the 1970s by Barham Smith and Donald Cox of the University of Wisconsin–Madison. A couple of years later Christopher F. McKee of the University of California at Berkeley and Jeremiah P. Ostriker of Princeton University argued that the hot phase should occupy 55 to 75 percent of interstellar space. Cooler neutral phases would be confined to isolated clouds within this ionized matrix—essentially the inverse of the traditional picture, in which the neutral gas dominates and the ionized gas is confined to small pockets.

Recent observations seem to support this upending of conventional wisdom. The nearby spiral galaxy M101, for example, has a circular disk of atomic hydrogen gas riddled with holes—presumably blown by massive stars. The interstellar medium of another galaxy, seven billion light-years distant, also looks like Swiss cheese. But the amount of hot gas and its influence on the structure of galactic atmospheres still occasion much debate.

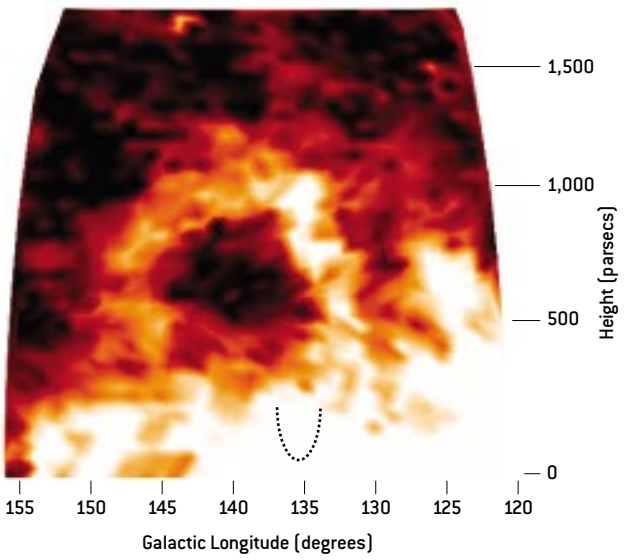
Chimneys and Fountains

THE SUN ITSELF APPEARS to be located within a hot bub-

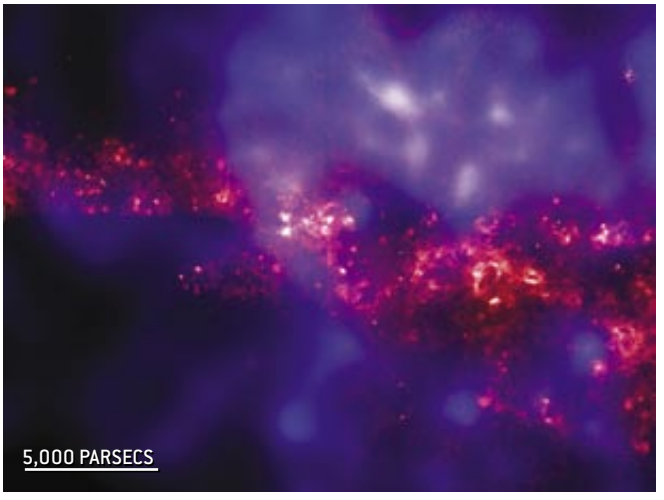


DON DIXON

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.



ARCHING OVER THE DISK of our galaxy is an enormous loop of warm ionized hydrogen. It is located just above the W4 Chimney (*dotted line*), shown on page 40. The same star cluster may account for both of these structures.



ENVELOPING THE DISK of the galaxy NGC 4631 is a hot plasma (*blue and purple*), seen by the Chandra X-ray Observatory. The Ultraviolet Imaging Telescope revealed massive stars within the disk (*orange*).

ble, which has revealed itself in x-rays emitted by highly ionized trace ions such as oxygen. Called the Local Bubble, this region of hot gas was apparently created by a nearby supernova about one million years ago.

An even more spectacular example lies 450 parsecs from the sun in the direction of the constellations Orion and Eridanus. It was the subject of a recent study by Carl Heiles of the University of California at Berkeley and his colleagues. The Orion-Eridanus Bubble was formed by a star cluster in the constellation Orion. The cluster is of an elite type called an OB association—a bundle of the hottest and most massive stars, the O- and B-type stars, which are 20 to 60 times heavier than the sun (a G-type star) and 10^3 to 10^5 times brighter. The spectacular deaths of these short-lived stars in supernovae over the past 10 million years have swept the ambient gas into a shell-like skin around the outer boundary of the bubble. In visible light the shell appears as a faint lacework of ionized loops and filaments. The million-degree gas that fills its interior gives off a diffuse glow of x-rays.

The entire area is a veritable thunderstorm of star formation, with no sign of letting up. Stars continue to precipitate from the giant molecular cloud out of which the OB association emerged. One of the newest O stars, theta^1 C Orionis, is ionizing a small piece of the cloud—producing the Orion Nebula. In time, however, supernovae and ionizing radiation will completely disrupt the molecular cloud and dissociate its molecules. The molecular hydrogen will turn back into atomic and ionized hydrogen, and star formation will cease. Because the violent conversion process will increase the pressure in the interstellar medium, the demise of this molecular cloud may mean the birth of stars elsewhere in the galaxy.

