Introduction to Cosmology

Callen Hogan

Introduction to Cosmology

Introduction to Cosmology

Callen Hogan

Published by White Word Publications,
5 Penn Plaza,
19th Floor,
New York, NY 10001, USA

Introduction to Cosmology
Callen Hogan

International Standard Book Number: 978-1-9789-7337-4

Cataloging-in-Publication Data

Introduction to cosmology / Callen Hogan.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-9789-7337-4
1. Cosmology. 2. Astronomy. I. Hogan, Callen.

QB981 .I58 2021
523.1--dc23

TABLE OF CONTENTS

PREFACE

The branch of astronomy, which deals with the origin and evolution of the universe is known as cosmology. There are three major divisions within cosmology, namely, physical cosmology, religious cosmology and philosophical cosmology. Physical cosmology focuses on the physical origins along with the evolution of the universe. The basis for religious cosmology is a body of different mythological, religious and esoteric literature. Philosophical cosmology is dedicated to the philosophical contemplation of the universe as a totality. Since the scope of cosmology extends to disciplines outside of science, theories in cosmology can be either scientific or non-scientific in nature. This book attempts to understand the multiple branches that fall under the discipline of cosmology and how such concepts have practical applications. The topics included in this book on cosmology are of utmost significance and bound to provide incredible insights to readers. It will serve as a valuable source of reference for graduate and postgraduate students.

To facilitate a deeper understanding of the contents of this book a short introduction of every chapter is written below:

Chapter 1- Cosmology is the branch of astronomy which deals with the study of origin and evolution of universe. It is divided into physical cosmology, philosophical cosmology and religious cosmology. The Big Bang theory forms the basis of origin of universe. This is an introductory chapter which will briefly introduce all the significant aspects of cosmology.

Chapter 2- There are various models in cosmology that deal with the study of evolution and cur-rent aspects of cosmology. It includes de Sitter Cosmological model, Einstein-de Sitter model, concordance model. This chapter has been carefully written to provide an easy understanding of these cosmological models.

Chapter 3- Dark matter is any substance that uses gravity to interact with visible matter such as stars and planets. Dark energy is an unknown form of energy that permeates through space which accelerates the expansion of the universe. The topics elaborated in this chapter will help in gaining a better perspective about dark matter and dark energy.

Chapter 4- Black hole is a spacetime region that exhibits a strong gravitational acceleration that no particles and electromagnetic radiations can escape from it. A wormhole is a speculative structure that links disparate points in spacetime. It is based on a special solution of the Einstein field equations. This chapter discusses the types of blackholes and wormholes in detail.

Chapter 5- A gravitationally bound system of stars, stellar remnants, interstellar gas, dust, and dark matter is termed as a galaxy. Some commonly studied galaxies are Milky Way galaxy, Andromeda galaxy, Magellanic cloud, Quasar, etc. This chapter closely examines these types of galaxies to provide an extensive understanding of the subject.

Chapter 6 - The universe is a combination of space and time. It includes planets, stars, galaxies, and all other forms of matter and energy. Early universe, Friedmann universe, mass, shape, size and density of universe, Hubble's law, steady state theory, cosmological redshift, etc. are some of the concepts studied to better understand our universe. All these concepts have been carefully analyzed in this chapter.

I would like to share the credit of this book with my editorial team who worked tirelessly on this book. I owe the completion of this book to the never-ending support of my family, who supported me throughout the project.

Callen Hogan

Chapter 1

Understanding Cosmology

Cosmology is the branch of astronomy which deals with the study of origin and evolution of universe. It is divided into physical cosmology, philosophical cosmology and religious cosmology. The Big Bang theory forms the basis of origin of universe. This is an introductory chapter which will briefly introduce all the significant aspects of cosmology.

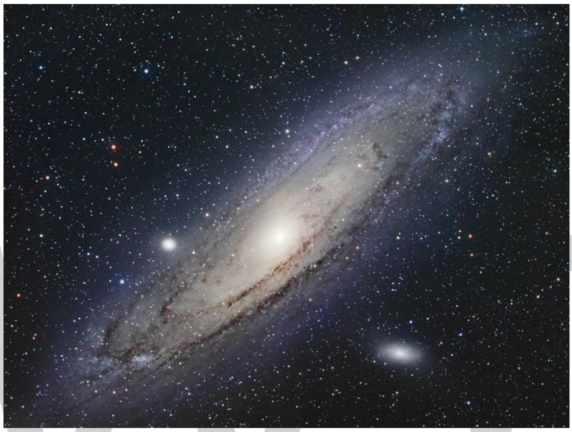
Cosmology is the field of study that brings together the natural sciences, particularly astronomy and physics, in a joint effort to understand the physical universe as a unified whole.

If one looks up on a clear night, one will see that the sky is full of stars. During the summer months in the Northern Hemisphere, a faint band of light stretches from horizon to horizon, a swath of pale white cutting across a background of deepest black. For the early Egyptians, this was the heavenly Nile, flowing through the land of the dead ruled by Osiris. The ancient Greeks likened it to a river of milk. Astronomers now know that the band is actually composed of countless stars in a flattened disk seen edge on. The stars are so close to one another along the line of sight that the unaided eye has difficulty discerning the individual members. Through a large telescope, astronomers find myriads of like systems sprinkled throughout the depths of space.

The Sun is a star around which Earth and the other planets revolve, and by extension every visible star in the sky is a sun in its own right. Some stars are intrinsically brighter than the Sun; others, fainter. Much less light is received from the stars than from the Sun because the stars are all much farther away. Indeed, they appear densely packed in the Milky Way only because there are so many of them. The actual separations of the stars are enormous, so large that it is conventional to measure their distances in units of how far light can travel in a given amount of time. The speed of light (in a vacuum) equals 3×10^{10} cm/sec (centimetres per second); at such a speed, it is possible to circle the Earth seven times in a single second. Thus in terrestrial terms the Sun, which lies 500 light-seconds from the Earth, is very far away; however, even the next closest star, Proxima Centauri, at a distance of 4.3 light-years (4.1×10^{18} cm), is 270,000 times farther yet. The stars that lie on the opposite side of the Milky Way from the Sun have distances that are on the order of 100,000 light-years, which is the typical diameter of a large spiral galaxy.

If the kingdom of the stars seems vast, the realm of the galaxies is larger still. The nearest galaxies to the Milky Way system are the Large and Small Magellanic Clouds, two irregular satellites of the Galaxy visible to the naked eye in the Southern Hemisphere. The Magellanic Clouds are relatively small (containing roughly 10^9 stars) compared to the Galaxy (with some 10^{11} stars), and they lie at a distance of about 200,000 light-years.

The nearest large galaxy comparable to the Galaxy is the Andromeda Galaxy, also called M31 because it was the 31st entry in a catalog of astronomical objects compiled by the French astronomer Charles Messier in 1781, and it lies at a distance of about 2,000,000 light-years. The Magellanic Clouds, the Andromeda Galaxy, and the Milky Way system all are part of an aggregation of two dozen or so neighbouring galaxies known as the Local Group. The Galaxy and M31 are the largest members of this group.



The Andromeda Galaxy, also known as the Andromeda Nebula or M31. It is the closest spiral galaxy to Earth, at a distance of 2.48 million light-years.

The Galaxy and M31 are both spiral galaxies, and they are among the brighter and more massive of all spiral galaxies. The most luminous and brightest galaxies, however, are not spirals but rather supergiant ellipticals (also called cD galaxies by astronomers for historical reasons that are not particularly illuminating). Elliptical galaxies have roundish shapes rather than the flattened distributions that characterize spiral galaxies, and they tend to occur in rich clusters (those containing thousands of members) rather than in the loose groups favoured by spirals. The brightest member galaxies of rich clusters have been detected at distances exceeding several thousand million light-years from the Earth. The branch of learning that deals with phenomena at the scale of many millions of light-years is called cosmology.

Physical Cosmology

Cosmology has developed into a science characterised by precision through ever more accurate measurements of temperature anisotropies in the Cosmic Microwave Background (CMB), along with studies of the expansion history of the Universe, as well as sky surveys providing detailed mapping of large-scale structures.

This exciting development has been possible thanks to ground-breaking discoveries in the theoretical framework of cosmology over the past half century. This year's Nobel Laureate James Peebles has made seminal contributions in this area. Through detailed modelling, with the help of analytic as well as numerical methods, he has explored fundamental properties of our Universe and uncovered unexpected new physics. We now have a unified model capable of describing the Universe from its earliest fraction of a second up until the present and into the distant future.

Modern cosmology is based on Einstein's theory of general relativity and assumes an early era, the Big Bang, when the Universe was extremely hot and dense. A little less than 400,000 years after the Big Bang, the temperature had decreased to about 3,000 K, enabling electrons to combine with nuclei into atoms. Because no charged particles were left that could easily interact with the photons, the Universe became transparent to light. This radiation is now visible as the CMB. Due to the cosmological redshift, its temperature is currently just 2.7 K — a factor of about 1,100 lower since the decoupling of matter and radiation. In figure below, the source of the CMB can be seen as a screen that prevents us from easily looking back in time further than to a few hundred thousand years after the Big Bang.

The first to formulate a mathematical theory for the expanding Universe, using Einstein's newly developed theory of general relativity, was the Russian mathematician and cosmologist Alexander Friedman in 1922. He further developed his theory in 1924. These ideas were rediscovered in 1927 by the Belgian Catholic priest and astronomer Georges Lemaître, who later introduced the notion of a "primeval atom". He argued that the galaxies were receding from each other, and that this could be explained if the Universe expanded. In 1924, the Swedish astronomer Knut Lundmark had made a similar observation, albeit with less rigor and accuracy. A more general acceptance that the Universe was in fact expanding came with the observations by the US astronomer Edwin Hubble in 1929.

It is easy to derive the basic equations that describe the expansion of the Universe, the Friedman equations, even without the use of general relativity. To see this, let us for simplicity assume a homogenous universe. We pick an arbitrary point, at rest relative to matter, draw a sphere around it with radius R, and assume the sphere will grow as the universe expands. On the surface of the sphere, we introduce a small test mass with mass m. The total energy of the test mass is given by,

$$E = \frac{mR^2}{2} - \frac{GMm}{R}$$

where, $M = \frac{4\pi}{3}\rho R^3$. A simple rearrangement gives:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{R^2}$$

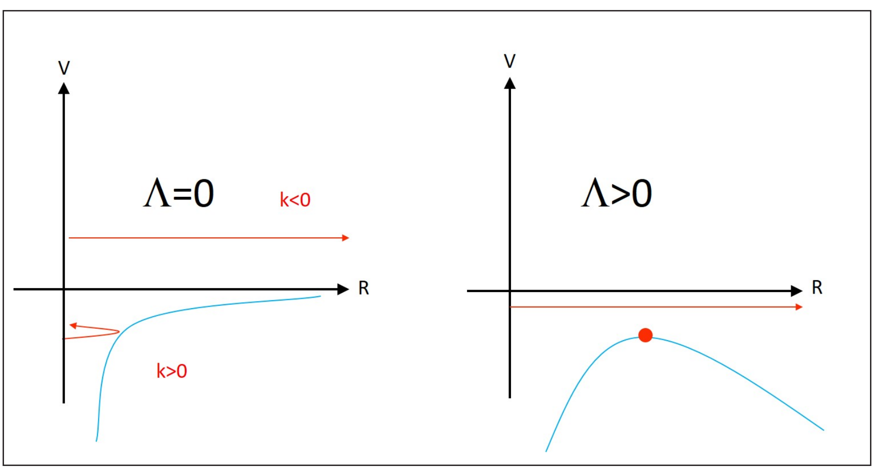
where, $k = -\frac{2E}{mc^2}$. Identifying $H = \frac{\dot{R}}{R}$ as the Hubble constant, this becomes the first Friedman equation. By rescaling R one can set k = ±1, 0. To correctly interpret the meaning of k, we need to appeal to general relativity, where it is identified as the spatial curvature. The value k = 0 corresponds to the critical density of a flat universe given by,

$$\rho_c = \frac{3H^2}{8\pi G}$$

Observations show that the total energy density of the Universe is very close to this value. Defining $\Omega = \frac{\rho}{\rho_c}$, we have Ω < 1 for a universe with negative curvature, Ω = 1 for a flat universe, and Ω > 1 for a universe with positive curvature.

There are several different components of energy in the Universe. Matter in the form of pressureless dust has an energy density that dilutes with volume, described by $1/R^3$, while radiation disperses according to $1/R^4$, due to the loss of energy caused by redshift. In the early Universe, radiation dominated the energy density of the Universe until a bit before recombination. Moreover, in the framework of general relativity, and to account for the possibility that the Universe could have been static, Einstein introduced an additional term in 1917, corresponding to a constant energy contributing to ρ, the cosmological constant, Λ.

Multiplying the Friedman equation with R^2, to think of it as energy conservation, makes it easy to figure out what is actually happening. On the left of figure below we see the effective gravitational potential in the case of matter or radiation. When k > 0, the Universe reaches a maximum size and then re-contracts. If k < 0, it may keep on expanding forever.



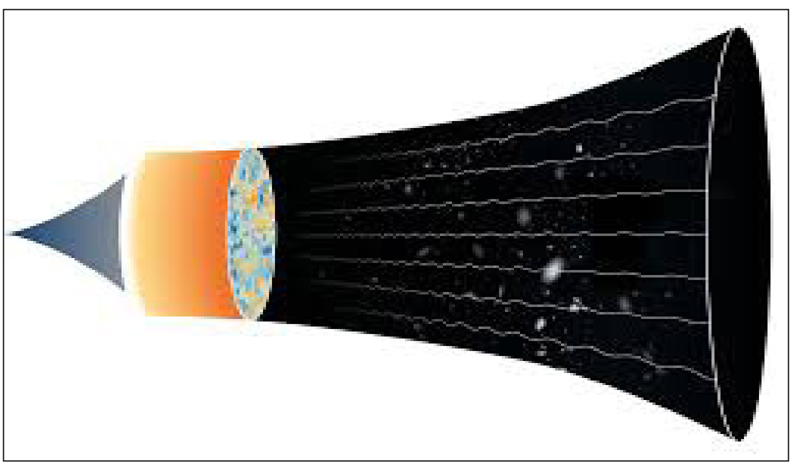
The effective gravitational potential V as a function of the scale factor R, without a cosmological constant on the left and with a cosmological constant contributing to the energy density on the right. The red arrows show what happens in a few cases. The red dot represents Einstein's static, and unstable, universe.

From observations, we know that the amount of ordinary baryonic matter in the form of nucleons, present in stars, clouds of gas, and the like, is no more than 5% of the total energy density at present. In addition, dark matter contributes 26% of the Universe's critical density. It might as well be called invisible matter, as it neither emits nor absorbs light and is so far only known through its gravitational effects.

The most important component is the cosmological constant, representing a constant energy density unaffected by expansion. It is often called dark energy to account for the possibility that it might vary over time and space. That is, dark energy is not necessarily the constant introduced and thought to be related to the vacuum energy in quantum field theory. Observations show that dark energy contributes the remaining 69% of the critical density. As the other components of matter become diluted by the expansion, the dark energy will become evermore important with time (unless its energy density starts to decrease).

In figure, a cosmological constant has been added to the potential with dramatic results. We note how the potential slopes downwards and gives rise to the accelerated expansion. The static universe of Einstein is marked as the unstable point at the maximum of the potential. That dark energy can force galaxies to accelerate away from each other may seem counterintuitive, but it is a direct consequence of the unusual properties of dark energy. The accelerated phase is indicated in the right-hand half of figure, where the number of galaxies thins out.

In figure, the hot Big Bang is the fire in the middle of the diagram, indicating that a preparatory phase, such as inflation, might exist before the hot Big Bang. Inflation is postulated as a scenario with a period of rapid acceleration, which would explain several properties of our Universe, such as its flatness.

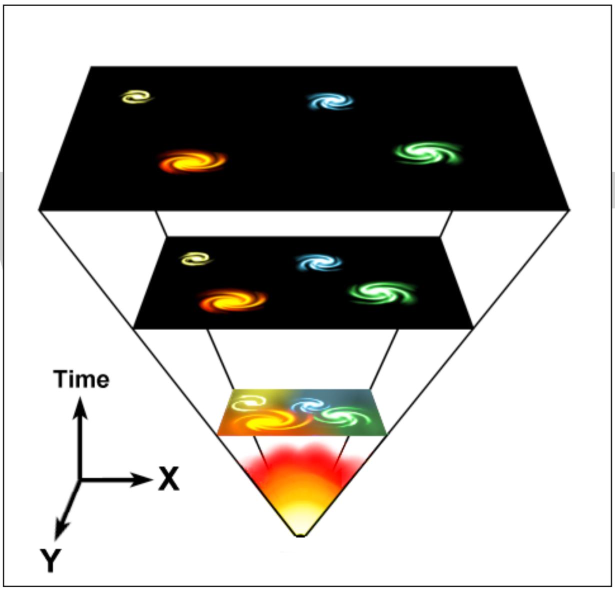


A timeline of our Universe extending from an unknown origin on the left to a darkening future on the right.

Big Bang

The Big Bang is the cosmological model of the universe whose primary assertion is that the universe has expanded into its current state from a primordial condition of enormous density and temperature. The term is also used in a narrower sense to describe the fundamental "fireball" that erupted at or close to an initial time-point in the history of our observed spacetime.

Theoretical support for the Big Bang comes from mathematical models, called Friedmann models. These models show that the Big Bang theory is consistent with the theory of general relativity and with the cosmological principle. The latter states that the properties of the universe should be independent of position or orientation.

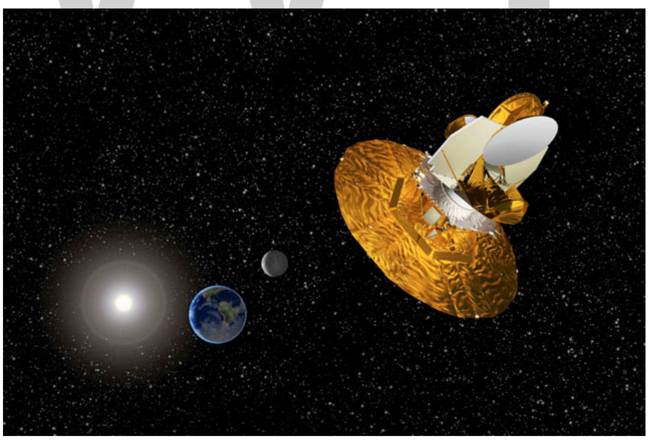


According to the Big Bang model, the universe developed from an extremely dense and hot state. Space itself has been expanding ever since, carrying galaxies (and all other matter) with it.

Observational evidence for the Big Bang includes analyses of the spectra of light from galaxies, which reveals a shift towards longer wavelengths proportional to each galaxy's distance in a relationship described by Hubble's law. Combined with the evidence that observers located anywhere in the universe make similar observations (the Copernican principle), this suggests that space itself is expanding. The next most important observational evidence was the discovery of cosmic microwave background radiation in 1964. This had been predicted as a relic from when hot, ionized plasma of the early universe first cooled sufficiently to form neutral hydrogen, allowing space to become transparent to light. This discovery led to general acceptance among physicists that the Big Bang is the best model for the origin and evolution of the universe. A third important line of evidence is the relative proportion of light chemical elements in the universe, which is a close match to predictions for the formation of light elements in the first minutes of the universe, according to Big Bang nucleosynthesis.

The Big Bang theory developed from observations of the structure of the universe and from theoretical considerations. In 1912 Vesto Slipher measured the first Doppler shift of a “spiral nebula” (spiral nebula is the obsolete term for spiral galaxies), and soon discovered that almost all such nebulae were receding from Earth. He did not grasp the cosmological implications of this fact, and indeed at the time it was highly controversial whether or not these nebulae were “island universes” outside our Milky Way. Ten years later, Alexander Friedmann, a Russian cosmologist and mathematician, derived the Friedmann equations from Albert Einstein’s equations of general relativity, showing that the universe might be expanding in contrast to the static universe model advocated by Einstein. In 1924, Edwin Hubble’s measurement of the great distance to the nearest spiral nebulae showed that these systems were indeed other galaxies. Independently deriving Friedmann’s equations in 1927, Georges Lemaître, a Belgian Roman Catholic priest, predicted that the recession of the nebulae was due to the expansion of the universe. In 1931 Lemaître went further and suggested that the universe began as a simple “primeval atom,” perhaps echoing previous speculations about the cosmic egg origin of the universe.

Starting in 1924, Hubble painstakingly developed a series of distance indicators, the forerunner of the cosmic distance ladder, using the 100 inch Hooker telescope at Mount Wilson Observatory. This allowed him to estimate distances to galaxies whose redshifts had already been measured, mostly by Slipher. In 1929, Hubble discovered a correlation between distance and recession velocity—now known as Hubble’s law. Lemaître had already shown that this was expected, given the cosmological principle.



Artist’s depiction of the WMAP satellite gathering data to help scientists understand the Big Bang.

During the 1930s other ideas were proposed as non-standard cosmologies to explain Hubble’s observations, including the Milne model, the oscillatory universe (originally suggested by Friedmann, but advocated by Einstein and Richard Tolman) and Fritz Zwicky’s tired light hypothesis.

After World War II, two distinct possibilities emerged. One was Fred Hoyle’s steady state model, whereby new matter would be created as the universe seemed to expand.

In this model, the universe is roughly the same at any point in time. The other was Lemaître's Big Bang theory, advocated and developed by George Gamow, who introduced big bang nucleosynthesis and whose associates, Ralph Alpher and Robert Herman, predicted the cosmic microwave background (CMB). It is an irony that it was Hoyle who coined the name that would come to be applied to Lemaître's theory, referring to it as "this big bang idea" during a 1950 BBC radio broadcast. For a while, support was split between these two theories. Eventually, the observational evidence, most notably from radio source counts, began to favor the latter. The discovery of the cosmic microwave background radiation in 1964 secured the Big Bang as the best theory of the origin and evolution of the cosmos. Much of the current work in cosmology includes understanding how galaxies form in the context of the Big Bang, understanding the physics of the universe at earlier and earlier times, and reconciling observations with the basic theory.

Huge strides in Big Bang cosmology have been made since the late 1990s as a result of major advances in telescope technology as well as the analysis of copious data from satellites such as COBE, the Hubble Space Telescope and Wilkinson Microwave Anisotropy Probe (WMAP). Cosmologists now have fairly precise measurement of many of the parameters of the Big Bang model, and have made the unexpected discovery that the expansion of the universe appears to be accelerating.

Extrapolation of the expansion of the universe backwards in time using general relativity yields an infinite density and temperature at a finite time in the past. This singularity signals the breakdown of general relativity. How closely we can extrapolate toward the singularity is debated—certainly not earlier than the Planck epoch. The early hot, dense phase is itself referred to as "the Big Bang", and is considered the "birth" of our universe. Based on measurements of the expansion using Type I_a supernovae, measurements of temperature fluctuations in the cosmic microwave background, and measurements of the correlation function of galaxies, the universe has a calculated age of 13.7 ± 0.2 billion years. The agreement of these three independent measurements strongly supports the ΛCDM model that describes in detail the contents of the universe.

The earliest phases of the Big Bang are subject to much speculation. In the most common models, the universe was filled homogeneously and isotropically with an incredibly high energy density, huge temperatures and pressures, and was very rapidly expanding and cooling. Approximately 10^{-35} seconds into the expansion, a phase transition caused a cosmic inflation, during which the universe grew exponentially. After inflation stopped, the universe consisted of a quark-gluon plasma, as well as all other elementary particles. Temperatures were so high that the random motions of particles were at relativistic speeds, and particle-antiparticle pairs of all kinds were being continuously created and destroyed in collisions. At some point an unknown reaction called baryogenesis violated the conservation of baryon number, leading to a very small excess of quarks and leptons over antiquarks and anti-leptons—of the order of 1 part in 30 million. This resulted in the predominance of matter over antimatter in the present universe.

The universe continued to grow in size and fall in temperature, hence the typical energy of each particle was decreasing. Symmetry breaking phase transitions put the fundamental forces of physics and the parameters of elementary particles into their present form. After about 10^{-11} seconds, the picture becomes less speculative, since particle energies drop to values that can be attained in particle physics experiments. At about 10^{-6} seconds, quarks and gluons combined to form baryons such as protons and neutrons. The small excess of quarks over antiquarks led to a small excess of baryons over antibaryons. The temperature was now no longer high enough to create new proton-antiproton pairs (similarly for neutrons-antineutrons), so a mass annihilation immediately followed, leaving just one in 10^{10} of the original protons and neutrons, and none of their antiparticles. A similar process happened at about 1 second for electrons and positrons. After these annihilations, the remaining protons, neutrons and electrons were no longer moving relativistically and the energy density of the universe was dominated by photons (with a minor contribution from neutrinos).

A few minutes into the expansion, when the temperature was about a billion Kelvin and the density was about that of air, neutrons combined with protons to form the universe's deuterium and helium nuclei in a process called Big Bang nucleosynthesis. Most protons remained uncombined as hydrogen nuclei. As the universe cooled, the rest mass energy density of matter came to gravitationally dominate that of the photon radiation. After about 380,000 years the electrons and nuclei combined into atoms (mostly hydrogen); hence the radiation decoupled from matter and continued through space largely unimpeded. This relic radiation is known as the cosmic microwave background radiation.

Over a long period of time, the slightly denser regions of the nearly uniformly distributed matter gravitationally attracted nearby matter and thus grew even denser, forming gas clouds, stars, galaxies, and other astronomical structures observable today. The details of this process depend on the amount and type of matter in the universe. The three possible types of matter are known as cold dark matter, hot dark matter, and baryonic matter. The best measurements available (from WMAP) show that the dominant form of matter in the universe is cold dark matter. The other two types of matter make up less than 20 percent of the matter in the universe.

The universe today appears to be dominated by a mysterious form of energy known as dark energy. Approximately 70 percent of the total energy density of today's universe is in this form. This dark energy causes the expansion of the universe to accelerate, observed as a slower than expected expansion at very large distances. Dark energy in its simplest formulation takes the form of a cosmological constant term in Einstein's field equations of general relativity, but its composition is unknown and, more generally, the details of its equation of state and relationship with the standard model of particle physics continue to be investigated both observationally and theoretically.

All these observations can be explained by the ΛCDM model of cosmology, which is a mathematical model of the Big Bang with six free parameters. As noted above, there is no compelling physical model for the first 10^{-11} seconds of the universe. To resolve the paradox of the initial singularity, a theory of quantum gravitation is needed. Understanding this period of the history of the universe is one of the greatest unsolved problems in physics.

The Big Bang theory depends on two major assumptions: The universality of physical laws, and the cosmological principle. The cosmological principle states that on large scales the universe is homogeneous and isotropic.

These ideas were initially taken as postulates, but today there are efforts to test each of them. For example, the first assumption has been tested by observations showing that largest possible deviation of the fine structure constant over much of the age of the universe is of order 10^{-5}. Also, General Relativity has passed stringent tests on the scale of the solar system and binary stars while extrapolation to cosmological scales has been validated by the empirical successes of various aspects of the Big Bang theory.

If the large-scale universe appears isotropic as viewed from the Earth, the cosmological principle can be derived from the simpler Copernican principle, which states that there is no preferred (or special) observer or vantage point. To this end, the cosmological principle has been confirmed to a level of 10^{-5} via observations of the CMB. The universe has been measured to be homogeneous on the largest scales at the 10 percent level.

FLRW Metric

General relativity describes spacetime by a metric, which determines the distances that separate nearby points. The points, which can be galaxies, stars, or other objects, themselves are specified using a coordinate chart or "grid" that is laid down over all spacetime. The cosmological principle implies that the metric should be homogeneous and isotropic on large scales, which uniquely singles out the Friedmann-Lemaître-Robertson-Walker metric (FLRW metric). This metric contains a scale factor, which describes how the size of the universe changes with time. This enables a convenient choice of a coordinate system to be made, called comoving coordinates. In this coordinate system, the grid expands along with the universe, and objects that are moving only due to the expansion of the universe remain at fixed points on the grid. While their coordinate distance (comoving distance) remains constant, the physical distance between two such comoving points expands proportionally with the scale factor of the universe.

The Big Bang is not an explosion of matter moving outward to fill an empty universe. Instead, space itself expands with time everywhere and increases the physical distance between two comoving points. Because the FLRW metric assumes a uniform distribution of mass and energy, it applies to our universe only on large scales—local concentrations of matter such as our galaxy are gravitationally bound and as such do not experience the large-scale expansion of space.

Horizons

An important feature of the Big Bang spacetime is the presence of horizons. Since the universe has a finite age, and light travels at a finite speed, there may be events in the past whose light has not had time to reach us. This places a limit or a *past horizon* on the most distant objects that can be observed. Conversely, because space is expanding, and more distant objects are receding ever more quickly, light emitted by us today may never "catch up" to very distant objects. This defines a *future horizon,* which limits the events in the future that we will be able to influence. The presence of either type of horizon depends on the details of the FRW model that describes our universe. Our understanding of the universe back to very early times suggests that there was a past horizon, though in practice our view is limited by the opacity of the universe at early times. If the expansion of the universe continues to accelerate, there is a future horizon as well.

Observational Evidence

The earliest and most direct kinds of observational evidence are the Hubble-type expansion seen in the redshifts of galaxies, the detailed measurements of the cosmic microwave background, and the abundance of light elements. These are sometimes called the three pillars of the Big Bang theory. Many other lines of evidence now support the picture, notably various properties of the large-scale structure of the cosmos which are predicted to occur due to gravitational growth of structure in the standard Big Bang theory.

Hubble's Law and the Expansion of Space

Observations of distant galaxies and quasars show that these objects are redshifted—the light emitted from them has been shifted to longer wavelengths. This can be seen by taking a frequency spectrum of an object and matching the spectroscopic pattern of emission lines or absorption lines corresponding to atoms of the chemical elements interacting with the light. These redshifts are uniformly isotropic, distributed evenly among the observed objects in all directions. If the redshift is interpreted as a Doppler shift, the recessional velocity of the object can be calculated. For some galaxies, it is possible to estimate distances via the cosmic distance ladder. When the recessional velocities are plotted against these distances, a linear relationship known as Hubble's law is observed:

$$v=H_oD$$

where,

- v is the recessional velocity of the galaxy or other distant object.

- D is the distance to the object.
- H_o is Hubble's constant, measured to be (70 +2.4/-3.2) (km/s)/Mpc by the WMAP probe.

Hubble's law has two possible explanations. We are at the center of an explosion of galaxies—which is untenable given the Copernican principle—or the universe is uniformly expanding everywhere. This universal expansion was predicted from general relativity by Alexander Friedman in 1922 and Georges Lemaître in 1927, well before Hubble made his 1929 analysis and observations, and it remains the cornerstone of the Big Bang theory as developed by Friedmann, Lemaître, Robertson and Walker.

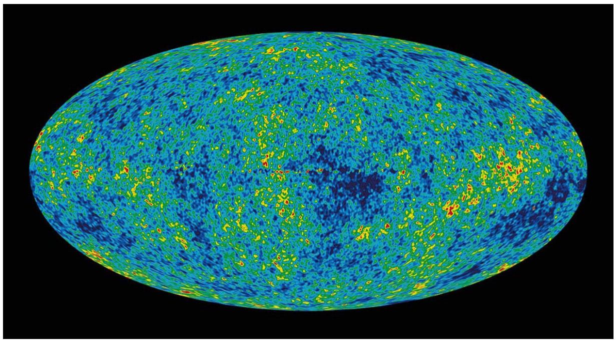
The theory requires the relation $V=H_oD$ to hold at all times, where D is the proper distance, v=dD/dt, and v, H, and D all vary as the universe expands (hence we write H_o to denote the present-day Hubble "constant"). For distances much smaller than the size of the observable universe, the Hubble redshift can be thought of as the Doppler shift corresponding to the recession velocity v. However, the redshift is not a true Doppler shift, but rather the result of the expansion of the universe between the time the light was emitted and the time that it was detected.

That space is undergoing metric expansion is shown by direct observational evidence of the Cosmological Principle and the Copernican Principle, which together with Hubble's law have no other explanation. Astronomical redshifts are extremely isotropic and homogeneous, supporting the Cosmological Principle that the universe looks the same in all directions, along with much other evidence. If the redshifts were the result of an explosion from a center distant from us, they would not be so similar in different directions.

Measurements of the effects of the cosmic microwave background radiation on the dynamics of distant astrophysical systems in 2000 proved the Copernican principle that the Earth is not in a central position, on a cosmological scale. Radiation from the Big Bang was demonstrably warmer at earlier times throughout the universe. Uniform cooling of the cosmic microwave background over billions of years is explainable only if the universe is experiencing a metric expansion, and excludes the possibility that we are near the unique center of an explosion.

Cosmic Microwave Background Radiation

During the first few days of the universe, the universe was in full thermal equilibrium, with photons being continually emitted and absorbed, giving the radiation a blackbody spectrum. As the universe expanded, it cooled to a temperature at which photons could no longer be created or destroyed. The temperature was still high enough for electrons and nuclei to remain unbound, however, and photons were constantly "reflected" from these free electrons through a process called Thomson scattering. Because of this repeated scattering, the early universe was opaque to light.



When the temperature fell to a few thousand Kelvin, electrons and nuclei began to combine to form atoms, a process known as recombination. Since photons scatter infrequently from neutral atoms, radiation decoupled from matter when nearly all the electrons had recombined, at the *epoch of last scattering,* 380,000 years after the Big Bang. These photons make up the cosmic microwave background radiation (CMB) that is observed today, and the observed pattern of fluctuations in the CMB is a direct picture of the universe at this early epoch. The energy of photons was subsequently redshifted by the expansion of the universe, which preserved the blackbody spectrum but caused its temperature to fall, meaning that the photons now fall into the microwave region of the electromagnetic spectrum. The radiation is thought to be observable at every point in the universe, and comes from all directions with (almost) the same intensity.

In 1964, Arno Penzias and Robert Wilson accidentally discovered the cosmic background radiation while conducting diagnostic observations using a new microwave receiver owned by Bell Laboratories. Their discovery provided substantial confirmation of the general CMB predictions—the radiation was found to be isotropic and consistent with a blackbody spectrum of about 3 K—and it pitched the balance of opinion in favor of the Big Bang hypothesis. Penzias and Wilson were awarded a Nobel Prize for their discovery.

In 1989, NASA launched the Cosmic Background Explorer satellite (COBE), and the initial findings, released in 1990, were consistent with the Big Bang's predictions regarding the CMB. COBE found a residual temperature of 2.726 K and in 1992 detected for the first time the fluctuations (anisotropies) in the CMB, at a level of about one part in 10^5. John C. Mather and George Smoot were awarded Nobel Prizes for their leadership in this work. During the following decade, CMB anisotropies were further investigated by a large number of ground-based and balloon experiments. In 2000–2001, several experiments, most notably BOOMERanG, found the universe to be almost geometrically flat by measuring the typical angular size (the size on the sky) of the anisotropies.

In early 2003, the first results of the Wilkinson Microwave Anisotropy Probe (WMAP) were released, yielding what were at the time the most accurate values for some of the cosmological parameters. This satellite also disproved several specific cosmic inflation models, but the results were consistent with the inflation theory in general. This satellite is still gathering data. Another satellite will be launched within the next few years, the Planck Surveyor, which will provide even more accurate measurements of the CMB anisotropies. Many other ground- and balloon-based experiments are also currently running; see Cosmic microwave background experiments.

The background radiation is exceptionally smooth, which presented a problem in that conventional expansion would mean that photons coming from opposite directions in the sky were coming from regions that had never been in contact with each other. The leading explanation for this far reaching equilibrium is that the universe had a brief period of rapid exponential expansion, called inflation. This would have the effect of driving apart regions that had been in equilibrium, so that all the observable universe was from the same equilibrated region.

Abundance of Primordial Elements

Using the Big Bang model it is possible to calculate the concentration of helium-4, helium-3, deuterium and lithium-7 in the universe as ratios to the amount of ordinary hydrogen, H. All the abundances depend on a single parameter, the ratio of photons to baryons, which itself can be calculated independently from the detailed structure of CMB fluctuations. The ratios predicted (by mass, not by number) are about 0.25 for ^{4}He/H, about 10^{-3} for ^{2}H/H, about 10^{-4} for ^{3}He/H and about 10^{-9} for ^{7}Li/H.

The measured abundances all agree at least roughly with those predicted from a single value of the baryon-to-photon ratio. The agreement is excellent for deuterium, close but formally discrepant for ^{4}He, and a factor of two off for ^{7}Li; in the latter two cases there are substantial systematic uncertainties. Nonetheless, the general consistency with abundances predicted by BBN is strong evidence for the Big Bang, as the theory is the only known explanation for the relative abundances of light elements, and it is virtually impossible to "tune" the Big Bang to produce much more or less than 20–30 percent helium. Indeed there is no obvious reason outside of the Big Bang that, for example, the young universe (i.e., before star formation, as determined by studying matter supposedly free of stellar nucleosynthesis products) should have more helium than deuterium or more deuterium than ^{3}He, and in constant ratios, too.

Galactic Evolution and Distribution

Detailed observations of the morphology and distribution of galaxies and quasars provide strong evidence for the Big Bang. A combination of observations and theory suggest that the first quasars and galaxies formed about a billion years after the Big Bang, and since then larger structures have been forming, such as galaxy clusters and superclusters. Populations of stars have been aging and evolving, so that distant galaxies (which are observed as they were in the early universe) appear very different from nearby galaxies (observed in a more recent state). Moreover, galaxies that formed relatively recently appear markedly different from galaxies formed at similar distances but shortly after the Big Bang. These observations are strong arguments against the steady-state model. Observations of star formation, galaxy and quasar distributions and larger structures agree well with Big Bang simulations of the formation of structure in the universe and are helping to complete details of the theory.