Galactic bubbles should buoyantly lift off from the galactic midplane, like a thermal rising above the heated ground on Earth. Numerical calculations, such as those recently made by Mordecai-Mark MacLow of the American Museum of Natural History in New York City and his colleagues, suggest that bubbles can ascend all the way up into the halo of the galaxy. The result is a cosmic chimney through which hot gas spewed by supernovae near the midplane can vent to the galaxy's upper atmosphere. There the gas will cool and rain back onto the galactic disk. In this case, the superbubble and chimney become a galactic-scale fountain.

Such fountains could perhaps be the source of the hot galactic corona and even the galaxy's magnetic field. According to calculations by Katia M. Ferrière of the Midi-Pyrénées Observatory in France, the combination of the updraft and the rotation of the galactic disk would act as a dynamo, much as motions deep inside the sun and Earth generate magnetic fields.

To be sure, observers have yet to prove the pervasive nature of the hot phase or the presence of fountains. The Orion-Eridanus bubble extends 400 parsecs from the midplane, and a similar superbubble in Cassiopeia rises 230 parsecs, but both have another 1,000 to 2,000 parsecs to go to reach the galactic corona. Magnetic fields and cooler, denser ionized gas could make it difficult or impossible for superbubbles to break out into the halo. But then, where did the hot corona come from? No plausible alternative is known.

Getting Warm

THE WARM (10^4 KELVINS) plasma is as mysterious as its hot relative. Indeed, in the traditional picture of the interstellar medium, the widespread presence of warm ionized gas is simply impossible. Such gas should be limited to very small regions of space—the emission nebulae, such as the Orion Nebula, that immediately surround ultramassive stars. These stars account for only one star in five million, and most of the interstellar gas (the atomic and molecular hydrogen) is opaque to their pho-

RONALD J. REYNOLDS [*top*]; NASA AND DANIEL WANG *University of Massachusetts at Amherst* [*x-ray imaging*] AND NASA/UIT [*ultraviolet imaging*] [*middle*]

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

tons. So the bulk of the galaxy should be unaffected.

Yet warm ionized gas is spread throughout interstellar space. One recent survey, known as WHAM, finds it even in the galactic halo, very far from the nearest O stars. Ionized gas is similarly widespread in other galaxies. This is a huge mystery. How did the ionizing photons manage to stray so far from their stars?

Bubbles may be the answer. If supernovae have hollowed out significant parts of the interstellar medium, ionizing photons may be able to travel large distances before being absorbed by neutral hydrogen. The Orion OB association provides an excellent example of how this could work. The O stars sit in an immense cavity carved out by earlier supernovae. Their photons now travel freely across the cavity, striking the distant bubble wall and making it glow. If galactic fountains or chimneys do indeed stretch up into the galactic halo, they could explain not only the hot corona but also the pervasiveness of warm ionized gas.

A new WHAM image of the Cassiopeia superbubble reveals a possible clue: a loop of warm gas arching far above the bubble, some 1,200 parsecs from the midplane. The outline of this loop bears a loose resemblance to a chimney, except that it has not (yet) broken out into the Milky Way's outer halo. The amount of energy required to produce this gigantic structure is enormous—more than that available from the stars in the cluster that formed the bubble. Moreover, the time required to create it is 10 times the age of the cluster. So the loop may be a multigenerational project, created by a series of distinct bursts of star formation predating the cluster we see today. Each burst reenergized and expanded the bubble created by the preceding burst.

Round and Round

THAT LARGE REGIONS of the galaxy can be influenced by the formation of massive stars in a few localized regions seems to require that star formation somehow be coordinated over long periods of time. It may all begin with a single O-type star or a cluster of such stars in a giant molecular cloud. The stellar radiation, winds and explosions carve a modest cavity out of the surrounding interstellar medium. In the process the parent cloud is probably destroyed. Perchance this disturbance triggers star formation in a nearby cloud, and so on, until the interstellar medium in this corner of the galaxy begins to resemble Swiss cheese. The bubbles then begin to overlap, coalescing into a superbubble. The energy from more and more O-type stars feeds this expanding superbubble until its natural buoyancy stretches it from the midplane up toward the halo, forming a chimney.

that large regions of the galaxy can be influenced by the formation of massive stars in a few localized regions seems to require that star formation somehow be coordinated over long periods of time.

The superbubble is now a pathway for hot interior gas to spread into the upper reaches of the galactic atmosphere, producing a widespread corona. Now, far from its source of energy, the coronal gas slowly begins to cool and condense into clouds. These clouds fall back to the galaxy's midplane, completing the fountainlike cycle and replenishing the galactic disk with cool clouds from which star formation may begin anew.

Even though the principal components and processes of our galaxy's atmosphere seem to have been identified, the details remain uncertain. Progress will be made as astronomers continue to study how the medium is cycled through stars, through the different phases of the medium, and between the disk and the halo. Observations of other galaxies give astronomers a bird's-eye view of the interstellar goings-on. SA

MORE TO EXPLORE

Ionizing the Galaxy. Ronald J. Reynolds in *Science,* Vol. 277, pages 1446–1447; September 5, 1997.