Other Lines of Evidence

After some controversy, the age of universe as estimated from the Hubble expansion and the CMB is now in good agreement with (i.e., slightly larger than) the ages of the oldest stars, both as measured by applying the theory of stellar evolution to globular clusters and through radiometric dating of individual Population II stars.

The prediction that the CMB temperature was higher in the past has been experimentally supported by observations of temperature-sensitive emission lines in gas clouds at high redshift. This prediction also implies that the amplitude of the Sunyaev-Zel'dovich effect in clusters of galaxies does not depend directly on redshift; this seems to be roughly true, but unfortunately the amplitude does depend on cluster properties which do change substantially over cosmic time, so a precise test is impossible.

Features, Issues and Problems

While very few researchers now doubt the Big Bang occurred, the scientific community was once divided between supporters of the Big Bang and those of alternative cosmological models. Throughout the historical development of the subject, problems with the Big Bang theory were posed in the context of a scientific controversy regarding which model could best describe the cosmological observations. With the overwhelming consensus in the community today supporting the Big Bang model, many of these problems are remembered as being mainly of historical interest; the solutions to them have been obtained either through modifications to the theory or as the result of better observations. Other issues, such as the cuspy halo problem and the dwarf galaxy problem of cold dark matter, are not considered to be fatal as it is anticipated that they can be solved through further refinements of the theory.

The core ideas of the Big Bang—the expansion, the early hot state, the formation of helium, the formation of galaxies—are derived from many independent observations including Big Bang nucleosynthesis, the cosmic microwave background, large scale structure, and Type Ia supernovae, and can hardly be doubted as important and real features of our universe.

Precise modern models of the Big Bang appeal to various exotic physical phenomena that have not been observed in terrestrial laboratory experiments or incorporated into the Standard Model of particle physics. Of these features, dark energy and dark matter are considered the most secure, while inflation and baryogenesis remain speculative: they provide satisfying explanations for important features of the early universe, but could be replaced by alternative ideas without affecting the rest of the theory. Explanations for such phenomena remain at the frontiers of inquiry in physics.

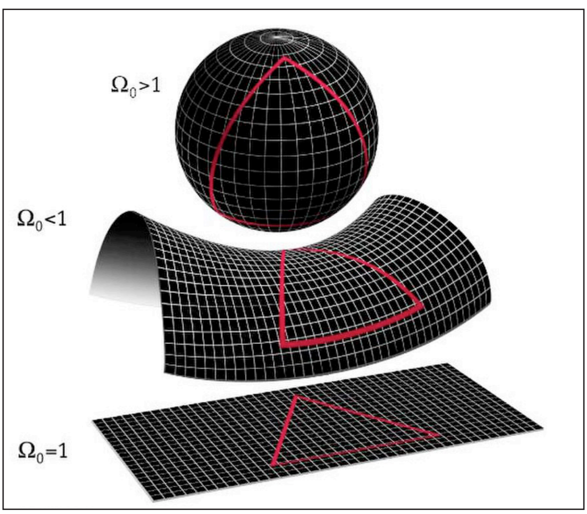
Horizon Problem

The horizon problem results from the premise that information cannot travel faster than light. In a universe of finite age, this sets a limit the particle horizon on the separation of any two regions of space that are in causal contact. The observed isotropy of the CMB is problematic in this regard: If the universe had been dominated by radiation or matter at all times up to the epoch of last scattering, the particle horizon at that time would correspond to about 2 degrees on the sky. There would then be no mechanism to cause these regions to have the same temperature.

A resolution to this apparent inconsistency is offered by inflationary theory in which a homogeneous and isotropic scalar energy field dominates the universe at some very early period. During inflation, the universe undergoes exponential expansion, and the particle horizon expands much more rapidly than previously assumed, so that regions presently on opposite sides of the observable universe are well inside each other's particle horizon. The observed isotropy of the CMB then follows from the fact that this larger region was in causal contact before the beginning of inflation.

Heisenberg's uncertainty principle predicts that during the inflationary phase there would be quantum thermal fluctuations, which would be magnified to cosmic scale. These fluctuations serve as the seeds of all current structure in the universe. Inflation predicts that the primordial fluctuations are nearly scale invariant and Gaussian, which has been accurately confirmed by measurements of the CMB.

Flatness/Oldness Problem



The overall geometry of the universe is determined by whether the Omega cosmological parameter is less than, equal to or greater than 1. From top to bottom: A closed universe with positive curvature, a hyperbolic universe with negative curvature and a flat universe with zero curvature.

The flatness problem (also known as the oldness problem) is an observational problem associated with a Friedmann-Lemaître-Robertson-Walker metric. The universe may have positive, negative or zero spatial curvature depending on its total energy density. Curvature is negative if its density is less than the critical density, positive if greater, and zero at the critical density, in which case space is said to be *flat*. The problem is that any small departure from the critical density grows with time, and yet the universe today remains very close to flat. Given that a natural timescale for departure from flatness might be the Planck time, 10^{-43} seconds, the fact that the universe has reached neither a Heat Death nor a Big Crunch after billions of years requires some explanation. For instance, even at the relatively late age of a few minutes (the time of nucleosynthesis), the universe must have been within one part in 10^{14} of the critical density, or it would not exist as it does today.

A resolution to this problem is offered by inflationary theory. During the inflationary period, spacetime expanded to such an extent that its curvature would have been smoothed out. Thus, it is believed that inflation drove the universe to a very nearly spatially flat state, with almost exactly the critical density.

Magnetic Monopoles

The magnetic monopole objection was raised in the late 1970s. Grand unification theories predicted topological defects in space that would manifest as magnetic monopoles. These objects would be produced efficiently in the hot early universe, resulting in a density much higher than is consistent with observations, given that searches have never found any monopoles. This problem is also resolved by cosmic inflation, which removes all point defects from the observable universe in the same way that it drives the geometry to flatness.

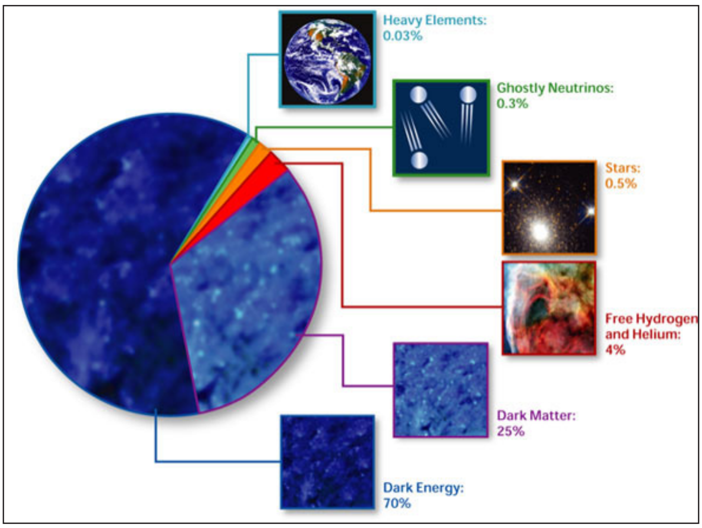
Baryon Asymmetry

It is not yet understood why the universe has more matter than antimatter. It is generally assumed that when the universe was young and very hot, it was in statistical equilibrium and contained equal numbers of baryons and anti-baryons. However, observations suggest that the universe, including its most distant parts, is made almost entirely of matter. An unknown process called “baryogenesis” created the asymmetry. For baryogenesis to occur, the Sakharov conditions must be satisfied. These require that baryon number is not conserved, that C-symmetry and CP-symmetry are violated and that the universe depart from thermodynamic equilibrium. All these conditions occur in the Standard Model, but the effect is not strong enough to explain the present baryon asymmetry.

Globular Cluster Age

In the mid-1990s, observations of globular clusters appeared to be inconsistent with the Big Bang. Computer simulations that matched the observations of the stellar populations of globular clusters suggested that they were about 15 billion years old, which conflicted with the 13.7-billion-year age of the universe. This issue was generally resolved in the late 1990s when new computer simulations, which included the effects of mass loss due to stellar winds, indicated a much younger age for globular clusters. There still remain some questions as to how accurately the ages of the clusters are measured, but it is clear that these objects are some of the oldest in the universe.

Dark Matter



A pie chart indicating the proportional composition of different energy-density components of the universe, according to the best CDM model fits. Roughly ninety-five percent is in the exotic forms of dark matter and dark energy.

During the 1970s and 1980s, various observations showed that there is not sufficient visible matter in the universe to account for the apparent strength of gravitational forces within and between galaxies. This led to the idea that up to 90 % of the matter in the universe is dark matter that does not emit light or interact with normal baryonic matter. In addition, the assumption that the universe is mostly normal matter led to predictions that were strongly inconsistent with observations. In particular, the universe today is far more lumpy and contains far less deuterium than can be accounted for without dark matter. While dark matter was initially controversial, it is now indicated by numerous observations: the anisotropies in the CMB, galaxy cluster velocity dispersions, large-scale structure distributions, gravitational lensing studies, and X-ray measurements of galaxy clusters.

The evidence for dark matter comes from its gravitational influence on other matter, and no dark matter particles have been observed in laboratories. Many particle physics candidates for dark matter have been proposed, and several projects to detect them directly are underway.

Dark Energy

Measurements of the redshift–magnitude relation for type Ia supernovae have revealed that the expansion of the universe has been accelerating since the universe was about half its present age. To explain this acceleration, general relativity requires that much of the energy in the universe consists of a component with large negative pressure, dubbed "dark energy." Dark energy is indicated by several other lines of evidence. Measurements of the cosmic microwave background indicate that the universe is very nearly spatially flat, and therefore according to general relativity the universe must have almost exactly the critical density of mass/energy. But the mass density of the universe can be measured from its gravitational clustering, and is found to have only about 30% of the critical density. Since dark energy does not cluster in the usual way it is the best explanation for the "missing" energy density. Dark energy is also required by two geometrical measures of the overall curvature of the universe, one using the frequency of gravitational lenses, and the other using the characteristic pattern of the large-scale structure as a cosmic ruler.

Negative pressure is a property of vacuum energy, but the exact nature of dark energy remains one of the great mysteries of the Big Bang. Possible candidates include a cosmological constant and quintessence. Results from the WMAP team in 2006, which combined data from the CMB and other sources, indicate that the universe today is 74 % dark energy, 22 % dark matter, and 4 % regular matter. The energy density in matter decreases with the expansion of the universe, but the dark energy density remains constant (or nearly so) as the universe expands. Therefore matter made up a larger fraction of the total energy of the universe in the past than it does today, but its fractional contribution will fall in the far future as dark energy becomes even more dominant.

In the Lambda-CDM model, the best current model of the Big Bang, dark energy is explained by the presence of a cosmological constant in the theory of General Relativity. However, the size of the constant that properly explains dark energy is surprisingly small relative to naive estimates based on ideas about quantum gravity. Distinguishing between the cosmological constant and other explanations of dark energy is an active area of current research.

The Future according to the Big Bang Theory

Before observations of dark energy, cosmologists considered two scenarios for the future of the universe. If the mass density of the universe were greater than the critical density, then the universe would reach a maximum size and then begin to collapse. It would become denser and hotter again, ending with a state that was similar to that in which it started—a Big Crunch. Alternatively, if the density in the universe were equal to or below the critical density, the expansion would slow down, but never stop. Star formation would cease as all the interstellar gas in each galaxy is consumed; stars would burn out leaving white dwarfs, neutron stars, and black holes. Very gradually, collisions between these would result in mass accumulating into larger and larger black holes. The average temperature of the universe would asymptotically approach absolute zero—a Big Freeze. Moreover, if the proton were unstable, then baryonic matter would disappear, leaving only radiation and black holes. Eventually, black holes would evaporate. The entropy of the universe would increase to the point where no organized form of energy could be extracted from it, a scenario known as heat death.

Modern observations of accelerated expansion imply that more and more of the currently visible universe will pass beyond our event horizon and out of contact with us. The eventual result is not known. The CDM model of the universe contains dark energy in the form of a cosmological constant. This theory suggests that only gravitationally bound systems, such as galaxies, would remain together, and they too would be subject to heat death, as the universe expands and cools. Other explanations of dark energy—so-called phantom energy theories—suggest that ultimately galaxy clusters, stars, planets, atoms, nuclei and matter itself will be torn apart by the ever-increasing expansion in a so-called Big Rip.

Speculative Physics beyond the Big Bang

While the Big Bang model is well established in cosmology, it is likely to be refined in the future. Little is known about the earliest moments of the universe's history. The Penrose-Hawking singularity theorems require the existence of a singularity at the beginning of cosmic time. However, these theorems assume that general relativity is correct, but general relativity must break down before the universe reaches the Planck temperature, and a correct treatment of quantum gravity may avoid the singularity.

There may also be parts of the universe well beyond what can be observed in principle. If inflation occurred this is likely, for exponential expansion would push large regions of space beyond our observable horizon.

Some proposals, each of which entails untested hypotheses, are:

- Models including the Hartle-Hawking boundary condition in which the whole of space-time is finite; the Big Bang does represent the limit of time, but without the need for a singularity.

- Brane cosmology models in which inflation is due to the movement of branes in string theory; the pre-big bang model; the ekpyrotic model, in which the Big Bang is the result of a collision between branes; and the cyclic model, a variant of the ekpyrotic model in which collisions occur periodically.

- Chaotic inflation, in which inflation events start here and there in a random quantum-gravity foam, each leading to a *bubble universe* expanding from its own big bang.

Philosophical Cosmology

As recently as 1960, cosmology was widely regarded as a branch of philosophy. It has transitioned to an extremely active area of mainstream physics and astronomy, particularly due to the application to the early universe of atomic and nuclear physics, on the one hand, and to a flood of data coming in from telescopes operating across the entire electromagnetic spectrum on the other. However, there are two main issues that make the philosophy of cosmology unlike that of any other science.

The first is: The uniqueness of the Universe: There exists only one universe, so there is nothing else similar to compare it with, and the idea of "Laws of the universe" hardly makes sense.

This means it is the historical science par excellence: It deals with only one unique object that is the only member of its class that exists physically; indeed there is no non-trivial class of such objects (except in theoreticians' minds) precisely for this reason.

The second is: Cosmology deals with the physical situation that is the context in the large for human existence: The universe has such a nature that our life is possible.

This means that although it is a physical science, it is of particular importance in terms of its implications for human life.

Cosmology's Standard Model

Physical cosmology has achieved a consensus Standard Model (SM), based on extending the local physics governing gravity and the other forces to describe the overall structure of the universe and its evolution. According to the SM, the universe has evolved from an extremely high temperature early state, by expanding, cooling, and developing structures at various scales, such as galaxies and stars. This model is based on bold extrapolations of existing theories—applying general relativity, for example, at length scales 14 orders of magnitude larger than the those at which it has been tested—and requires several novel ingredients, such as dark matter and dark energy. The last few decades have been a golden age of physical cosmology, as the SM has been developed in rich detail and substantiated by compatibility with a growing body of observations.

Spacetime Geometry

Gravity is the dominant interaction at large length scales. General relativity introduced a new way of representing gravity: Rather than describing gravity as a force deflecting bodies from inertial motion, bodies free from non-gravitational forces move along the analog of straight lines, called geodesics, through a curved spacetime geometry. The spacetime curvature is related to the distribution of energy and matter through GR's fundamental equations (Einstein's field equations, EFE). The dynamics of the theory are non-linear: matter curves spacetime, and the curvature of spacetime determines how matter moves; and gravitational waves interact with each other gravitationally, and act as gravitational sources. The theory also replaces the single gravitational potential, and associated field equation, of Newton's theory, with a set of 10 coupled, non-linear equations for ten independent potentials. This complexity is an obstacle to understanding the general features of solutions to EFE, and to finding exact solutions to describe specific physical situations. Most exact solutions have been found based on strong idealizations, introduced to simplify the mathematics.

Remarkably, much of cosmology is based on an extremely simple set of solutions found within a decade of Einstein's discovery of GR. These Friedman-Lemaître-Robertson-Walker (FLRW) solutions have, in a precise sense, the most symmetry possible. The spacetime geometry is constrained to be uniform, so that there are no preferred locations or directions. They have a simple geometric structure, consisting of a "stack" of three-dimensional spatial surfaces Σ(t) labeled by values of the cosmic time t (topologically, $\Sigma \times \mathbb{R}$). The surfaces Σ(t) are three-dimensional spaces (Riemannian manifolds) of constant curvature, with three possibilities: (1) spherical space, for the case of positive curvature; (2) Euclidean space, for zero curvature; and (3) hyperbolic space, for negative curvature.

These models describe an expanding universe, characterized fully by the behavior of the scale factor R(t). The worldlines of "fundamental observers", defined as at rest with respect to matter, are orthogonal to these surfaces, and the cosmic time corresponds to the proper time measured by the fundamental observers. The scale factor R(t) represents the spatial distance in Σ between nearby fundamental observers as a function of cosmic time. The evolution of these models is described by a simple set of equations governing R(t), implied by Einstein's field equations (EFE): the Friedmann equation,

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

and the isotropic form of the Raychaudhuri equation:

$$3\frac{\ddot{R}}{R} = -4\pi G(\rho + 3p) + \Lambda.$$

The curvature of surfaces Σ(t) of constant cosmic time is given by $k/R^2(t)$, where k={−1,0,1}for negative, flat, and positive curvature (respectively). The assumed symmetries force the matter to be described as a perfect fluid with energy density ρ and pressure p, which obey the energy conservation equation:

$$\dot{\rho}+(\rho+p)3\frac{\dot{R}}{R}=0$$

The unrelenting symmetry of the FLRW models makes them quite simple geometrically and dynamically. Rather than a set of coupled partial differential equations, which generically follow from EFE, in the FLRW models one only has to deal with 2 ordinary differential equations just given above which are determinate once an equation of state p=p(ρ) is given.

These equations reveal three basic features of these models. First, these are dynamical models: It is hard to arrange an unchanging universe, with $\dot{R}(t)=0$. "Ordinary" matter has positive total stress-energy density, in the sense that $\rho_{grav}:=\rho+3p>0$. From $3\frac{\ddot{R}}{R}=-4\pi G(\rho+3p)+\Lambda$ the effect of such ordinary matter is to decelerate cosmic expansion, $\ddot{R}<0$ gravity is a force of attraction. This is only so for ordinary matter: A positive cosmological constant, or matter with negative gravitational-energy density ρ_{grav} leads, conversely, to accelerating expansion $\ddot{R}>0$. Einstein was only able to construct a static model by delicately balancing the attraction of ordinary matter with a precisely chosen value of Λ; he unfortunately failed to notice that the solution was unstable, and overlooked the dynamical implications of his own theory.

Second, the expansion rate varies as different types of matter come to dominate the dynamics. As shown by ($\dot{\rho}+(\rho+p)3\frac{\dot{R}}{R}=0$), the energy density for different types of matter and radiation dilutes at different rates: for example, pressureless dust (p=0) dilutes as $\propto R^{-3}$, radiation (p=ρ/3) as $\propto R^{-4}$, and the cosmological constant (p=−ρ) remains (as the name suggests) constant. The SM describes the early universe as having a much higher energy density in radiation than matter. This radiation-dominated phase eventually transitions to a matter-dominated phase as radiation dilutes more rapidly, followed eventually, if Λ>0, by a transition to a Λ-dominated phase; if k≠0 there may also be a curvature dominated phase.

Third, FLRW models with ordinary matter have a singularity at a finite time in the past. Extrapolating back in time, given that the universe is currently expanding, equation $3\frac{\ddot{R}}{R}=-4\pi G(\rho+3p)+\Lambda$ implies that the expansion began at some finite time in the past. The current rate of expansion is given by the Hubble parameter, $H_0=\left(\frac{\dot{R}}{R}\right)0$. Simply extrapolating this expansion rate backward, from equation $3\frac{\ddot{R}}{R}=-4\pi G(\rho+3p)+\Lambda$.

the expansion rate must increase at earlier times, so $R(t)\rightarrow 0$ at a time less than the Hubble time Hubble time H_o^{-1} before now, if $\rho_{grav}\geq 0$. As this "big bang" is approached, the energy density and curvature increase without bound provided $\rho_{inert}:=(\rho+p)>0$ (which condition guarantees that $\rho\rightarrow\infty$ as $R\rightarrow 0$). This reflects gravitational instability: as R(t) decreases, the energy density and pressure both increase, and they both appear with the same sign on the right hand side of equation $3\frac{\ddot{R}}{R}=-4\pi G(\rho+3p)+\Lambda.$, hence pressure p>0 does not help avoid the singularity.

The SM adds small departures from strict uniformity in order to account for the formation and evolution of structure. Due to gravitational instability, such perturbations are enhanced dynamically—the density contrast of an initial region that differs from the average density grows with time. Sufficiently small fluctuations can be treated as linear perturbations to a background cosmological model, governed by an evolution equation that follows from EFE. Yet as the fluctuations grow larger, linearized perturbation theory no longer applies. According to the SM, structure grows hierarchically with smaller length scales going non-linear first and larger structures forming via later mergers. Models of evolution of structures at smaller length scales (e.g., the length scales of galaxies) include physics other than gravity, such as gas dynamics, to describe the collapsing clumps of matter. Cold dark matter (CDM) also plays a crucial role in the SM's account of structure formation: It clumps first, providing scaffolding for clumping of baryonic matter.

A full account of structure formation requires integrating physics over an enormous range of dynamical scales and including a cosmological constant as well as baryonic matter, radiation, and dark matter. This is an active area of research, primarily pursued using sophisticated NN-body computer simulations to study features of the galaxy distribution produced by the SM, given various assumptions.

Observations

There are two main ways in which cosmological observations support perturbed FLRW models. First, cosmologists use matter and radiation in the universe to probe the background spacetime geometry and its evolution. The universe appears to be isotropic at sufficiently large scales, as indicated by background radiation (most notably the cosmic microwave background radiation (CMB)) and discrete sources (e.g., galaxies). Isotropy observed along a single worldline is, however, not sufficient to establish the universe is well described by an FLRW geometry. A further assumption that our worldline is not the only vantage point from which the universe appears isotropic, often called the Copernican principle, is needed. Granting this principle, there are theorems establishing that observations of almost isotropic background radiation implies that the spacetime geometry is almost FLRW. The principle itself cannot be established directly via observations. Given that we live in an almost FLRW models, we need to determine its parameters such as the Hubble constant H_o and the deceleration parameter $q_0:=-\ddot{R}/(RH_0^2)$

which measures how the rate of expansion is changing, and the normalized density parameters $\Omega_m := \rho m/(3H_2 0)$ for each matter or energy density component mm. There are a variety of ways to determine the accuracy of the background evolution described by the FLRW models, which depends on these parameters. For this purpose, cosmologists seek effective standard candles and standard rulers—objects with a known intrinsic luminosity and length, respectively, which can then be used to measure the expansion history of the universe.

The second main avenue of testing focuses on the SM's account of structure formation, which describes the evolution of small perturbations away from the background FLRW geometry in terms of a small number of parameters such as the tilt n_s and the scalar to tensor ratio r. Observations from different epochs, such as temperature anisotropies in the CMB and the matter power spectrum based on galaxy surveys, can be used as independent constraints on these parameters as well as on the background parameters (indeed such observations turn out to give the best constraints on the background model parameters). These two routes to testing almost FLRW space time geometry are closely linked because the background model provides the context for the evolution of perturbations under the dynamics described by general relativity.

The remarkable success of perturbed FLRW models in describing the observed universe has led many cosmologists to focus almost exclusively on them, yet there are drawbacks to such a myopic approach. For example, the observations at best establish that the observed universe can be well-approximated by an almost FLRW model within some (large) domain. But they are not the only models that fit the data: there are other cosmological models that mimic FLRW models in the relevant domain, yet differ dramatically elsewhere (and elsewhen). Specifically, on the one hand there are a class of spatially homogeneous and anisotropic models (Bianchi models) that exhibit "intermediate isotropization": namely, they have physical properties that are arbitrarily close to (isotropic) FLRW models over some time scale T. Agreement over the time interval T does not imply global agreement, however, as these models have large anisotropies at other times. Relying on the FLRW models in making extrapolations to the early or late universe requires some justification for ignoring models, such as these Bianchi models, that mimic their behavior for a finite time interval. On the other hand there are inhomogeneous spherically symmetric models that can reproduce exactly the background model observations (number counts versus redshifts and angular diameter distance versus redshift, for example) with or without a cosmological constant. These can be excluded by direct observations with good enough standard candles or by observations of structure formation features in such universes but that exclusion cannot take place unless one indeed examines such models and their observational consequences.

Lack of knowledge of the full space of solutions to EFE makes it difficult to assess the fragility of various inferences cosmologists make based on perturbed FLRW models. A fragile inference depends on the properties of the model holding exactly, contrasted with robust inferences that hold even if the models are good approximations (up to some tolerable error) that will hold even if the model is perturbed. The singularity theorems, for example, establish that the existence of an initial singularity is robust: rather than being features specific to the FLRW models, or other highly symmetric models, singularities are generic in models satisfying physically plausible assumptions. The status of various other inferences cosmologists make is less clear. For example, how sensitively does the observational case in favor of dark energy, which contributes roughly 70% of the total energy density of the universe in the SM, depend upon treating the universe as having almost-FLRW spacetime geometry? Studies along these lines are needed to evaluate the possibility that subtle dynamical effects, absent in the FLRW models, provide alternative explanations of observed phenomena. The deduction also depends on the assumption that the EFE hold at cosmological scales - which may not be true: maybe for example some form of scalar-tensor theory should be used. More generally, an assessment of the reliability of a variety of cosmological inferences requires detailed study of a larger space of cosmological models.

Historical Epochs

The SM's account of the evolution of the matter and radiation in the universe reflects the dynamical effect of expansion. Consider a cube of spacetime in the early universe, filled with matter and radiation. The dynamical effects of the universe's expansion are locally the same as slowly stretching the cube. For some stages of evolution the contents of the cube interact sufficiently quickly that they reach and stay in local thermal equilibrium as the cube changes volume. (Because of isotropy, equal amounts of matter and radiation enter and leave the cube from neighboring cubes.) But when the interactions are too slow compared to the rate of expansion, the cube changes volume too rapidly for equilibrium to be maintained. As a result, particle species "freeze out" and decouple, and entropy increases. Without a series of departures from equilibrium, cosmology would be boring the system would remain in equilibrium with a state determined solely by the temperature, without a trace of things past. The rate of expansion of the cube varies with cosmic time. Because radiation, matter, and a cosmological constant term (or dark energy) dilute with expansion at different rates, an expanding universe naturally falls into separate epochs, characterized by different expansion rates.

There are several distinctive epochs in the history of the universe, according to the SM, including the following:

- Quantum Gravity: Classical general relativity is expected to fail at early times, when quantum effects will be crucial in describing the gravitational degrees of freedom. There is considerable uncertainty regarding physics at this scale.
- Inflation: A period of exponential, quasi-De Sitter expansion driven by an "inflaton" field (or fields), leading to a uniform, almost flat universe with Gaussian linear nearly scale invariant density perturbations. During inflation pre-existing matter and radiation are rapidly diluted; the universe is repopulated with

matter and energy by the decay of the inflaton field into other fields at the end of inflation (“re-heating”).

- Big Bang Nucleosynthesis: At t≈1 second, the constituents of the universe include neutrons, protons, electrons, photons, and neutrinos, tightly coupled and in local thermal equilibrium. Synthesis of light elements occurs during a burst of nuclear interactions that transpire as the universe falls from a temperature of roughly 10^9 K to 10^8 K after neutrinos fall out of equilibrium and consequent onset of neutron decay. The predicted light-element abundances depend on physical features of the universe at this time, such as the total density of baryonic matter and the baryon to photon ratio. Agreement between theory and observation for a specific baryon to photon ratio is a great success of the SM.

- Decoupling: As the temperature drops below ≈4,000K, electrons become bound in stable atoms, and photons decouple from the matter with a black-body spectrum. With the expansion of the universe, the photons cool adiabatically but retain a black-body spectrum with a temperature $T \propto 1/R$. This “cosmic background radiation” (CBR) has been aptly called the cosmic Rosetta stone because it carries so much information about the state of the universe at decoupling.

- Dark Ages: After decoupling, baryonic matter consists almost entirely of neutral hydrogen and helium. Once the first generation of stars form, the dark ages come to an end with light from the stars, which re-ionize the universe.

- Structure Formation: Cold dark matter dominates the early stages of the formation of structure. Dark matter halos provide the scaffolding for hierarchical structure formation. The first generation of stars aggregate into galaxies, and galaxies into clusters. Massive stars end their lives in supernova explosions and spread through space heavy elements that have been created in their interiors, enabling formation of second generation stars surrounded by planets.

- Dark Energy Domination: Dark energy (or a non-zero cosmological constant) eventually comes to dominate the expansion of the universe, leading to accelerated expansion. This expansion will be never-ending if the dark energy is in fact a cosmological constant.

Status of the Standard Model

The development of a precise cosmological model compatible with the rich set of cosmological data currently available is an impressive achievement. Cosmology clearly relies very heavily on theory; the cosmological parameters that have been the target of observational campaigns are only defined given a background model. The strongest case for accepting the SM rests on the evidence in favor of the underlying physics, in concert with the over determination of cosmological parameters. The SM includes several free parameters, such as the density parameters characterizing the abundance of different types of matter, each of which can be measured several ways. These methods have distinctive theoretical assumptions and sources of error. For example, the abundance of deuterium produced during big bang nucleosynthesis depends sensitively on the baryon density. Nucleosynthesis is described using well-tested nuclear physics, and the light element abundances are frozen in within the "first three minutes". The amplitudes of the acoustic peaks in the CMB angular power spectrum depend on the baryon density at the time of decoupling. Current measurements fix the baryon density to an accuracy of one percent, and the values determined by these two methods agree within observational error. This agreement is one of many consistency checks for the SM. There are important discrepancies, such as that between local versus global measurements of the Hubble parameter H_o. The significance and further implications of these discrepancies is not clear.

The SM from nucleosynthesis on can be regarded as well supported by many lines of evidence. The independence and diversity of the measurements provides some assurance that the SM will not be undermined by isolated theoretical mistakes or undetected sources of systematic error. But the SM is far from complete, and there are three different types of significant open issues.

First, we do not understand three crucial components of the SM that require new physics. We do not have a full account of the nature, or underlying dynamics, of dark matter, dark energy or the inflaton field. These are well-recognized problems that have inspired active theoretical and observational work.

The second set of open questions regards structure formation. While the account of structure formation matches several significant observed features, such as the correlations among galaxies in large scale surveys, there are a number of open questions about how galaxies form. Many of these, such as the cusp-core problem, and the dark halos problem (a great many more small dark halos are predicted around galaxies than observed) regard features of galaxies on relatively small scales, which require detailed modeling of a variety of astrophysical processes over an enormous dynamical range. This is also a very active area of research, driven in particular by a variety of new lines of observational research and large-scale numerical simulations.

The third and final set of open issues regards possible observations that would show that the SM is substantially wrong. Any scientific theory should be incompatible with at least some observations, and that is the case for the SM. In the early days of relativistic cosmology, the universe was judged to be younger than some stars or globular clusters. This conflict arose due to a mistaken value of the Hubble constant. There is currently no such age problem for the SM, but obviously discovering an object older than 13.7 Gyr would force a major re-evaluation of current cosmological models. Another example would be if there was not a dipole in matter number counts that agrees with the CMB dipole.

Local vs. Global Interplay in Cosmology

Although cosmology is generally seen as fitting into the general physics paradigm of everything being determined in a bottom up manner, there is another tradition that sees the effect of the global on the local in cosmology.

The traditional issues of this kind are:

- Mach's Principle: The idea that the origin of inertia is due to the very distant matter in the universe nowadays understood as being due to the fact that the vorticity ω of the universe is very low at present (it could have been otherwise).
- Olber's Paradox: The issue of why the sky is dark at night, resolved by evolution of the universe together with the redshift factor of about 1000 since the surface of last scattering (which determines that the temperature of the night sky is the 2.73K of the CMB everywhere except for the small fraction of the sky covered by stars and galaxies).
- The Arrow of Time: Where does the arrow of time come from, if the underlying physics is time symmetric? This has to be due to special initial conditions at the start of the universe. This is related to the Sommerfeld outgoing radiation condition and Penrose's Weyl curvature hypothesis.

In each case, global boundary conditions have an important effect on local physics. More recent ones relate to:

- Nucleosynthesis, where the course of nuclear reactions is determined by the $T(t)$ relation that is controlled by cosmological evolution (the temperature T being a coarse grained variable with evolution determined by the average density ρ of matter in the universe through the Friedmann equation).
- Structure formation due to gravitational instability which is affected crucially by the expansion of the universe, which turns what would have been an exponential growth of inhomogeneity (in a static universe) to a power law growth. It is because of this effect that studies of structure such as the BAO and CMB anisotropies give us strong limits on the parameters of the background model.
- The Anthropic Principle, whereby large-scale conditions in the universe (such as the value of the cosmological constant and the initial amplitude of inhomogeneities in the early universe) provide local conditions suitable for life to come into being.

Underdetermination

Many philosophers hold that evidence is not sufficient to determine which scientific theory we should choose. Scientific theories make claims about the natural world that extend far beyond what can be directly established through observations or experiments. Rival theories may fare equally well with regard to some body of data, yet give quite different accounts of the world. Philosophers often treat the existence of such rivals as inevitable: For a given theory, it is always possible to construct rival theories that have "equally good fit" with available data. Duhem gave an influential characterization of the difficulty in establishing physical theories conclusively, followed a half century later by Quine's arguments for a strikingly general version of underdetermination. The nature of this proposed under determination of theory by evidence, and appropriate responses to it, have been central topics in philosophy of science. Although philosophers have identified a variety of distinct senses of underdetermination, they have generally agreed that underdetermination poses a challenge to justifying scientific theories.

There is a striking contrast with discussions of underdetermination among scientists, who often emphasize instead the enormous difficulty in constructing compelling rival theories. This contrast reflects a disagreement regarding how to characterize the empirical content of theories. Suppose that the empirical content of theory consists of a set of observational claims implied by the theory. Philosophers then take the existence of rival theories to be straightforward. Van Fraassen, for example, defines a theory as "empirically adequate" if what it says about observable phenomena is true, and argues that for any successful theory there are rival theories that disagree about theoretical claims. If we demand more of theories than empirical adequacy in this sense, it is possible to draw distinctions among theories that philosophers would regard as underdetermined. Furthermore, even when scientists do face a choice among competing theories, they are almost never rivals in the philosopher's sense. Instead, they differ in various ways: intended domain of applicability, explanatory scope, importance attributed to particular problems, and so on.

The scientists' relatively dismissive attitude towards alleged underdetermination threats may be based on a more demanding conception of empirical success. Scientists demand much more of their theories than mere compatibility with some set of observational claims: they must fit into a larger explanatory scheme, and be compatible with other successful theories. Given a more stringent account of empirical success it is much more challenging to find rival theories.

One aspect of underdetermination is of more direct relevance to scientific debates: current theories may be indistinguishable, within a restricted domain, from a successor theory, even though the successor theory makes different predictions for other domains. This raises the question of how far we can rely on extrapolating a theory to a new domain. For example, despite its success in describing objects moving with low relative velocities in a weak gravitational field, where it is nearly indistinguishable from general relativity, Newtonian gravity does not apply to other regimes. The obstacles to making such reliable inferences reflect the specific details of particular domains of inquiry. Below we will focus on the obstacles to answering theoretical questions in cosmology due to the structure of the universe and our limited access to phenomena.

Underdetermination in Cosmology

Given the grand scope of cosmology, one might expect that many questions must remain unresolved. Basic features of the SM impose two fundamental limits to the ambitions of cosmological theorizing. First, the finitude of the speed of light ensures that we have a limited observational window on the universe due to existence of the visual horizon, representing the most distant matter from which we can receive and information by electromagnetic radiation, and the particle horizon, representing the most distant matter with which we can have had any causal interaction (matter up to that distance can influence what we see at the visual horizon). Recent work has precisely characterized what can be established via idealized astronomical observations, regarding spacetime geometry within, or outside, our past light cone (the observationally accessible region). Second, in addition to enormous extrapolations of well-tested physics in the SM, cosmologists have explored speculative ideas in physics that can only be tested through their implications for cosmology; the energies involved are too high to be tested by any accelerator on Earth. Ellis has characterized these speculative aspects of cosmology as falling on the far side of a “physics horizon”.

Global Structure

To what extent can observations determine the spacetime geometry of the universe directly? The question can be posed more precisely in terms of the region that is, in principle, accessible to an observer at a location in spacetime p—the *causal past*, $J^-(p)$, of that point. This set includes all regions of spacetime from which signals traveling at or below the speed of light can reach pp. What can observations confined to $J^-(p)$, assuming that GR is valid, reveal about the spacetime geometry of $J^-(p)$ itself, and the rest of spacetime?