Far Ultraviolet Spectroscopic Explorer Observations of O VI Absorption in the Galactic Halo. Blair D. Savage et al. in *Astrophysical Journal Letters,* Vol. 538, No. 1, pages L27–L30; July 20, 2000. Preprint available at **arXiv.org/abs/astro-ph/0005045**

Gas in Galaxies. Joss Bland-Hawthorn and Ronald J. Reynolds in *Encyclopaedia of Astronomy & Astrophysics.* MacMillan and Institute of Physics Publishing, 2000. Preprint available at **arXiv.org/abs/astro-ph/0006058**

Detection of a Large Arc of Ionized Hydrogen Far Above the CAS OB6 Association: A Superbubble Blowout into the Galactic Halo? Ronald J. Reynolds, N. C. Sterling and L. Matthew Haffner in *Astrophysical Journal Letters,* Vol. 558, No. 2, pages L101–L104; September 10, 2001. Preprint available at **arXiv.org/abs/astro-ph/0108046**

The Interstellar Environment of Our Galaxy. K. M. Ferrière in *Reviews of Modern Physics,* Vol. 73, No. 4 (in press). Preprint available at **arXiv.org/abs/astro-ph/0106359**

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

A PICTURE LIKE THIS could not have been drawn with any confidence a decade ago, because no one had yet figured out what causes gamma-ray bursts—flashes of high-energy radiation that light up the sky a couple of times a day. Now astronomers think of them as the ultimate stellar swan song. A black hole, created by the implosion of a giant star, sucks in debris and sprays out some of it. A series of shock waves emits radiation.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

The Brightest Explosions in the Universe

Every time a gamma-ray burst goes off, a black hole is born

originally published in December 2002

By Neil Gehrels, Luigi Piro and Peter J. T. Leonard

Early in the morning of January 23, 1999, a robotic telescope in New Mexico picked up a faint flash of light in the constellation Corona Borealis. Though just barely visible through binoculars, it turned out to be the most brilliant explosion ever witnessed by humanity. We could see it nine billion light-years away, more than halfway across the observable universe. If the event had instead taken place a few thousand light-years away, it would have been as bright as the midday sun, and it would have dosed Earth with enough radiation to kill off nearly every living thing.

The flash was another of the famous gamma-ray bursts, which in recent decades have been one of astronomy's most intriguing mysteries. The first sighting of a gamma-ray burst (GRB) came on July 2, 1967, from military satellites watching for nuclear tests in space. These cosmic explosions proved to be rather different from the man-made explosions that the

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

satellites were designed to detect. For most of the 35 years since then, each new burst merely heightened the puzzlement. Whenever researchers thought they had the explanation, the evidence sent them back to square one.

The monumental discoveries of the past several years have brought astronomers closer to a definitive answer. Before 1997, most of what we knew about GRBs was based on observations from the Burst and Transient Source Experiment (BATSE) onboard the Compton Gamma Ray Observatory. BATSE revealed that two or three GRBs occur somewhere in the observable universe on a typical day. They outshine everything else in the gamma-ray sky. Although each is unique, the bursts fall into one of two rough categories. Bursts that last less than two seconds are "short," and those that last longer—the majority—are "long." The two categories differ spectroscopically, with short bursts having relatively more high-energy gamma rays than long bursts do. The January 1999 burst emitted gamma rays for a minute and a half.

Arguably the most important result from BATSE concerned the distribution of the bursts. They occur isotropically—that is, they are spread evenly over the entire sky. This finding cast doubt on the prevailing wisdom, which held that bursts came from sources within the Milky Way; if they did, the shape of our galaxy, or Earth's off-center position within it, should have caused them to bunch up in certain areas of the sky. The uniform distribution led most astronomers to conclude that the instruments were picking up some kind of event happening throughout the universe. Unfortunately, gamma rays alone did not provide enough information to settle the question for sure. Researchers would need to detect radiation from the bursts at other wavelengths. Visible light, for example, could reveal the galaxies in which the bursts took place, allowing their distances to be measured. Attempts were made to detect these burst counterparts, but they proved fruitless.

...gamma rays alone did not provide enough information to settle the question for sure. Researchers would need to detect radiation from the bursts at other wavelengths.

A Burst of Progress

THE FIELD TOOK a leap forward in 1996 with the advent of the x-ray spacecraft BeppoSAX, built and operated by the Italian Space Agency with the participation of the Netherlands Space Agency. BeppoSAX was the first satellite to localize GRBs precisely and to discover their x-ray "afterglows." The afterglow appears when the gamma-ray signal disappears. It persists for days to months, diminishing with time and degrading from x-rays into less potent radiation, including visible light and radio waves. Although BeppoSAX detected afterglows for only long bursts—no counterparts of short bursts have yet been identified—it made follow-up observations possible at last. Given the positional information from BeppoSAX, optical and radio telescopes were able to identify the galaxies in which the GRBs took place. Nearly all lie billions of light-years away, meaning that the bursts must be enormously powerful [see "Gamma-Ray Bursts," by Gerald J. Fishman and Dieter H. Hartmann; SCIENTIFIC AMERICAN, July 1997]. Extreme energies, in turn, call for extreme causes, and researchers began to associate GRBs with the most extreme objects they knew of: black holes.