The observational cosmology program clarifies the extent to which a set of ideal observations can determine the spacetime geometry directly with minimal cosmological assumptions. (By contrast, the standard approach starts by assuming a background cosmological model and then finding an optimal parameter fit.) Roughly put, the ideal data set consists of a set of astrophysical objects that can be used as standard candles and standard rulers. If the intrinsic properties and evolution of a variety of sources are given, observations can directly determine the area (or luminosity) distance of the sources, and the distortion of distant images determines lensing effects. These observations thus directly constrain the spacetime geometry of the past light cone $C^-(p)$. Number counts of discrete sources (such as galaxies or clusters) can be used to infer the total amount of baryonic matter, again granting various assumptions. Ellis proved the remarkable result that an appropriate idealized data set of this kind is sufficient, if we grant that EFE hold, to fully fix the spacetime geometry and distribution of matter on the past light cone $C^-(p)$ and from that, in the causal past $J^-(p)$ of the observation point pp. Observers do not have access to anything like the ideal data set, obviously, and in practice cosmologists face challenges in understanding the nature of sources and their evolution with sufficient clarity that they can be used to determine spacetime geometry, so this is the ideal situation.

What does $J^-(p)$ reveal about the rest of spacetime? In classical GR, we would not expect the physical state on to determine that of other regions of spacetime—even the causal past of a point just to the future of p. There are some models in which $J^-(p)$ does reveal more: "small universe" models are closed models with a finite maximum length in all directions that is smaller than the visual horizon. Observers in such a model would be able to "see around the universe" in all directions, and establish some global properties via direct observation because they would be able to see all matter that exists.

Unless this is the case, the causal past for a single observer, and even a collection of causal pasts, place very weak constraints on the global properties of spacetime. The global properties of a spacetime characterize its causal structure, such as the presence or absence of singularities. General relativity tolerates a wide variety of global properties, since EFE impose only a local constraint on the spacetime geometry. One way to make this question precise is to consider whether there are any global properties shared by spacetimes that are constructed as follows. For a given spacetime, construct an indistinguishable counterpart that includes the collection of causal pasts $J^-(p)$ for all points in the original spacetime. The constructed spacetime is indistinguishable from the first, because for any observer in the first spacetime there is a "copy" of their causal past in the counterpart. It is possible, however, to construct counterparts that do not have the same global properties as the original spacetime. The property of having a Cauchy surface, for example, need not be shared by an indistinguishable counterpart. More generally, the only properties that are guaranteed to hold for an indistinguishable counterpart are those that can be established based on the causal past of a single point.

Establishing FLRW Geometry

The case of global spacetime geometry is not a typical instance of underdetermination of theory by evidence, as discussed by philosophers, for two reasons. First, this whole discussion assumes that classical GR holds; the question regards discriminating among models of a given theory, rather than a choice among competing theories. Second, these results establish that all observations available to us that are compatible with a given spacetime, with some appealing global property, are equally compatible with its indistinguishable counterparts. But as is familiar from more prosaic examples of the problem of induction, evidence of past events is compatible, in a similar sense, with many possible futures. Standard accounts of inductive inference aim to justify some expectations about the future as more reasonable, e.g., those based on extending past uniformities. The challenge in this case is to articulate an account of inductive inferences that justifies accepting one spacetime over its indistinguishable counterparts.

As a specific instance of this challenge, consider the status of the cosmological principle, the global symmetry assumed in the derivation of the FLRW models. The results above show that all evidence available to us is equally compatible with models in which the cosmological principle does or does not hold. One might take the principle as holding *a priori*, or as a pre-condition for cosmological theorizing. A recent line of work aims to justify the FLRW models by appealing to a weaker general principle in conjunction with theorems relating homogeneity and isotropy. Global isotropy around every point implies global homogeneity, and it is natural to seek a similar theorem with a weaker antecedent formulated in terms of observable quantities. The Ehlers-Geren-Sachs theorem shows that if all geodesic fundamental observers in an expanding model find that freely propagating background radiation is exactly isotropic, then their spacetime is an FLRW model. If our causal past is "typical", observations along our worldline will constrain what other observers should see. This is often called the Copernican principle—namely, no point pp is distinguished from other points qq by any spacetime symmetries or lack thereof (there are no "special locations"). There are indirect ways of testing this principle empirically: The Sunyaev-Zel'dovich effect can be used to indirectly measure the isotropy of the CBR as observed from distant points. Other tests are direct tests with a good enough set of standard candles, and an indirect test based on the time drift of cosmological redshift. This line of work provides an empirical argument that the observed universe is well-approximated by an FLRW model, thus changing that assumption from a philosophically based starting point to an observationally tested foundation.

Physics Horizon

The Standard Model of particle physics and classical GR provide the structure and framework for the SM. But cosmologists have pursued a variety of questions that extend beyond these core theories. In these domains, cosmologists face a form of underdetermination: Should a phenomena be accounted for by extending the core theories, or by changing physical or astrophysical assumptions?

The Soviet physicist Yakov Zel'dovich memorably called the early universe the "poor man's accelerator", because relatively inexpensive observations of the early universe may reveal features of high-energy physics well beyond the reach of even the most lavishly funded earth-bound accelerators. For many aspects of fundamental physics, including quantum gravity in particular, cosmology provides the only feasible way to assess competing ideas. This ambitious conception of cosmology as the sole testing ground for new physics extends beyond the standard model of particle physics (which is generally thought to be incomplete, even though there are no observations that contradict it). Big bang nucleosynthesis, for example, is an application of well-tested nuclear physics to the early universe, with scattering cross-sections and other relevant features of the physics fixed by terrestrial experiments. While working out how nuclear physics applied in detail required substantial effort, there was little uncertainty regarding the underlying physics. By contrast, in some domains cosmologists now aim to explain the universe's history while at the same time evaluating new physics used in constructing it.

This contrast can be clarified in terms of the "physics horizon", which delimits the physical regime accessible to terrestrial experiments and observations, roughly in terms of energy scales associated with different interactions. The horizon can be characterized more precisely for a chosen theory, by specifying the regions of parameter space that can be directly tested by experiments and observations. Aspects of cosmological theories that extend past the physics horizon cannot be independently tested through non-cosmological experiments or observations; the only empirical route to evaluating these ideas is through their implications for cosmology. (This is not to deny that there may be strong theoretical grounds to favor particular proposals, as extensions of the core theories).

Cosmological physics extending beyond the physics horizon faces an underdetermination threat due to the lack of independent lines of relevant evidence. The case of dark matter illustrates the value of such independent evidence. Dark matter was first proposed to account for the dynamical behavior of galaxy clusters and galaxies, which could not be explained using Newtonian gravitational theory with only the luminous matter observed. Dark matter also plays a crucial role in accounts of structure formation, as it provides the scaffolding necessary for baryonic matter to clump, without conflicting with the uniformity of the CMB. Both inferences to the existence of dark matter rely on gravitational physics, raising the question of whether we should take these phenomena as evidence that our gravitational theory fails, rather than as evidence for a new type of matter. There is an active research program (MOND, for *Mo*dified *N*ewtonian *D*ynamics) devoted to accounting for the relevant phenomena by modifying gravity. Regardless of one's stance on the relative merits of MOND vs. dark matter (obviously MOND needs to be extended to a relativistic theory), direct evidence of existence of dark matter, or indirect evidence via decay products, would certainly reshape the debate. Efforts have been underway for some time to find dark matter particles through direct interactions with a detector, mediated by the weak force. A positive outcome of these experiments would provide evidence of the existence of dark matter that does not depend upon gravitational theory.

Such independent evidence is not available for two prominent examples of new physics motivated by discoveries in cosmology. "Dark energy" was introduced in studies of structure formation, which employed a non-zero cosmological constant to fit observational constraints (the ΛCDM models). Subsequent observations of the redshift-distance relation, using supernovae (type I_a) as a standard candle, led to the discovery that the expansion of the universe is accelerating. (For $\ddot{R} < 0$ in an FLRW model, there must be a contribution that appears in equations like a positive Λ term.) Rather than treating these observations as simply determining the value of a parameter in the SM, many cosmologists have developed phenomenological models of "dark energy" that leads to an effective Λ. Unlike dark matter, however, the properties of dark energy insure that any attempt at non-cosmological detection would be futile: the energy density is so small, and uniform, that any local experimental study of its properties is practically impossible. Furthermore these models are not based in well-motivated physics: They have the nature of 'saving the phenomena' in that they are tailored to fitting the cosmological observations by curve fitting.

Inflationary cosmology originally promised a powerful unification of particle physics and cosmology. The earliest inflationary models explored the consequences of specific scalar fields introduced in particle physics (the then supposed Higgs field for the strong interactions). Yet theory soon shifted to treating the scalar field responsible for inflation as the "inflaton" field, leaving its relationship to particle physics unresolved, and the promise of unification unfulfilled. If the properties of the inflaton field are unconstrained, inflationary cosmology is extremely flexible; it is possible to construct an inflationary model that matches any chosen evolutionary history of the early universe. Specific models of inflation, insofar as they specify the features of the field or fields driving inflation and its initial state, do have predictive content. In principle, cosmological observations could determine some of the properties of the inflaton field and so select among them. This could in principle then have implications for a variety of other experiments or observations; yet in practice the features of the inflaton field in most viable models of inflation guarantee that it cannot be detected in other regimes. The one exception to this is if the inflaton were the electroweak Higgs particle detected at the LHC. This remains a viable inflaton candidate, so testing if it is indeed the inflaton is an important task.

The physics horizon poses a challenge because one particularly powerful type of evidence—direct experimental detection or observation, with no dependence on cosmological assumptions—is unavailable for the physics relevant at earliest times (before inflation, and indeed even for baryosynthesis after inflation). Yet this does not imply that competing theories, such as dark matter vs. modified gravity, should be given equal credence. The case in favor of dark matter draws on diverse phenomena, and it has been difficult to produce a compelling modified theory of gravity, consistent with GR, that captures the full range of phenomena as an alternative to dark matter. Cosmology typically demands a more intricate assessment of background assumptions, and the degree of independence of different tests, in evaluating proposed extensions of the core theories.

Cosmic Variance

There is a distinctive form of underdetermination regarding the use of statistics in cosmology, due to the uniqueness of the universe. To compare the universe with the statistical predictions of the SM, we conceptualize it as one realization of a family of possible universes, and compare what we actually measure with what is predicted to occur in the ensemble of hypothetical models.

Origins of the Universe

Cosmology confronts a distinctive challenge in accounting for the origin of the universe. In most other branches of physics the initial or boundary conditions of a system do not call out for theoretical explanation. They may reflect, for example, the impact of the environment, or an arbitrary choice regarding when to cut off the description of a subsystem of interest. But in cosmology there are heated debates regarding what form a "theory of the initial state" should take, and what it should contribute to our understanding of the universe.

The Initial State

Contemporary cosmology at least has a clear target for a theory of origins: the SM describes the universe as having expanded and evolved over 13.7 billion years from an initial state where many physical quantities diverged. In the FLRW models, the cosmic time tt can be measured by the total proper time elapsed along the worldline of a fundamental observer, from the "origin" of the universe until the present epoch. Extrapolating backwards from the present, various quantities diverge as the cosmic time $t \rightarrow 0$ for example, $R(t) \rightarrow 0$ and the matter density goes to infinity. The worldlines of observers cannot be extended arbitrarily far into the past. Although there is no "first moment" of time, because the very concept of time breaks down as $\tau \rightarrow 0$, the age of the universe is the maximum length of these worldlines.

Singularity Theorems

The singularity theorems proved in the 60s show that the universe is finite to the past in a broad class of cosmological models. Past singularities, signaled by the existence of inextendible geodesics with bounded length, must be present in models with a number of plausible features. (Geodesics are the curves of extreme length through curved spacetime, and freely falling bodies follow timelike geodesics.) Intuitively, extrapolating backwards from the present, an inextendible geodesic reaches, within finite distance, an "edge" beyond which it cannot be extended. There is not a uniquely defined "cosmic time", in general, but the maximum length of these curves reflects the finite age of the universe. The singularity theorems plausibly apply to the observed universe, within the domain of applicability of general relativity. There are various related theorems differing in detail, but one common ingredient is an assumption that there is sufficient matter and energy present to guarantee that our past light cone refocuses. The energy density of the CMB alone is sufficient to justify this assumption. The theorems also require an energy condition: a restriction on the types of matter present in the model, guaranteeing that gravity leads to focusing of nearby geodesics.

The prediction of singularities is usually taken to be a deep flaw of GR. One potential problem with singularities is that they may lead to failures of determinism, because the laws "break down" in some sense. This concern only applies to some kinds of singularities, however. Relativistic spacetimes that are globally hyperbolic have Cauchy surfaces, and appropriate initial data posed on such surfaces fix a unique solution throughout the spacetime. Global hyperbolicity does not rule out the existence of singularities, and in particular the FLRW models are globally hyperbolic in spite of the existence of an initial singularity. The threat to determinism is thus more qualified: the laws do not apply "at the singularity itself" even though the subsequent evolution is fully deterministic, and there are some types of singularities that pose more serious threats to determinism.

Another common claim is that the presence of singularities establish that GR is incomplete, since it fails to describe physics "at the singularity". This is difficult to spell out fully without a local analysis of singularities, which would give precise meaning to talk of "approaching" or being "near" the singularity. In any case, it is clear that the presence of a singularity in a cosmological model indicates that spacetime, as described by GR, comes to an end: there is no way of extending the spacetime through the singularity, without violating mathematical conditions needed to insure that the field equations are well-defined. Any description of physical conditions "before the big bang" must be based on a theory that supersedes GR, and allows for an extension through the singularity.

There are two limitations regarding what we can learn about the origins of the universe based on the singularity theorems. First, although these results establish the existence of an initial singularity, they do not provide much guidance regarding its structure. The spacetime structure near a "generic" initial singularity has not yet been fully characterized. Partial results have been established for restricted classes of solutions; for example, numerical simulations and a number of theorems support the BKL conjecture, which holds that isotropic, inhomogeneous models exhibit a complicated form of chaotic, oscillatory behavior. The resulting picture of the approach to the initial singularity contrasts sharply with that in FLRW models. It is also possible to have non-scalar singularities.

Second, classical general relativity does not include quantum effects, which are expected to be relevant as the singularity is approached. Crucial assumptions of the singularity theorems may not hold once quantum effects are taken into account. The standard energy conditions do not hold for quantum fields, which can have negative energy densities. This opens up the possibility that a model including quantum fields may exhibit a "bounce" rather than collapse to a singularity. More fundamentally, GR's classical spacetime description may fail to approximate the description provided by a full theory of quantum gravity. According to recent work applying loop quantum gravity to cosmology, spacetime collapses to a minimum finite size rather than reaching a true singularity. On this account, GR fails to provide a good approximation in the region of the bounce, and the apparent singularity is an artifact. Classical spacetime "emerges" from a state to which familiar spacetime concepts do not apply. There are several accounts of the early universe, motivated by string theory and other approaches that similarly avoid the initial singularity due to quantum gravity effects.

Puzzling Features of the Initial State

In practice, cosmologists often take the physical state at the expected boundary of the domain of applicability of GR as the "initial state". (For example, this might be taken as the state specified on a spatial hypersurface at a very early cosmic time. However, the domain of applicability of GR is not well understood, given uncertainty about quantum gravity.) Projecting observed features of the universe backwards leads to an initial state with three puzzling features:

- Uniformity*:* The FLRW models have a finite particle horizon distance, much smaller than the scales at which we observe the CMB. Yet the isotropy of the CMB, among other observations, indicate that distant regions of the universe have uniform physical properties.
- Flatness: An FLRW model close to the "flat" model, with nearly critical density at some specified early time is driven rapidly away from critical density under FLRW dynamics if $\Lambda=0$ and $\rho+3p>0$. Given later observations, the initial state has to be very close to the flat model (or, equivalently, very close to critical density, $\Omega=1$) at very early times.
- Perturbations: The SM includes density perturbations that are coherent on large scales and have a specific amplitude, constrained by observations. It is challenging to explain both properties dynamically. In the standard FLRW models, the perturbations have to be coherent on scales much larger than the Hubble radius at early times.

On a more phenomenological approach, the gravitational degrees of freedom of the initial state could simply be chosen to fit with later observations, but many proposed "theories of initial conditions" aim to account for these features based on new physical principles.

Theories of the Initial State

There are three main approaches to theories of the initial state, all of which have been pursued by cosmologists since the late 60s in different forms. Expectations for what a theory of initial conditions should achieve have been shaped, in particular, by inflationary cosmology. Inflation provided a natural account of the three otherwise puzzling features of the initial state emphasized. Prior to inflation, these features were regarded as "enigmas", but after inflation, accounting for these features has served as an eligibility requirement for any proposed theory of the early universe.

The first approach aims to reduce dependence on special initial conditions by introducing a phase of attractor dynamics. This phase of dynamical evolution "washes away" the traces of earlier states, in the sense that a probability distribution assigned over initial states converges towards an equilibrium distribution. Misner introduced a version of this approach (his "chaotic cosmology program"), proposing that free-streaming neutrinos could isotropize an initially anisotropic state. Inflationary cosmology was initially motivated by a similar idea: a "generic" or "random" initial state at the Planck time would be expected to be "chaotic", far from a flat FLRW model. During an inflationary stage, arbitrary initial states are claimed to converge towards a state with the three features described above.

The second approach regards the initial state as extremely special rather than generic. Penrose, in particular, has argued that the initial state must be very special to explain time's arrow; the usual approaches fail to take seriously the fact that gravitational degrees of freedom are not excited in the early universe like the others. Penrose treats the second law as arising from a law-like constraint on the initial state of the universe, requiring that it has low entropy. Rather than introducing a subsequent stage of dynamical evolution that erases the imprint of the initial state, we should aim to formulate a "theory of initial conditions" that accounts for its special features. Penrose's conjecture is that the Weyl curvature tensor approaches zero as the initial singularity is approached; his hypothesis is explicitly time asymmetric, and implies that the early universe approaches an FLRW solution.

A third approach rejects the framework accepted by the other two proposals, and regards the "initial state" as a misnomer: It should instead by regarded as a "branch point" where our pocket universe separated off from a larger multiverse. (There are still, of course, questions regarding the initial state of the multiverse ensemble, if one exists.)

A dynamical approach, even if it is successful in describing a phase of the universe's evolution, arguably does not offer a complete solution to the problem of initial conditions: it collapses into one of the other two approaches. For example, an inflationary stage can only begin in a region of spacetime if the inflaton field and the geometry are uniform over a sufficiently large region, such that the stress-energy tensor is dominated by the potential term (implying that the derivative terms are small) and the gravitational entropy is small. There are other model-dependent constraints on the initial state of the inflaton field. One way to respond is to adopt Penrose's point of view, namely that this reflects the need to choose a special initial state, or to derive one from a previous expansion phase. The majority of those working in inflationary cosmology instead appeal to the third approach: Rather than treating inflation as an addition to standard big-bang evolution in a single universe, we should treat the observed universe as part of a multiverse.

The Limits of Science

Cosmology provokes questions about the limits of scientific explanation because it lacks many of the features that are present in other areas of physics. Physical laws are usually regarded as capturing the features of a type of system that remain invariant under some changes, and explanations often work by placing a particular event in larger context. Theories of the initial state cannot appeal to either idea: We have access to only one universe, and there is no larger context to appeal to in explaining its properties. This contrast between the types of explanation available in cosmology and other areas of physics has often led to dissatisfaction.

One challenge to establishing theories of the initial state is entirely epistemic. As emphasized, we lack independent experimental probes of physics at the relevant scales, so the extensions of core theories described above are only tested indirectly through their implications for cosmology. This limitation reflects contingent facts about the universe, namely the contrast between the energy scales of the early universe and those accessible to us, and does not follow from the uniqueness of the universe per se. Yet this limitation does not imply that it would be impossible to establish laws. There are cases in the history of physics, such as celestial mechanics, where confidence in a theory's laws is based primarily on successful application under continually improving standards of precision.

A further conceptual challenge regards whether it even makes sense to seek "laws" in cosmol. Laws are usually taken to cover multiple instances of some type of phenomena, or family of objects.

Competing philosophical analyses of laws of nature render different verdicts on the possibility of cosmological laws. Cosmological laws, if possible, differ from local physical laws in a variety of ways—they do not apply to subsystems of the universe, they lack multiple instances, and etc. Philosophical accounts of laws take different features to be essential to law-hood. For example, the influential Mill-Ramsey-Lewis account takes the laws to be axioms of the deductive system capturing some body of physical knowledge that optimally balances strength (the scope of derived claims) and simplicity (the number of axioms). It is quite plausible that a constraint on the initial state, such as Penrose's Weyl curvature hypothesis, would count as a law on this account. By contrast, accounts that take other features, such as governing evolution, as essential, reach the opposite verdict.

Finally, there are a number of conceptual pitfalls regarding what would count as an adequate "explanation" of the origins of the universe. What is the target of such explanations, and what can be used in providing an explanation? The target might be the state defined at the earliest time when extrapolations based on the SM can be trusted. The challenge is that this state then needs to be explained in terms of a physical theory, quantum gravity, whose basic concepts are still obscure to us. This is a familiar challenge in physics, where substantial work is often required to clarify how central concepts (such as space and time) are modified by a new theory. An explanation of origins in this first sense would explain how it is that classical spacetime emerges from a quantum gravity regime. While any such proposals remain quite speculative, the form of the explanation is similar to other cases in physics: What is explained is the applicability of an older, less fundamental theory within some domain. Such an explanation does not address ultimate questions regarding why the universe exists—instead, such questions are pushed back one step, into the quantum gravity regime.

Anthropic Reasoning and Multiverse

Anthropic Reasoning

The physical conditions necessary for our existence impose a selection effect on what we observe. The significance of this point for cosmological theorizing is exemplified by

Dicke's criticism of Dirac's speculative "large number hypothesis". Dirac noted the age of the universe expressed in terms of fundamental constants in atomic physics is an extremely large number (roughly 10^{39}), which coincides with other large, dimensionless numbers defined in terms of fundamental constants. Inspired by this coincidence, he proposed that the large numbers vary to maintain this order of magnitude agreement, implying that the gravitational "constant" G is a function of cosmic time. Dicke noted that creatures like us, made of carbon produced in an earlier generation of red giants and sustained by the light and heat of a main sequence star, can only exist within a restricted interval of cosmic times, and that Dirac's coincidence holds for observations made within this interval. Establishing that the coincidence holds at a randomly chosen tt would support Dirac's hypothesis, however slightly, but Dicke's argument shows that our evidence does not do so.

Dicke's reasoning illustrates how taking selection effects into account can mitigate surprise, and undermine the apparent implications of facts like those noted by. These facts reflect biases in the evidence available to us, rather than supporting his hypothesis. It is also clear that Dicke's argument is "anthropic" in only a very limited sense: His argument does not depend on a detailed characterization of human observers. All that matters is that we can exist at a cosmic time constrained by the time scales of stellar evolution.

How to account for selection effects, within a particular approach to confirmation theory, is one central issue in discussions of anthropic reasoning. This question is intertwined with other issues that are more muddled and contentious. Debates among cosmologists regarding "anthropic principles" ignited in the 70s, prompted by the suggestion that finely-tuned features of the universe such as the universe's isotropy can be explained as necessary conditions for the existence of observers. More recently, a number of cosmologists have argued that cosmological theories should be evaluated based on predictions for what a "typical" observer should expect to see. These ideas have dovetailed with work in formal epistemology. A number of philosophers have developed extensions of Bayesianism to account for "self-locating" evidence, for example. This kind of evidence includes indexical information characterizing an agent's beliefs about their identity and location. At present work in this area has not reached a consensus, and we will present a brief overview of some of the considerations that have motivated different positions in these debates.

In cosmology the most famous example of an "anthropic prediction" is Weinberg prediction for Λ. One part of Weinberg's argument is similar to Dicke's: He argued that there are anthropic bounds on Λ, due to its impact on structure formation. The existence of large, gravitationally bound structures such as galaxies is only possible if Λ falls within certain bounds. Weinberg went a step further than Dicke, and considered what value of Λ a "typical observer" should see. He assumed that observers occupy different locations within a multiverse, and that the value of Λ varies across different regions. Weinberg further argues that the prior probability assigned to different values of Λ

should be uniform within the anthropic bounds. Typical observers should expect to see a value close to the mean of the anthropic bounds, leading to Weinberg's prediction for Λ.

Essential to Weinberg's argument is an appeal to the principle of indifference, applied to a class of observers. We should calculate what we expect to observe, that is, as if we are a "random choice" among all possible observers. Bostrom argues that indifference-style reasoning is necessary to respond to the problem of "freak observers". As Bostrom formulates it, the problem is that in an infinite universe, any observation O is true for some observer (even if only for an observer who has fluctuated into existence from the vacuum). His response is that we should evaluate theories based not on the claim that some observer sees O, but on an indexical claim: That is, we make the observation O. He assumes that we are a "random" choice among the class of possible observers. (How to justify such a strong claim is a major challenge for this line of thought.) If we grant the assumption, then we can assign low probability to the observations of the "freak" observers, and recover the evidential value of O.

There are three immediate questions regarding this proposal. The first is called the "reference class" problem. The assignments of probabilities to events requires specifying how they are grouped together. Obviously, what is typical with respect to one reference class will not be typical with respect to another (compare, for example, "conscious observers" with "carbon-based life"). Second, the principle of indifference has been thoroughly criticized as a justification for probability in other contexts; what justifies the use of indifference in this case? Why should we take ourselves as "randomly chosen" among an appropriate reference class? The third problem reflects the intended application of these ideas: Bostrom and other authors in this line of work are particularly concerned with observes that may occupy an infinite universe. There is no proof that the universe is in fact infinite. These are all pressing problems for those who hold that the principle of indifference is essential to making cosmological predictions.

Furthermore, one way of implementing this approach leads to absurd consequences. The Doomsday Argument, for example, claims to reach a striking conclusion about the future of the human species without any empirical input. Suppose that we are "typical" humans, in the sense of having a birth rank that is randomly selected among the collection of all humans that have ever lived. We should then expect that there are nearly as many humans before and after us in overall birth rank. For this to be true, given current rates of population growth, there must be a catastrophic drop in the human population ("Doomsday") in the near future. The challenge to advocates of indifference applied to observers is to articulate principles that avoid such consequences, while still solving (alleged) problems such as that of freak observers.

In sum, one approach to anthropic reasoning aims to clarify the rules of reasoning applicable to predictions made by observers in a large or infinite universe. This line of work is motivated by the idea that without such principles we face a severe skeptical predicament, as observations would not have any bearing on the theory. Yet there is still not general agreement on the new principles required to handle these cases, which are of course not scientifically testable principles: They are philosophically based proposals. According to an alternative approach, selection effects can and should be treated within the context of a Bayesian approach to inductive inference. On this line of thought, “predictions” like those that Bostrom and others hope to analyze play no direct role in the evaluation of cosmological theories, so further principles governing anthropic reasoning are simply not necessary. There is much further work to be done in clarifying and assessing these (and other) approaches to anthropic reasoning.

Fine-tuning

Fine-tuning arguments start from a conflict between two different perspectives on certain features of cosmology (or other physical theories). On the first perspective, the existence of creatures like us seems to be sensitive to a wide variety of aspects of cosmology and physics. To be more specific, the prospects for life depend sensitively on the values of the various fundamental constants that appear in these theories. The SM includes about 10 constants, and the particle physics standard model includes about 20 more. Tweaking the SM, or the standard model of particle physics, by changing the values of these constants seems to lead to a barren cosmos. Focusing on the existence of “life” runs the risk of being too provincial; we don’t have a good general account of what physical systems can support intelligent life. Yet it does seem plausible that intelligence requires an organism with complex structural features, living in a sufficiently stable environment.

At a bare minimum, the existence of life seems to require the existence of complex structures at a variety of scales, ranging from galaxies to planetary systems to macro-molecules. Such complexity is extremely sensitive to the values of the fundamental constants of nature. From this perspective, the existence of life in the universe is fragile in the sense that it depends sensitively on these aspects of the underlying theory.

This view contrasts sharply with the status of the constants from the perspective of fundamental physics. Particle physicists typically regard their theories as effective field theories, which suffice for describing interactions at some specified energy scale. These theories include various constants, characterizing the relative strength of the interactions they describe, that cannot be further explained by the effective field theory. The constants can be fixed by experimental results, but are not derivable from fundamental physical principles. (If the effective field theory can be derived from a more fundamental theory, the value of the constants can in principle be determined by integrating out higher-energy degrees of freedom. But this merely pushes the question back one step: the constants appearing in the more fundamental theory are determined experimentally.) Similarly, the constants appearing in the SM are treated as contingent features of the universe. There is no underlying physical principle that sets, for example, the cosmological densities of different kinds of matter, or the value of the Hubble constant.

So features of our theories that appear entirely contingent, from the point of view of physics, are necessary to account for the complexity of the observed universe and the very possibility of life. The fine-tuning argument starts from a sense of unease about this situation: Shouldn't something as fundamental as the complexity of the universe be explained by the laws or basic principles of the theory, and not left to brute facts regarding the values of various constants? The unease develops into serious discomfort if the specific values of the constants are taken to be extremely unlikely: how could the values of all these constants be just right, by sheer coincidence?

In many familiar cases, our past experience is a good guide to when an apparent coincidence calls for further explanation. As Hume emphasized, however, intuitive assessments from everyday life of whether a given event is likely, or requires a further explanation, do not extend to cosmology. Recent formulations of fine-tuning arguments often introduce probabilistic considerations. The constants are "fine-tuned", meaning that the observed values are "improbable" in some sense. Introducing a well-defined probability over the constants would provide a response to Hume: Rather than extrapolating our intuitions, we would be drawing on the formal machinery of our physical theories to identify fine-tuning. Promising though this line of argument may be, there is not an obvious way to define physical probabilities over the values of different constants, or over other features of the laws. There is nothing like the structure used to justify physical probabilities in other contexts, such as equilibrium statistical mechanics.

There are four main responses to fine-tuning:

- Empiricist Denial*:* This response follows Hume in denying that a clear problem has even been identified. One form of this response challenges appeals to probability, undermining the claim that there are unexplained coincidences. Alternatively, fine-tuning is taken to reveal that the laws alone are not sufficient to account for some features of nature; these features are properly explained by the laws in conjunction with various contingent facts.
- Designer: Newton famously argued, for example, that the stability of the solar system provides evidence of providential design. For the hypothesized Designer to be supported by fine-tuning evidence, we require some way of specifying what kind of universe the Designer is likely to create; only such a specific Design hypothesis, based in some theory of the nature of the Designer, can offer an explanation of fine-tuning.
- New Physics: The fine-tuning can be eliminated by modifying physical theory in a variety of ways: altering the dynamical laws, introducing new constraints on the space of physical possibilities (or possible values of the constants of nature), etc.
- Multiverse: Fine-tuning is explained as a result of selection, from among a large space of possible universes (or multiverse).

Multiverse

The multiverse response replaces a single, apparently finely-tuned universe within an ensemble of universes, combined with an appeal to anthropic selection. Suppose that all possible values of the fundamental constants are realized in individual elements of the ensemble. Many of these universes will be inhospitable to life. In calculating the probabilities that we observe specific values of the fundamental constants, we need only consider the subset of universe compatible with the existence of complexity (or some more specific feature associated with life). If we have some way of assigning probabilities over the ensemble, we could then calculate the probability associated with our measured values. These calculations will resolve the fine-tuning puzzles if they show that we observe typical values for a complex (or life-permitting) universe.

Many cosmologists have argued in favor of a specific version of the multiverse called eternal inflation (EI). On this view, the rapid expansion hypothesized by inflationary cosmology continues until arbitrarily late times in some regions, and comes to an end (with a transition to slower expansion) in others. This leads to a global structure of "pocket" universes embedded within a larger multiverse.

On this line of thought, the multiverse should be accepted for the same reason we accept many claims about what we cannot directly observe—namely, as an inevitable consequence of an established physical theory. It is not clear, however, that EI is inevitable, as not all inflationary models, arguably including those favored by CMB observations, have the kind of potential that leads to EI. Accounts of how inflation leads to EI rely on speculative physics. Furthermore, if inflation does lead to EI, that threatens to undermine the original reasons for accepting inflation: rather than the predictions regarding the state produced at the end of inflation taken to provide evidence for inflation, EI seems to imply that, as Guth put it, in EI "anything that can happen will happen; in fact, it will happen an infinite number of times".

There have been two distinct approaches to recovering some empirical content in this situation. First, there may be traces of the early formation of the pocket universes, the remnants of collisions between neighboring "bubbles", left on the CMB sky. Detection of a distinctive signature that cannot be explained by other means would provide evidence for the multiverse. However, there is no expectation that a multiverse theory would generically predict such traces; for example, if the collision occurs too early the imprint is erased by subsequent inflationary expansion.

The other approach regards predictions for the fundamental constants, such as Weinberg's prediction of Λ discussed above. The process of forming the pocket universes is assumed to yield variation in the local, low-energy physics in each pocket. Predictions for the values of the fundamental constants follow from two things: (1) a specification of the probabilities for different values of the constant over the ensemble, and (2) a treatment of the selection effect imposed by restricting consideration to pocket universes with observers and then choosing a “typical” observer.

The aim is to obtain probabilistic predictions for what a typical observer should see in the EI multiverse. The assumption that the formation of pocket universes leads to variation in constants is just an assumption, which is not yet justified by a plausible, well-tested dynamical theory. The most widely discussed challenge in the physics literature is the “measure problem”: Roughly, how to assign “size” to different regions of the multiverse, as a first step towards assigning probabilities. It is difficult to define a measure because the EI multiverse is usually taken to be an infinite ensemble, lacking in the kinds of structure used in constructing a measure. On our view, these unmet challenges undercut the hope that the EI multiverse yields probabilistic predictions. And without such an account, the multiverse proposal does not have any testable consequences. If everything happens somewhere in the ensemble, then any potential observation is compatible with the theory.

Supposing that we grant a successful resolution of all these challenges, the merits of a multiverse solution of fine-tuning problems could then be evaluated by comparison with competing ideas. The most widely cited evidence in favor of a multiverse is Weinberg’s prediction for the value of Λ, discussed above. There are other proposals to explain the observed value of Λ; Wang, Zhu, and Unruh, for example, treat the quantum vacuum as extremely inhomogeneous, and argue that resonance among the vacuum fluctuations leads to a small Λ.

The unease many have about multiverse proposals are only reinforced by the liberal appeals to “infinities”. Many have argued, for example, that we must formulate an account of anthropic reasoning that applies to a truly infinite, rather than merely very large, universe. Claims that we occupy one of infinitely many possible pocket universes, filled with an infinity of other observers, rest on an enormous and speculative extrapolation. Such claims fail to take seriously the concept of infinity, which is not merely a large number. Hilbert emphasized that while infinity is required to complete mathematics, it does not occur anywhere in the accessible physical universe. One response is to require that infinities in cosmology should have a restricted use. It may be useful to introduce infinity as part of an explanatory account of some aspect of cosmology, as is common practice in mathematical models that introduce various idealizations. Yet this infinity should be eliminable, such that the explanation of the phenomena remains valid when the idealization is removed. Even for those who regard this demand as too stringent, there certainly needs to be more care in clarifying and justifying claims regarding infinities.

In sum, interest in the multiverse stems primarily from speculations about the consequences of inflation for the global structure of the universe. The main points of debate regard whether EI is a disaster for inflation, undermining the possibility of testing inflation at all, and how much predictions such as that for Λ lend credence to these speculations. Resolution of these questions is needed to decide whether the multiverse can be tested in a stronger sense, going beyond the special cases (such as bubble collisions) that may provide more direct evidence.

Religious Cosmology

A Religious cosmology (also mythological cosmology) is a way of explaining the origin, the history and the evolution of the universe based on the religious mythology of a specific tradition. Religious cosmologies usually include an act or process of creation by a creator deity or a larger pantheon.

Buddhism

In Buddhism, the universe comes into existence dependent upon the actions (karma) of its inhabitants. Buddhists posit neither an ultimate beginning nor final end to the universe, but see the universe as something in flux, passing in and out of existence, parallel to an infinite number of other universes doing the same thing.

The Buddhist universe consists of a large number of worlds which correspond to different mental states, including passive states of trance, passionless states of purity, and lower states of desire, anger, and fear. The beings in these worlds are all coming into existence or being born, and passing out of existence into other states, or dying. A world comes into existence when the first being in it is born, and ceases to exist, as such, when the last being in it dies. The universe of these worlds also is born and dies, with the death of the last being preceding a universal conflagration that destroys the physical structure of the worlds; then, after an interval, beings begin to be born again and the universe is once again built up. Other universes, however, also exist, and there are higher planes of existence which are never destroyed, though beings that live in them also come into and pass out of existence.

As well as a model of universal origins and destruction, Buddhist cosmology also functions as a model of the mind, with its thoughts coming into existence based on preceding thoughts, and being transformed into other thoughts and other states.

Hebrew Bible

The main Judeo-Christian religious text, the Bible, opens with the story of creation. The first two chapters of the Book of Genesis describe the creation of heaven and earth by God (called both Elohim and Yhvh) in six successive days.

- First day: God creates light ("Let there be light!")—the first divine command. The light is divided from the darkness, and "day" and "night" are named.

- Second day: God creates a firmament ("Let a firmament be!")—the second command—to divide the waters above from the waters below. The firmament is named "heaven" (shamayim).
- Third day: God commands the waters below to be gathered together in one place, and dry land to appear (the third command). "Earth" and "sea" are named. God commands the earth to bring forth grass, plants, and fruit-bearing trees (the fourth command).
- Fourth day: God creates lights in the firmament (the fifth command) to separate light from darkness and to mark days, seasons and years. Two great lights are made (most likely the Sun and Moon, but not named), and the stars.
- Fifth day: God commands the sea to "teem with living creatures", and birds to fly across the heavens (sixth command). He creates birds and sea creatures, and commands them to be fruitful and multiply.
- Sixth day: God commands the land to bring forth living creatures (seventh command); He makes wild beasts, livestock and "every thing that creepeth upon the earth". He then creates humanity in His "image" and "likeness" (eighth command). They are told to "be fruitful, and multiply, and fill the earth, and subdue it." The totality of creation is described by God as "very good."
- Seventh day: God, having completed the heavens and the earth, rests from His work, and blesses and sanctifies the seventh day.