Among the first GRBs pinpointed by BeppoSAX was GRB970508, so named

Overview/*Gamma-Ray Bursts*

- For three decades, the study of gamma-ray bursts was stuck in first gear—astronomers couldn't settle on even a sketchy picture of what sets off these cosmic fireworks.
- Over the past five years, however, observations have revealed that bursts are the birth throes of black holes. Most of the holes are probably created when a massive star collapses, releasing a pulse of radiation that can be seen billions of light-years away.
- Now the research has shifted into second gear—fleshing out the theory and probing subtle riddles, especially the bursts' incredible diversity.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

because it occurred on May 8, 1997. Radio observations of its afterglow provided an essential clue. The glow varied erratically by roughly a factor of two during the first three weeks, after which it stabilized and then began to diminish. The large variations probably had nothing to do with the burst source itself; rather they involved the propagation of the afterglow light through space. Just as Earth's atmosphere causes visible starlight to twinkle, interstellar plasma causes radio waves to scintillate. For this process to be visible, the source must be so small and far away that it appears to us as a mere point. Planets do not twinkle, because, being fairly nearby, they look like disks, not points.

Therefore, if GRB970508 was scintillating at radio wavelengths and then stopped, its source must have grown from a mere point to a discernible disk. "Discernible" in this case means a few light-weeks across. To reach that size, the source must have been expanding at a considerable rate—close to the speed of light.

The BeppoSAX and follow-up observations have transformed astronomers' view of GRBs. The old concept of a sudden release of energy concentrated in a few brief seconds has been discarded. Indeed, even the term "afterglow" is now recognized as misleading: the energy radiated during both phases is comparable. The spectrum of the afterglow is characteristic of electrons moving in a magnetic field at or very close to the speed of light.

The January 1999 burst (GRB990123) was instrumental in demonstrating the immense power of the bursts. If the burst radiated its energy equally in all directions, it must have had a luminosity of a few times 10^{45} watts, which is 10^{19} times as bright as our sun. Although the other well-known type of cosmic cataclysm, a supernova explosion, releases almost as much energy, most of that energy escapes as neutrinos, and the remainder leaks out more gradually than in a GRB. Consequently, the luminosity of a supernova at any given moment is a tiny fraction of that of a GRB. Even quasars, which are famously brilliant, give off only about 10^{40} watts.

If the burst beamed its energy in particular directions rather than in all directions, however, the luminosity estimate would be lower. Evidence for beaming comes from the way the afterglow of GRB990123, among others, dimmed over time. Two days into the burst, the rate of dimming increased suddenly, which would happen naturally if the observed radiation came from a narrow jet of material moving at close to the speed of light. Because of a relativistic effect, the observer sees more and more of the jet as it slows down. At some point, there is no more to be seen, and the apparent brightness begins to fall off more rapidly [*see illustration on next page*]. For GRB990123 and several other bursts, the inferred jet-opening angle is a few degrees. Only if the jet is aimed along our line of sight do we see the burst. This beaming effect reduces the overall energy emitted by the burst approximately in proportion to the square of the jet angle. For example, if the jet subtends 10 degrees, it covers about one 500th of the sky, so the energy requirement goes down by a factor of 500; moreover, for every GRB that is observed, another 499 GRBs go unseen. Even after taking beaming into account, however, the luminosity of GRB990123 was still an impressive 10^{43} watts.

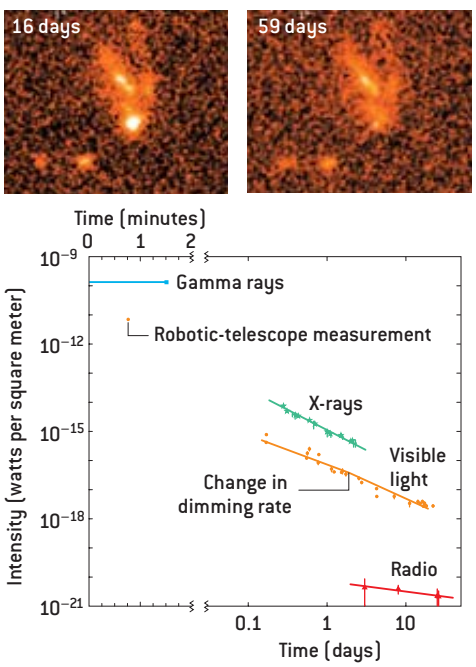
GRB-Supernova Connection

ONE OF THE MOST interesting discoveries has been the connection between GRBs and supernovae. When telescopes went to look at GRB980425, they also found a supernova, designated SN1998-bw, that had exploded at about the same time as the burst. The probability of a chance coincidence was one in 10,000 [see "Bright Lights, Big Mystery," by George Musser; News and Analysis, Scientific American, August 1998].

A link between GRBs and supernovae has also been suggested by the detection of iron in the x-ray spectra of several bursts. Iron atoms are known to be synthesized and dumped into interstellar space by supernova explosions. If these atoms are stripped of their electrons and later hook up with them again, they give off light at distinctive wavelengths, referred to as emission lines. Early, marginal detections of such lines by BeppoSAX and the Japanese x-ray satellite ASCA in 1997 have been followed by more solid measurements. Notably, NASA's Chandra X-ray Observatory detected iron lines in GRB991216, which yielded a direct dis-

FADING AWAY

BRIGHTEST GAMMA-RAY BURST yet recorded went off on January 23, 1999. Telescopes tracked its brightness in gamma rays (*blue in graph*), x-rays (*green*), visible light (*orange*) and radio waves (*red*). At one point, the rate of dimming changed abruptly—a telltale sign that the radiation was coming from narrow jets of high-speed material. About two weeks into the burst, after the visible light had dimmed by a factor of four million, the Hubble Space Telescope took a picture and found a severely distorted galaxy. Such galaxies typically have high rates of star formation. If bursts are the explosions of young stars, they should occur in just such a place.