Christianity

It is a tenet of Christian faith (Roman Catholic, Orthodox and Protestant) that God is the creator of all things from nothing, and has made human beings in the Image of God, who by direct inference is also the source of the human soul. In Chalcedonian Christology, Jesus is the Word of God, which was in the beginning and, thus, is uncreated, and hence is God, and consequently identical with the Creator of the world ex nihilo.

The New Testament claims that God created everything by the eternal Word, Jesus Christ his beloved Son. In him:

> "All things were created, in heaven and on earth, all things were created through him and for him. He is before all things, and in him all things hold together."

Mormon

Mormon cosmology draws from Biblical cosmology, but has many unique elements provided by Latter Day Saint movement founder Joseph Smith, Jr.

According to Mormon cosmology, there was a pre-existence, better described as a pre-mortal life, in which human spirits were literal children of heavenly parents.

Though their spirits were created, the essential "intelligence" of these spirits is considered eternal, and without beginning. During this pre-existence, two plans were said to have been presented, one championed by Lucifer (Satan) that would have involved loss of moral agency, and another championed by God the Father. When his plan was not accepted, Lucifer is said to have rebelled and taken a third of the hosts of heaven with him to the earth to serve as tempters. According to a plan of salvation as described by God the Father, Jesus would create the earth, under the direction of God the Father, as a place where humanity would be tested. After the resurrection all men and women except spirits that followed Lucifer and the sons of perdition would be assigned one of three degrees of glory. Within the highest degree, the Celestial Kingdom, there are three divisions, and those in the highest of these divisions would become gods and goddesses through a process called exaltation or "eternal progression". This would involve having spirit children and populating new worlds.

The Earth's creation, according to Mormon scripture, was not ex nihilo, but organized from existing matter. The faith teaches that this earth is just one of many inhabited worlds, and that there are many governing heavenly bodies, including a planet or star Kolob which is said to be nearest the throne of God. According to some Mormon sources, God the Father himself was once like a human, and lived on a planet with his own higher god.

Hinduism

The Hindu cosmology and timeline is the closest to modern scientific timelines and even more which might indicate that the Big Bang is not the beginning of everything but just the start of the present cycle preceded by an infinite number of universes and to be followed by another infinite number of universes.

The Rig Veda questions the origin of the cosmos in: "Neither being (sat) nor non-being was as yet. What was concealed? And where? And in whose protection? Who really knows? Who can declare it? Whence was it born, and whence came this creation? The devas were born later than this world's creation, so who knows from where it came into existence? None can know from where creation has arisen, and whether he has or has not produced it. He who surveys it in the highest heavens, he alone knows-or perhaps does not know."

The Rig Veda's view of the cosmos also sees one true divine principle self-projecting as the divine word, Vaak, 'birthing' the cosmos that we know, from the monistic Hiranyagarbha or Golden Womb. The Hiranyagarbha is alternatively viewed as Brahma, the creator who was in turn created by God, or as God (Brahman) himself. The universe is considered to constantly expand since creation and disappear into a thin haze after billions of years. An alternate view is that the universe begins to contract after reaching its maximum expansion limits until it disappears into a fraction of a millimeter. The creation begins anew after billions of years (Solar years) of non-existence.

The puranic view asserts that the universe is created, destroyed, and re-created in an eternally repetitive series of cycles. In Hindu cosmology, a universe endures for about 4,320,000,000 years (one day of Brahma, the creator or kalpa) and is then destroyed by fire or water elements. At this point, Brahma rests for one night, just as long as the day. This process, named pralaya (Cataclysm), repeats for 100 Brahma years (311 Trillion, 40 Billion Human Years) that represents Brahma's lifespan. It must be noted that Brahma is the creator but not necessarily regarded as God in Hinduism. He is mostly regarded as a creation of God/Brahman.

We are currently believed to be in the 51st year of the present Brahma and so about 156 trillion years have elapsed since He was born as Brahma. After Brahma's "death", it is necessary that another 100 Brahma years (311 Trillion, 40 Billion Years) pass unti a new Brahma is born and the whole creation begins anew. This process is repeated again and again, forever.

Brahma's day is divided in one thousand cycles (Maha Yuga, or the Great Year). Maha Yuga, during which life, including the human race appears and then disappears, has 71 divisions, each made of 14 Manvantara (1000) years. Each Maha Yuga lasts for 4,320,000 years. Manvantara is Manu's cycle, the one who gives birth and governs the human race.

Each Maha Yuga consists of a series of four shorter yugas, or ages. The yugas get progressively worse from a moral point of view as one proceeds from one yuga to another. As a result, each yuga is of shorter duration than the age that preceded it. The current Kali Yuga (Iron Age) began at midnight 17 February/18 February in 3102 BC in the proleptic Julian calendar.

Space and time are considered to be maya (illusion). What looks like 100 years in the cosmos of Brahma could be thousands of years in other worlds, millions of years in some other worlds and 311 trillion and 40 billion years for our solar system and earth.

Islam

Islam preaches that God, or Allah, created the universe, including Earth's physical environment and human beings. The highest goal is to visualize the cosmos as a book of symbols for meditation and contemplation for spiritual upliftment or as a prison from which the human soul must escape to attain true freedom in the spiritual journey to God. Islam elaborates on cosmology in many instances. The Quran in the following terms describes the expansion of the universe, "We have built the heaven with might, and we are Steadily Expanding it".

Jainism

Jain cosmology considers the loka, or universe, as an uncreated entity, existing since infinity, having no beginning or an end. Jain texts describe the shape of the universe as similar to a man standing with legs apart and arm resting on his waist. This Universe, according to Jainism, is narrow at the top, broad at the middle and once again becomes broad at the bottom.

Mahāpurāṇa of Ācārya Jinasena is famous for this quote: "Some foolish men declare that a creator made the world. The doctrine that the world was created is ill advised and should be rejected. If God created the world, where was he before the creation? If you say he was transcendent then and needed no support, where is he now? How could God have made this world without any raw material? If you say that he made this first, and then the world, you are faced with an endless regression."

Taoism

The cosmology of Taoism beliefs is a complex mixture of different beliefs. There is a "primordial universe" Wuji (philosophy), and Hongjun Laozu, water or qi. It transformed into Taiji and multiplied into everything. The Pangu legend tells a formless chaos coalesced into a cosmic egg. Pangu emerged (or woke up) and separated Yin from Yang with a swing of his giant axe, creating the Earth (murky Yin) and the Sky (clear Yang). To keep them separated, Pangu stood between them and pushed up the Sky. After Pangu died, he became everything.

References

- Cosmology-astronomy, science: britannica.com, Retrieved 02 May, 2019
- Big-Bang: newworldencyclopedia.org, Retrieved 28 June, 2019
- Cosmology: plato.stanford.edu, Retrieved 25 August, 2019
- Religious-cosmology: religion.wikia.org, Retrieved 26 January, 2019
- Extra-solar planets, Rep. Progr. Phys. 63, 1209 (2000)

Chapter 2

Cosmological Models

There are various models in cosmology that deal with the study of evolution and current aspects of cosmology. It includes de Sitter Cosmological model, Einstein-de Sitter model, concordance model. This chapter has been carefully written to provide an easy understanding of these cosmological models.

de Sitter Cosmological Model

A static cosmological model made isotropic and homogeneous by removing all mass from the universe. The resulting expanding universe had density of zero and constant curvature. In this model, the radius as a function of time is given by,

$$R \propto e^{Ht}$$

Where, H is the Hubble constant, equal to:

$$H = \sqrt{\frac{\Lambda}{3}}$$

Where, is Λ the cosmological constant.

Geometry of the de Sitter Universe

By making use of the fact that the de-Sitter metric corresponds to a hyperquadric in a five-dimensional flat space, it is shown that the three Robertson-Walker metrics for empty spacetime and positive cosmological constant, corresponding to 3-space of positive, negative and zero curvative, are geometrically equivalent. The 3-spaces correspond to intersections of the hyperquadric by hyperplanes, and the time-like geodesics perpendicular to them correspond to intersections by planes, in all three cases.

In Einstein's theory these solutions correspond to empty universes and are devoid of physical interest. If the real universe were such that $|\Lambda| \gg \kappa\rho c^2$ ($\kappa = 8\pi G / c^4$) (and this inequality seems highly unlikely on the basis of the observational data), an empty space solution (with matter treated as test particles) could be a good approximation to the actual geometry of the universe. A more plausible physical justification for the study of cosmological solutions of comes from Hoyle-Narlikar creation-field cosmology; if the density is required to remain constant this theory reduces to with $\Lambda = \kappa\rho c^2 / 2$. In any case the geometrical aspects of are sufficiently interesting to warrant attention apart from their possible relevance to physics. The Robertson-Walker metrics are of the form,

$$dt^2 - S^2(t)\left(1+\frac{k\overline{r}^2}{4}\right)^{-2}(d\overline{r}^2+\overline{r}^2(d\theta^2+\sin^2\theta d\varphi^2))$$

and equation $R_{\mu\nu} = \Lambda g_{\mu\nu}$ reduces to,

$$\dot{S}^2 = -k + \Lambda S^2 / 3$$

The curves of constant $(\overline{r}, \theta, \varphi)$ are time-like geodesics orthogonal to the space-like hypersurfaces of constant *t*. These space-like hypersurfaces have positive, negative or zero curvature according as,

k = +-1 or zero.

The solutions of $\dot{S}^2 = -k + \Lambda S^2 / 3$ are:

(a) Λ negative. Choose units so that Λ = -3,

k = -1 ; *S* = cos *t*

(b) Λ positive. Choose units so that Λ = 3.

$$\begin{cases} k = +-1; & S = \cosh t \\ k = 0; & S = \exp t \\ k = -1; & S = \sinh t \end{cases}$$

Make the following change of radial variable:

$$\begin{cases} k = +-1; & \tan(\rho/2) = \overline{r}/2 \\ k = 0; & \rho = \overline{r} \\ k = -1; & \tanh(\rho/2) = \overline{r}/2 \end{cases}$$

and we obtain the solutions $dt^2 - S^2(t)\left(1+\frac{k\overline{r}^2}{4}\right)^{-2}(d\overline{r}^2+\overline{r}^2(d\theta^2+\sin^2\theta d\varphi^2))$ in the form,

(a) $\dfrac{\Lambda = -3}{k = -1}$; $dt^2 - \cos^2 t(d\rho^2 + \sinh^2\rho d\Omega^2)$

$$\text{(b)} \begin{cases} \dfrac{\Lambda = 3}{k = +1}; & dt^2 - \cosh^2 t(d\rho^2 + \sin^2 \rho\, d\Omega^2) \\ k = 0; & dt^2 - e^{2t}(d\rho^2 + \rho^2 d\Omega^2) \\ k = -1; & dt^2 - \sinh^2 t(d\rho^2 + \sinh^2 \rho\, d\Omega^2) \end{cases}$$

The metric $dt^2 - e^{2t}(d\rho^2 + \rho^2 d\Omega^2)$ is the well-known de Sitter metric. It is, of course, equivalent to the static de Sitter metric,

$$(1 - r^2)\,d\tau^2 - \frac{dr^2}{1 - r^2} - r^2 d\Omega^2$$

the coordinate transformation connecting $dt^2 - e^{2t}(d\rho^2 + \rho^2 d\Omega^2)$ and $(1 - r^2)\,d\tau^2 - \frac{dr^2}{1 - r^2} - r^2 d\Omega^2$ being,

$$e^{2t} = e^{2\tau}(1 - r^2), \qquad \rho = re^{-\tau}$$

As is well known, the metric $(1 - r^2)\,d\tau^2 - \frac{dr^2}{1 - r^2} - r^2 d\Omega^2$ is that of the hyperquadric,

$$\eta_1{}^2 + \eta_2{}^2 + \eta_3{}^2 - \eta_4{}^2 + \eta_5{}^2 = r^2 - \eta_4{}^2 + \eta_5{}^2 = 1$$

in a flat 5-space with signature (+++-+), metric,

$$d\eta_1{}^2 + d\eta_2{}^2 + d\eta_3{}^2 - d\eta_4{}^2 + d\eta_5{}^2$$

The parameters $\eta_1 \ldots.. \eta_5$ are defined by,

$$\eta_1 = r\sin\theta\cos\theta, \quad \eta_2 = r\sin\theta\sin\varphi, \quad \eta_3 = r\cos\theta$$

$$\eta_1 \pm \eta_4 = e^{\pm\tau}(1 - r^2)^{1/2}$$

The identity $\eta_1{}^2 + \eta_2{}^2 + \eta_3{}^2 - \eta_4{}^2 + \eta_5{}^2 = r^2 - \eta_4{}^2 + \eta_5{}^2 = 1$ and the equality (apart from an overall minus sign) of $(1 - r^2)\,d\tau^2 - \frac{dr^2}{1 - r^2} - r^2 d\Omega^2$ and $d\eta_1{}^2 + d\eta_2{}^2 + d\eta_3{}^2 - d\eta_4{}^2 + d\eta_5{}^2$ are easily verified.

In fact, all three metrics are the metric of the hyperquadric in different coordinate systems. This equivalence is truly remarkable since in conjunction with the interpretation of the $d\rho = d\theta = d\varphi = 0$ lines as world-lines of matter the three metrics refer to completely different cosmological situations (respectively: 3-space positively curved (closed); flat; negatively curved/infinite in time with expansion preceded by contraction; infinite in time continually expanding; reaction at t = 0 followed by expansion).

Geometry of the Hyperquadric

Take a section of the hyperquadric $\eta_1^2+\eta_2^2+\eta_3^2-\eta_4^2+\eta_5^2=r^2-\eta_4^2+\eta_5^2=1$ defined by θ = const, φ = const. We get a two-dimensional subspace--a quadric in a flat 3-space with coordinates (r,η_4,η_5). We actually get only half the quadric since r takes only the positive values. We include the negative values by adopting the convention that $(-r,\theta,\varphi,t)$ means $(r,\pi-\theta,\pi+\varphi,t)$. The equation of the quadric is,

$$r^2-\eta_4^2+\eta_5^2=1$$

(a hyperboloid of one sheet), and the metric of the 3-space is,

$$dr^2-d\eta_4^2+d\eta_5^2$$

In terms of the parameters (t, ρ) given by,

$$\left.\begin{aligned} r &= \cosh t \sin \rho \\ \eta_4 &= \sinh t \\ \eta_5 &= \cosh t \sin \rho \end{aligned}\right\}$$

the metric $dr^2-d\eta_4^2+d\eta_5^2$ is,

$$\cosh^2 t . d\rho^2 - dt^2$$

Any plane through the origin of (r,η_4,η_5) -space can, by a suitable $SO(2,1)$ rotation, be taken to be one of the two planes,

or $$\left.\begin{aligned} \eta_5 &= 0 \quad (t=0) \\ \eta_4 &= 0 \quad (\rho=0) \end{aligned}\right\}$$

From the metric $\cosh^2 t . d\rho^2 - dt^2$ we readily obtain the geodesic equations,

$$\left.\begin{aligned} & dt^2/ds^2+\sinh t \cosh t (d\rho^2/ds^2)=0 \\ & d\rho^2/ds^2+2\tanh t\,(dt/ds)(d\rho/ds)=0 \end{aligned}\right\}$$

which are clearly satisfied by the curves t = 0 (p a linear function of s) and ρ = 0 (t a linear function of s), so that the curves obtained by intersection of the quadric and the planes $\left.\begin{aligned} \eta_5 &= 0 \quad (t=0) \\ \eta_4 &= 0 \quad (\rho=0) \end{aligned}\right\}$ are geodesics. Hence the intersection of the quadric by any plane through the origin of (r,η_4,η_5) -space is a geodesic. Conversely, every geodesic on the quadric lies in some plane through the origin of (r,η_4,η_5) -space.

We are, of course, interested not so much in the quadric $r^2 - \eta_4{}^2 + \eta_5{}^2 = 1$ as in the hyperquadric $\eta_1{}^2 + \eta_2{}^2 + \eta_3{}^2 - \eta_4{}^2 + \eta_5{}^2 = r^2 - \eta_4{}^2 + \eta_5{}^2 = 1$, which in terms of t and ρ has metric:

$$-dt^2 + \cosh t(d\rho^2 + \sin^2 \rho \, d\Omega^2) = 0$$

The geodesic equations for this metric reduce to,

$$\left.\begin{aligned} &dt^2 / ds^2 + \sinh t \cosh t (d\rho^2 / ds^2) = 0 \\ &d\rho^2 / ds^2 + 2\tanh t\,(dt/ds)(d\rho / ds) = 0 \end{aligned}\right\}$$ under the restriction to constant θ and φ.

Thus every geodesic of the quadric is also a geodesic of the hyperquadric.

Coordinate Systems on the Quadric

Parametrise the quadric,

$$r^2 - \eta_4{}^2 + \eta_5{}^2 = 1$$

by the coordinates (r, t) where,

$$t = \tanh^{-1}(\eta_4 / \eta_5)$$

The coordinate curves r = constant and t = constant are then, respectively, the intersections of the quadric with the planes parallel to the (η_4, η_5)-plane, and the planes $\eta_4 + K\eta_5 = 0$. The generators of the quadric are null lines. The two generators through $(r, \eta_4, \eta_5) = (\pm 1, 0, 0)$ are the intersections of the quadric with the planes $\eta_4 = \pm\eta_5$ (corresponding to $t = \pm\infty$). Thus the null generators of (1, 0, 0) and (-1,0, 0) are coordinate singularities.