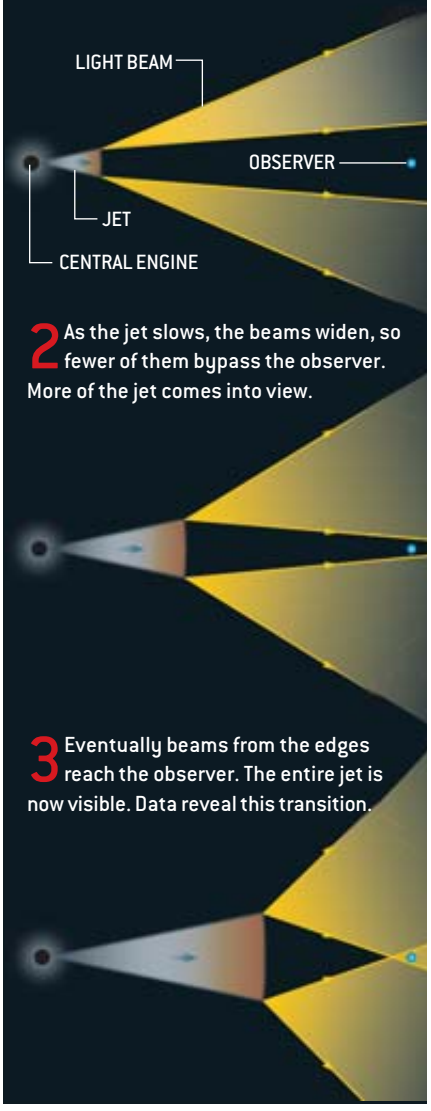


CORNELIA BLIK (*graph*); ANDREW FRUCHTER *Space Telescope Science Institute* AND NASA

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

BEAM LINES

RELATIVITY PLAYS TRICKS on observers' view of jets from gamma-ray bursts.



tance measurement of the GRB. The figure agreed with the estimated distance of the burst's host galaxy.

Additional observations further support the connection between GRBs and supernovae. An iron-absorption feature appeared in the x-ray spectrum of GRB-990705. In the shell of gas around another burst, GRB011211, the European Space Agency's X-ray Multi-Mirror satellite found evidence of emission lines from silicon, sulfur, argon and other elements commonly released by supernovae.

Although researchers still debate the matter, a growing school of thought holds that the same object can produce, in some cases, both a burst and a supernova. Because GRBs are much rarer than supernovae—every day a couple of GRBs go off somewhere in the universe, as opposed to hundreds of thousands of supernovae—not every supernova can be associated with a burst. But some might be. One version of this idea is that supernova explosions occasionally squirt out jets of material, leading to a GRB. In most of these cases, astronomers would see either a supernova or a GRB, but not both. If the jets were pointed toward Earth, light from the burst would swamp light from the supernova; if the jets were aimed in another direction, only the supernova would be visible. In some cases, however, the jet would be pointed just slightly away from our line of sight, letting observers see both. This slight misalignment would explain GRB980425.

Whereas this hypothesis supposes that most or all GRBs might be related to supernovae, a slightly different scenario attributes only a subset of GRBs to supernovae. Roughly 90 of the bursts seen by BATSE form a distinct class of their own, defined by ultralow luminosities and long spectral lags, meaning that the high- and low-energy gamma-ray pulses arrive several seconds apart. No one knows why the pulses are out of sync. But whatever the reason, these strange GRBs occur at the same rate as a certain type of supernova, called Type Ib/c, which occurs when the core of a massive star implodes.

GREAT BALLS OF FIRE

EVEN LEAVING ASIDE the question of how the energy in GRBs might be generated, their sheer brilliance poses a paradox. Rapid brightness variations suggest that the emission originates in a small region: a luminosity of 10^{19} suns comes from a volume the size of one sun. With so much radiation emanating from such a compact space, the photons must be so densely packed that they should interact and prevent one another from escaping. The situation is like a crowd of people who are running for the exit in such a panic that that nobody can get out. But if the gamma rays are unable to escape, how can we be seeing GRBs?

The resolution of this conundrum, developed over the past several years, is that the gammas are not emitted immediately. Instead the initial energy release of the explosion is stored in the kinetic energy of a shell of particles—a fireball—moving at close to the speed of light. The particles include photons as well as electrons and their antimatter counterpart, positrons. This fireball expands to a diameter of 10 billion to 100 billion kilometers, by which point the photon density has dropped enough for the gamma rays to escape unhindered. The fireball then converts some of its kinetic energy into electromagnetic radiation, yielding a GRB.

The initial gamma-ray emission is most likely the result of internal shock waves within the expanding fireball. Those shocks are set up when faster blobs in the expanding material overtake slower blobs. Because the fireball is expanding so close to the speed of light, the timescale witnessed by an external observer is vastly compressed, according to the principles of relativity. So the observer sees a burst of gamma rays that lasts only a few seconds, even if it took a day to produce. The fireball continues to expand, and eventu-

THE AUTHORS

NEIL GEHRELS, LUIGI PIRO and PETER J. T. LEONARD bring both observation and theory to the study of gamma-ray bursts. Gehrels and Piro are primarily observers—the lead scientists, respectively, of the Compton Gamma Ray Observatory and the BeppoSAX satellite. Leonard is a theorist, and like most theorists, he used to think it unlikely that the bursts were bright enough to be seen across the vastness of intergalactic space. "I have to admit that the GRBs really had me fooled," he says. Gehrels is head of the Gamma Ray, Cosmic Ray and Gravitational Wave Astrophysics Branch of the Laboratory for High Energy Astrophysics at the NASA Goddard Space Flight Center. Piro is a member of the Institute of Space Astrophysics and Cosmic Physics of the CNR in Rome. Leonard works for Science Systems and Applications, Inc., in support of missions at Goddard.

JUAN VELASCO

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

ally it encounters and sweeps up surrounding gas. Another shock wave forms, this time at the boundary between the fireball and the external medium, and persists as the fireball slows down. This external shock nicely accounts for the GRB afterglow emission and the gradual degradation of this emission from gamma rays to x-rays to visible light and, finally, to radio waves.

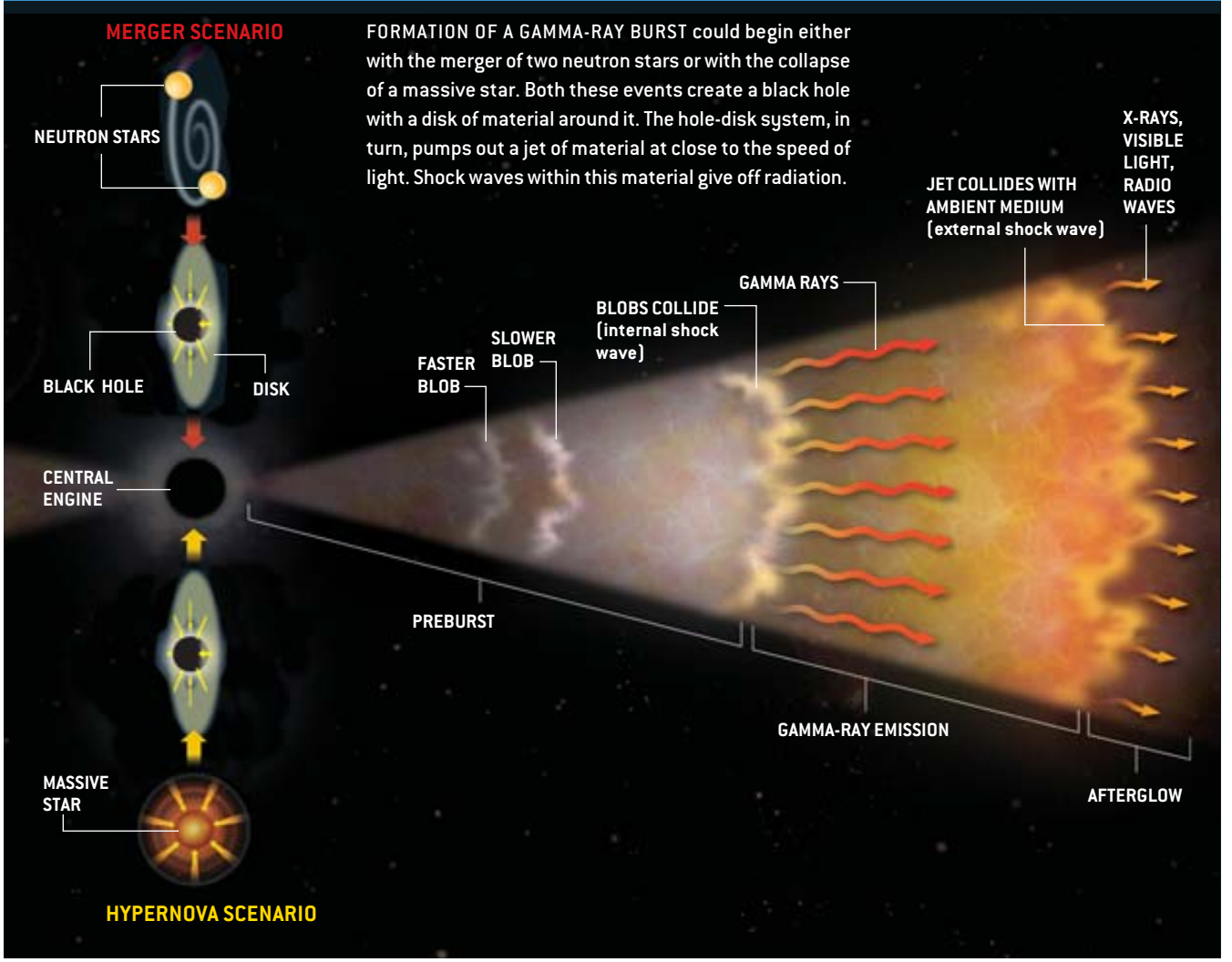
Although the fireball can transform the explosive energy into the observed radiation, what generates the energy to begin with? That is a separate problem, and astronomers have yet to reach a consensus. One family of models, referred to as hypernovae or collapsars, involves stars born with masses greater than about 20 to 30 times that of our sun. Simulations show that the central core of such a star eventually collapses to form a rapidly rotating black hole encircled by a disk of leftover material.

A second family of models invokes binary systems that consist of two compact objects, such as a pair of neutron stars (which are ultradense stellar corpses) or a neutron star paired with a black hole. The two objects spiral toward each other and merge into one. Just as in the hypernova scenario, the result is the formation of a single black hole surrounded by a disk.

Many celestial phenomena involve a hole-disk combination. What distinguishes this particular type of system is the sheer mass of the disk (which allows for a gargantuan release of energy) and the lack of a companion star to resupply the disk (which means that the energy release is a one-shot event). The black hole and disk have two large reservoirs of energy: the gravitational energy of the disk and the rotational energy of the hole. Exactly how these would be converted into gamma radiation is not fully understood. It is possible that a magnetic field, 10^{15} times more intense than Earth's magnetic field, builds up during the formation of the disk. In so doing, it heats the disk to such high temperatures that it unleashes a fireball of gamma rays and plasma. The fireball is funneled into a pair of narrow jets that flow out along the rotational axis.

Because the GRB emission is equally well explained by both hypernovae and compact-object mergers, some other qualities of the bursts are needed to decide be-

BURSTING OUT



JUAN VELASCO

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

tween these two scenarios. The association of GRBs with supernovae, for example, is a point in favor of hypernovae, which, after all, are essentially large supernovae. Furthermore, GRBs are usually found just where hypernovae would be expected to occur—namely, in areas of recent star formation within galaxies. A massive star blows up fairly soon (a few million years) after it is born, so its deathbed is close to its birthplace. In contrast, compact-star coalescence takes much longer (billions of years), and in the meantime the objects will drift all over the galaxy. If compact objects were the culprit, GRBs should not occur preferentially in star-forming regions.

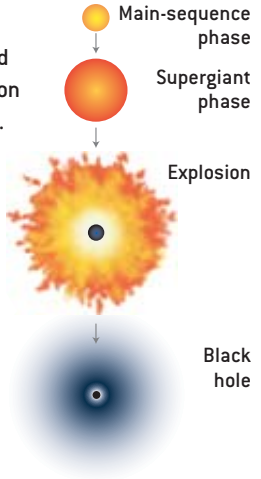
Although hypernovae probably explain most GRBs, compact-star coalescence could still have a place in the big picture. This mechanism may account for the poorly understood short-duration GRBs. Moreover, additional models for GRBs are still in the running. One scenario produces the fireball via the extraction of energy from an electrically charged black hole. This model suggests that both the immediate and the afterglow emissions are consequences of the fireball sweeping up the external medium. Astronomers have come a long way in understanding gamma-ray bursts, but they still do not know precisely what causes these explosions, and they know little about the rich variety and subclasses of bursts.

All these recent findings have shown that the field has the potential for answering some of the most fundamental questions in astronomy: How do stars end their lives? How and where are black holes formed? What is the nature of jet outflows from collapsed objects?

The Destinies of Massive Stars

STARS SPEND MOST OF THEIR LIVES in the relatively unexciting main-sequence evolutionary phase, during which they casually convert hydrogen into helium in their cores via nuclear fusion. Our sun is in this phase. According to basic stellar theory, stars more massive than the sun shine more brightly and burn their fuel more quickly. A star 20 times as massive as the sun can keep going for only a thousandth as long.

As the hydrogen in the core of a star runs out, the core contracts, heats up and starts to fuse heavier elements, such as helium, oxygen and carbon. The star thus evolves into a giant and then, if sufficiently massive, a supergiant star. If the initial mass of the star is at least eight times that of the sun, the star successively fuses heavier and heavier elements in its interior until it produces iron. Iron fusion does not release energy—on the contrary, it uses up energy. So the star suddenly finds itself without any useful fuel.



The result is a sudden and catastrophic collapse. The core is thought to turn into a neutron star, a stellar corpse that packs at least 40 percent more mass than the sun into a ball with a radius of only 10 kilometers. The remainder of the star is violently ejected into space in a powerful supernova explosion.

There is a limit to how massive a neutron star can be—namely, two to three times as massive as the sun. If it is any heavier, theory predicts it will collapse into a black hole. It can be pushed over the line if enough matter falls onto it. It is also possible that a black hole can be formed directly during the collapse. Stars born with masses exceeding roughly 20 solar masses may be destined to become black holes. The creation of these holes provides a natural explanation for gamma-ray bursts. *—N.G., L.P. and P.J.T.L.*

Blasts from the Past

ONE OUTSTANDING question concerns the dark, or "ghost," GRBs. Of the roughly 30 GRBs that have been localized and studied at wavelengths other than gamma rays, about 90 percent have been seen in x-rays. In contrast, only about 50 percent have been seen in visible light. Why do some bursts fail to shine in visible light?

One explanation is that these GRBs lie in regions of star formation, which tend to be filled with dust. Dust would block visible light but not x-rays. Another intriguing possibility is that the ghosts are GRBs that happen to be very far away. The relevant wavelengths of light produced by the burst would be absorbed by intergalactic gas. To test this hypothesis, measurement of the distance via x-ray spectra will be crucial. A third possibility is that ghosts are optically faint by nature. Currently the evidence favors the dust explanation. High-sensitivity optical and radio investigations have identified the probable host galaxies of two dark GRBs, and each lies at a fairly moderate distance.

Another mystery concerns a class of events known as the x-ray-rich GRBs, or simply the x-ray flashes. Discovered by BeppoSAX and later confirmed by reanalysis of BATSE data, these bursts are now known to represent 20 to 30 percent of GRBs. They give off more x-radiation than gamma radiation; indeed, extreme cases exhibit no detectable gamma radiation at all.

One explanation is that the fireball is loaded with a relatively large amount of baryonic matter such as protons, making for a "dirty fireball." These particles increase the inertia of the fireball, so that it moves more slowly and is less able to boost photons into the gamma-ray range. Alternatively, the x-ray flashes might come from very distant galaxies—even more distant than the galaxies proposed to explain the ghost GRBs. Cosmic expansion would then shift the gamma rays into the x-ray range, and intergalactic gas would block any visible afterglow. In fact, none of these x-ray flashes has a detectable visible-light counterpart, a finding that is consistent with this scenario. If either x-ray flashes or ghost GRBs are

CORNELIA BLIK

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.

Classes of Gamma-Ray Bursts

BURST CLASS (SUBCLASS)	PERCENTAGE OF ALL BURSTS	TYPICAL DURATION OF INITIAL EMISSION (SECONDS)	INITIAL GAMMA-RAY EMISSION	AFTERGLOW X-RAY EMISSION	AFTERGLOW VISIBLE EMISSION	HYPOTHETICAL CENTRAL ENGINE	EXPLANATION FOR PECULIAR PROPERTIES
Long (normal)	25	20	✓	✓	✓	Energetic explosion of massive star	Not applicable
Long (ghosts or dark)	30	20	✓	✓	✗	Energetic explosion of massive star	Extremely distant, obscured by dust, or intrinsically faint
Long (x-ray-rich or x-ray flashes)	25	30	✗	✓	✗	Energetic explosion of massive star	Extremely distant or weighed down by extra particles
Short	20	0.3	✓	?	?	Merger of pair of compact objects	Does not occur in a star-forming region, so ambient gas is less dense and external shocks are weaker

located in extremely distant galaxies, they could illuminate an era in cosmic history that is otherwise almost invisible.

The next step for GRB astronomy is to flesh out the data on burst, afterglow and host-galaxy characteristics. Observers need to measure many hundreds of bursts of all varieties: long and short, bright and faint, bursts that are mostly gamma rays, bursts that are mostly x-rays, bursts with visible-light afterglows and those without. Currently astronomers are obtaining burst positions from the second High Energy Transient Explorer satellite, launched in October 2000, and the Interplanetary Network, a series of small gamma-ray detectors piggybacking on planetary spacecraft. The Swift mission, scheduled for launch next fall, will offer multiwavelength observations of hundreds of GRBs and their afterglows. On discovering a GRB, the gamma-ray instrument will trigger automatic onboard x-ray and optical observations. A rapid response will determine whether the GRB has an x-ray or visible afterglow. The mission will be sensitive to short-duration bursts, which have barely been studied so far.

Another goal is to probe extreme gamma-ray energies. GRB940217, for example, emitted high-energy gamma rays for more than an hour after the burst, as observed by the Energetic Gamma Ray Experiment Telescope instrument on the Compton Gamma Ray Observatory. Astronomers do not understand how such extensive and energetic afterglows can be produced. The Italian Space Agency's AGILE satellite, scheduled for launch in 2004, will observe GRBs at these high energies. The supersensitive Gamma-Ray Large Area Space Telescope mission, expected to launch in 2006, will also be key for studying this puzzling phenomenon.

Other missions, though not designed solely for GRB discovery, will also contribute. The International Gamma-Ray Astrophysics Laboratory, launched on October 17, is expected to detect 10 to 20 GRBs a year. The Energetic X-ray Imaging Survey Telescope, planned for launch a decade from now, will have a sensitive gamma-ray instrument capable of detecting thousands of GRBs.

The field has just experienced a series of breakthrough years, with the discovery that GRBs are immense explosions occurring throughout the universe. Bursts provide us with an exciting opportunity to study new regimes of physics and to learn what the universe was like at the earliest epochs of star formation. Space- and ground-based observations over the coming years should allow us to uncover the detailed nature of these most remarkable beasts. Astronomers can no longer talk of bursts as utter mysteries, but that does not mean the puzzle is completely solved. SA

MORE TO EXPLORE

Gamma-Ray Bursts of Doom. Peter J. T. Leonard and Jerry T. Bonnell in *Sky & Telescope,* Vol. 95, No. 2, pages 28–34; February 1998.

Observation of X-ray Lines from a Gamma-Ray Burst (GRB991216): Evidence of Moving Ejecta from the Progenitor. Luigi Piro et al. in *Science,* Vol. 290, pages 955–958; November 3, 2000. Preprint available at **arXiv.org/abs/astro-ph/0011337**

Gamma-Ray Bursts: Accumulating Afterglow Implications, Progenitor Clues, and Prospects. Peter Mészáros in *Science,* Vol. 291, pages 79–84; January 5, 2001. **arXiv.org/abs/astro-ph/0102255**

Blinded by the Light. Stan Woosley in *Nature,* Vol. 414, pages 853–854; December 20, 2001.

The Biggest Bangs: The Mystery of Gamma-Ray Bursts, the Most Violent Explosions in the Universe. Jonathan I. Katz. Oxford University Press, 2002.

Flash! The Hunt for the Biggest Explosions in the Universe. Govert Schilling. Cambridge University Press, 2002.

COPYRIGHT 2005 SCIENTIFIC AMERICAN, INC.