From $r^2 - \eta_4{}^2 + \eta_5{}^2 = 1$ and $t = \tanh^{-1}(\eta_4 / \eta_5)$,

$$\left.\begin{aligned} \eta_4{}^2 &= (1 - r^2)\sinh^2 t \\ \eta_5{}^2 &= (1 - r^2)\cosh^2 t \end{aligned}\right\}$$

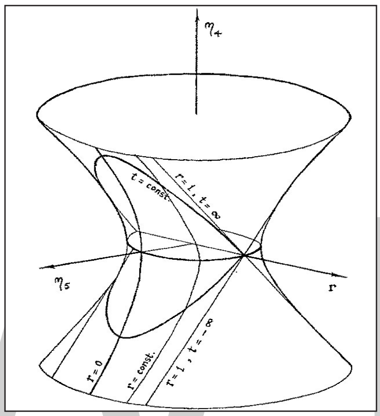
so that,

$$d\eta_5{}^2 - d\eta_4{}^2 = \frac{r^2 dr}{1 - r^2}^2 - (1 - r^2)dt^2$$

and the metric of the quadric in terms of (r, t) is,

$$\frac{dr}{1 - r^2}^2 - (1 - r^2)dt^2$$

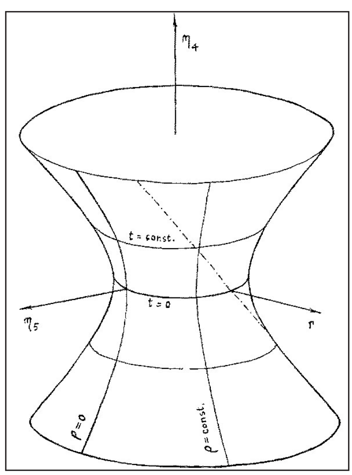
The corresponding metric of the hyperquadric is obtained simply by restoring the $r^2 d\Omega^2$ term, and we arrive at the static de Sitter metric $(1-r^2)\,d\tau^2 - \dfrac{dr^2}{1-r^2} - r^2 d\Omega^2$. Thus, we have a visualisation of the coordinate system we are employing when we say that the static de Sitter metric is the metric of a hyperquadric. Note that the 4-space of the static de Sitter metric is not the whole of the hyperquadric but only that portion bounded by null generators through the points ($\pm$1,0,0). Note that only one of the r = constant curves is a geodesic (r = 0)-world-lines of matter in the form of 'test particles' in a static de Sitter world will not be simply related to the coordinate system. In fact a static de Sitter world cannot contain a static distribution of test particles.



De Sitter's static universe.

If we extend the coordinate system we have set up to the regions $|r| > 1$ the r = constant curves become space-like and the t = constant curves become time –like. We therefore change the notation by what amounts to a rotation of the whole coordinate net through a right angle about the η_4 - axis. Define,

$$\rho = \tanh^{-1}(r/\eta_4)$$



Universe with k =-1.

And express the metric in terms of $(\eta_{5,}\rho)$. The manipulations are formally the same as those leading to $\frac{dr^2}{1-r^2} \; -(1-r^2)dt^2$ and we find that the metric is:

$$-\frac{d\eta_5^2}{\eta_5^2-1}+(\eta_5^2-1)d\rho^2$$

In terms of the new parameter t defined by:

$$\eta_5 = \cosh t$$

We get:

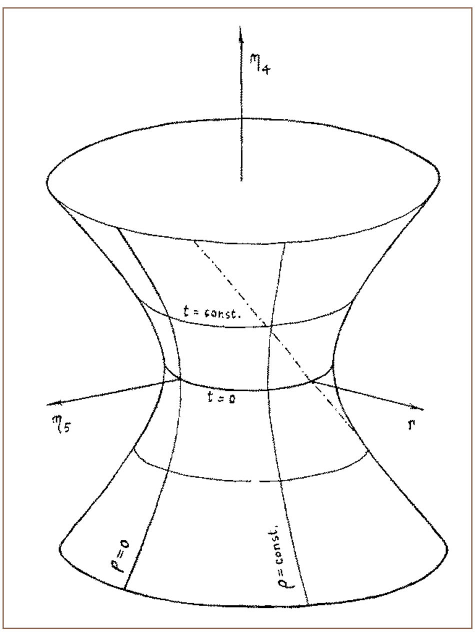
$$\text{-dt}^2 + \sinh^2\text{t}\, d\rho^2$$

We have only to include $r^2 d\Omega^2 = (\eta_5^2-1)\sinh^2\rho d\Omega^2 = \sinh^2\text{t}\sinh^2 \rho d\Omega^2$

to obtain the corresponding form for the metric of the hyperquadric:

$$\text{-dt}^2 + \sinh^2\text{t}(d\rho^2 + \sinh^2\rho)d\Omega^2$$

Thus we have obtained $\text{dt}^2 - \sinh^2\text{t}\ (d\rho^2 + \sinh^2\rho d\Omega^2)$ as the metric of the hyperquadric. Note that the Robertson-Walker 4-space corresponding to $\text{dt}^2 - \sinh^2\text{t}\ (d\rho^2 + \sinh^2\rho d\Omega^2)$ corresponds only to the part of the hyperquadric for which r/s > 1. Note that the p = constant curves in figure are intersections of the quadric by planes through the origin ($r + K\eta_4 = 0$), so are geodesics.



Universe with k = +1.

The most natural coordinate systems on the quadric are obtained by taking the coordinate curves to be intersection of the quadric by planes η_4 = constant and η_5+ Kr = 0 (i.e. through the symmetry axis of the quadric). Such a system covers the whole quadric without singularities.

Define,

$$\rho = \tan^{-1}(r / \eta_5)$$

and use (p, η_4) as coordinates. Then,

$$r^2 = (1+\eta_4^2)\sin^2\rho$$
$$\eta_5^2 = (1+\eta_4^2)\cos^2\rho$$
$$dr^2 + d\eta_5^2 = -\eta_4^2\, d\eta_4^2 / (1+\eta_4^2)^2 + (1+\eta_4)^2 d\rho^2$$

and the metric is therefore:

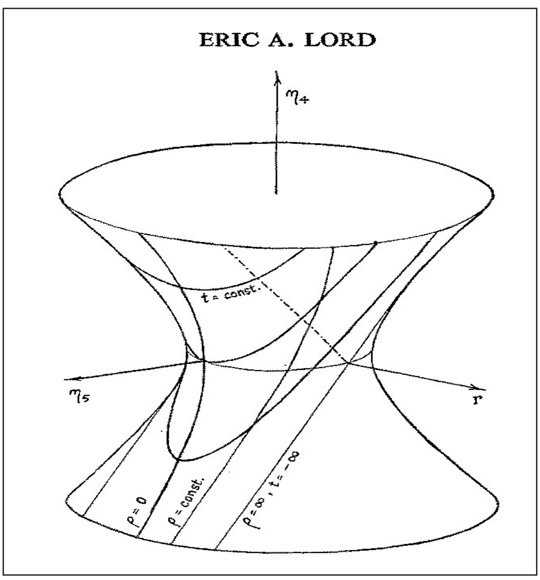
$$-\frac{d\eta_4^2}{1+\eta_4^2} + (1+\eta_4^2)d\rho^2$$

In terms of the new variable t defined by,

$$\eta_4 = \sinh t$$

We get,

$$-dt^2 + \cosh^2 t.d\rho^2$$



Universe with k=0.

For the corresponding metric of the hyperquadric we add on:

$$r^2 d\Omega^2 = (1+\eta_4^2)\sin^2\rho\, d\Omega^2 = \cosh^2 t \sin^2 \rho d\Omega^2$$

to obtain:

$$-dt^2 + \cosh^2 t(d\rho^2 + \sin^2 \rho d\Omega^2)$$

Note that it is not possible for a light ray to circumnavigate the closed 3-space in this model, since the null geodesics are represented by straight lines on the quadric.

The conventional (expanding, flat 3-space) de-Sitter metric corresponds to the intersection of the quadric by the planes $\eta_4 + \eta_5 = \text{constant}$ (which are parabolae in the $(r,\eta_4,\eta_5) - space$) and the planes $\eta_4 + \eta_5 + Kr = 0$ (which pass through the origin and therefore yield geodesics). We define the parameters p and t:

$$\rho = r / (\eta_4 + \eta_5)$$
$$e^t = \eta_4 + \eta_5$$

then,

$$r = \rho e^t$$

and

$$\eta_4 - \eta_5 = \frac{1-r^2}{\eta_4 + \eta_5} = (1-\rho^2 e^{2t})e^{-t}$$

Differentiating these expressions and forming the metric:

$$dr^2 + (d\eta_5 + d\eta_4)(d\eta_5 - d\eta_4)$$

we get,

$$-dt^2 + e^{2t} d\rho^2$$

to which we have to add $r^2 d\Omega^2 = \rho^2 e^{2t} d\Omega^2$ to obtain just $\dot{S}^2 = -k + \Lambda S^2 / 3$. Note that the whole 4-space corresponds to the part of the quadric above the plane $\eta_4 = r$. Note also that the null generator through (1,0,0) shown broken in figure is an event horizon in the sense that a generator through any point in the region r > 1 will never intersect the curve ρ = 0 (an observer moving along ρ = 0 will never observe events for which $|r| > 1$).

The Group SO(2,1)

The group of rotations SO(2,1) in the $(r,\eta_4,\eta_5) - space$ (metric (+-+)) leaves the quadric invariant (the quadric is the analogue of a 'sphere' in the space with indefinite metric). It is illuminating to demonstrate those particular SO(2,1) rotations that leave the various coordinate lines unchanged. Consider the coordinate system corresponding to the static de-Sitter metric and rotate about the r-axis:

$$\begin{pmatrix}\eta_4\\ \eta_5\end{pmatrix} \to \begin{pmatrix}\cosh\chi & \sinh\chi\\ \sinh\chi & \cosh\chi\end{pmatrix}\begin{pmatrix}\eta_4\\ \eta_5\end{pmatrix}$$

then,

$$\tanh t = \frac{\eta_4}{\eta_5} \to \frac{(\eta_4/\eta_5)+\tanh\chi}{1+(\eta_4/\eta_5)+\tanh\chi} = \tanh(1+\chi)$$

so that the ratation $\begin{pmatrix}\eta_4\\ \eta_5\end{pmatrix} \to \begin{pmatrix}\cosh\chi & \sinh\chi\\ \sinh\chi & \cosh\chi\end{pmatrix}\begin{pmatrix}\eta_4\\ \eta_5\end{pmatrix}$ correponds to:

$$r \to r,\ t \to t+\chi$$

This corresponds to the fact that the metric is static-there is no 'privileged' value of t (this is not apparent in figure of De-Sitter's static universe where t = 0 is a circle and other t = constant curves appear as ellipses. In fact they are all circles; the distortion is due to the impossibility of representing the geometry of space with metric (+-+) adequately in Euclidean 3-space). Applying to figure of universe with k=-1 a rotation about the η_5 -axis:

$$\begin{pmatrix}r\\ \eta_4\end{pmatrix} \to \begin{pmatrix}\cosh\chi & \sinh\chi\\ \sinh\chi & \cosh\chi\end{pmatrix}\begin{pmatrix}r\\ \eta_4\end{pmatrix}$$

we get,

$$\tanh\rho = \frac{r}{\eta_4} \to \tanh(\rho+\chi)$$

So the rotation $\begin{pmatrix}r\\ \eta_4\end{pmatrix} \to \begin{pmatrix}\cosh\chi & \sinh\chi\\ \sinh\chi & \cosh\chi\end{pmatrix}\begin{pmatrix}r\\ \eta_4\end{pmatrix}$ has the effect:

$$\rho \to \rho+\chi,\ t \to t$$

This illustrates the absence of a privileged p = constant curve. For Figure we carry out the rotation about the η_4 -axis:

$$\begin{pmatrix}r\\ \eta_5\end{pmatrix} \to \begin{pmatrix}\cos\theta & \sinh\theta\\ -\sin\theta & \cosh\theta\end{pmatrix}\begin{pmatrix}r\\ \eta_5\end{pmatrix}$$

Thus (recalling $\tan\rho = r/\eta_5, \sinh t = \eta_4$),

$$\rho \to \rho+\theta,\ t \to t$$

The absence of a privileged ρ = constant curve is demonstrated by carrying out a rotation about the common intersection of the planes:

$$\eta_4 + \eta_5 + Kr = 0$$

This transformation is of the form:

$$\begin{pmatrix} r \\ \eta_4 \\ \eta_5 \end{pmatrix} \to \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1+\alpha^2/2 & \alpha^2/2 \\ -\alpha & -\alpha^2/2 & 1-\alpha^2/2 \end{pmatrix} \begin{pmatrix} r \\ \eta_4 \\ \eta_5 \end{pmatrix}$$

We find that $\eta_4 + \eta_5$ (and therefore t) remains invariant and that:

$$\rho \to \frac{r}{\eta_4 + \eta_5} \to \frac{r + \alpha(\eta_4 + \eta_5)}{\eta_4 + \eta_5} = \rho + \alpha$$

Finally, we note that the shape of the coordinate net is unchanged by a time translation combined with a dilatation of 3-space. This transformation is simply rotation about the r-axis:

$$\begin{pmatrix} \eta_4 \\ \eta_5 \end{pmatrix} \to \begin{pmatrix} \cosh\chi & \sinh\chi \\ \sinh\chi & \cosh\chi \end{pmatrix} \begin{pmatrix} \eta_4 \\ \eta_5 \end{pmatrix}$$

so that,

$$t \to t + \chi, \quad \rho \to e^x \rho$$

The (3 + 2) de-Sitter Space

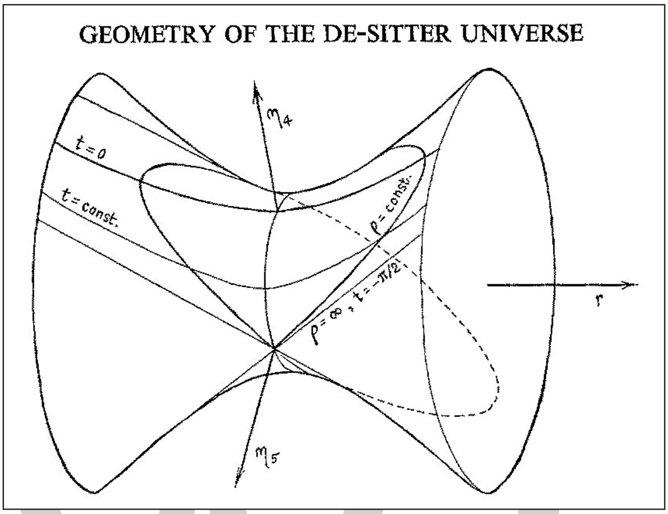
We have shown that the three metric $dt^2 - S^2(t)\left(1 + \frac{k\overline{r}^2}{4}\right)^{-2} (d\overline{r}^2 + \overline{r}^2(d\theta^2 + \sin^2\theta d\varphi^2))$ to $k = -1; S = cost$ correspond to different coordinate systems on the hyperquadric:

$$\eta^{AB}\eta_A\eta_B = 1 = r^2 + \eta_4^2 + \eta_5^2$$

In a 5-space with metric η_{AB} = dg(+++-+). To deal similarly with negative Λ we make use of the hyperquadric:

$$\zeta^{AB}\eta_A\eta_B = 1 = -r^2 + \eta_4^2 + \eta_5^2$$

in a 5-space with metric ζ_{AB} = dg(---++). The corresponding quadric in (r, η_4, η_5)-space is different from the previous case in that the roles of r and η_4 are interchanged. It is immediately apparent that this quadric contains closed time-like geodesics. This is usually taken to mean that negative A is non-physical. However, since a metric gives no information about the topological properties of the hyperquadric it is quite possible to consider it as a hypersurface with infinitely many sheets, so that in fact a closed time-like line passing once round the hyperboloid would return to a different event having the same (η_4, η_5, r)-coordinates as its initial event.



Universe with k = -1 and negative Λ.

We set up a coordinate system having the same form as that of Figure, defining (ρ, t) by:

$$\rho = \tanh^{-1}(r/\eta_5)$$
$$\eta_4 = \sin t$$

The metric for the hyperquadric becomes:

$$dt^2 - \cos^2 t d\rho^2$$

to which we must add:

$$-r^2 d\Omega^2 = -\cos^2 t \sinh^2 \rho d\Omega^2$$

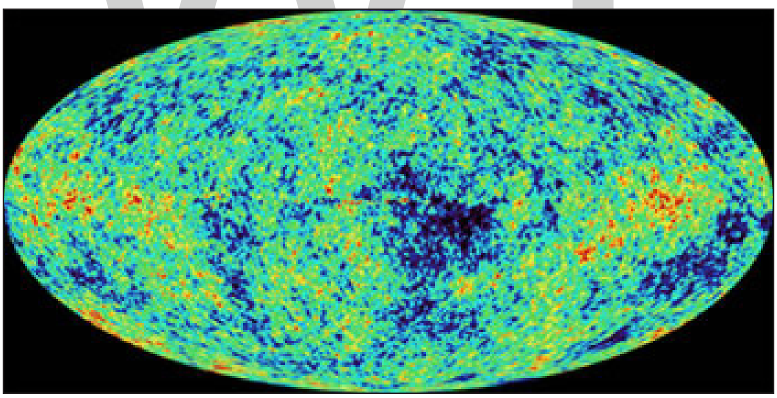
to obtain ($R_{\mu v} = Ag_{\mu v}$). We see that the question of closed time-like lines does not arise because the universe implied by ($R_{\mu v} = Ag_{\mu v}$) has a beginning and an end ($t = \pm\pi/2$) so the physically relevant portions of the time-like geodesics are only halves of the closed curves.

The Einstein-de Sitter Model

In 1932 Einstein and de Sitter proposed that the cosmological constant should be set equal to zero, and they derived a homogeneous and isotropic model that provides the separating case between the closed and open Friedmann models; i.e., Einstein and de Sitter assumed that the spatial curvature of the universe is neither positive nor negative but rather zero. The spatial geometry of the Einstein–de Sitter universe is Euclidean (infinite total volume), but space-time is not globally flat (i.e., not exactly the space-time of special relativity). Time again commences with a big bang and the galaxies recede forever, but the recession rate (Hubble's "constant") asymptotically coasts to zero as time advances to infinity. Because the geometry of space and the gross evolutionary properties are uniquely defined in the Einstein–de Sitter model, many people with a philosophical bent long considered it the most fitting candidate to describe the actual universe.

Concordance Model

The term 'concordance model' is used in cosmology to indicate the currently accepted and most commonly used cosmological model. It is important to identify a concordance model because the measurement of many astrophysical quantities (e.g. distance, radius, luminosity and surface brightness) depends upon the cosmological model used. Consequently, for ease of comparison if nothing else, the models assumed in different studies should at least be similar, if not identical.



WMAP observations of the cosmic microwave background radiation have been used to estimate the age, mass and composition of the Universe to unprecedented accuracy. The results form the current concordance model.

Currently, the concordance model is the Lambda CDM model (which includes cold dark matter and a cosmological constant). In this model the Universe is 13.7 billion years old and made up of 4% baryonic matter, 23% dark matter and 73% dark energy. The Hubble constant for this model is 71 km/s/Mpc and the density of the Universe is very close to the critical value for re-collapse. These values were derived from WMAP (Wilkinson Microwave Anisotropy Probe) satellite observations of the cosmic microwave background radiation.

References

- DeSitterCosmologicalModel, physics: scienceworld.wolfram.com, Retrieved 13 February, 2019
- Geometry-of-the-De-Sitter-Universe- 260833504: researchgate.net, Retrieved 14 April, 2019
- The-Einstein-de-Sitter-universe, science-cosmology-astronomy- 1069808: britannica.com, Retrieved 09 June, 2019
- Concordance-Model, cosmos: astronomy.swin.edu.au, Retrieved 17 March, 2019

Chapter 3

Dark Matter and Dark Energy

Dark matter is any substance that uses gravity to interact with visible matter such as stars and planets. Dark energy is an unknown form of energy that permeates through space which accelerates the expansion of the universe. The topics elaborated in this chapter will help in gaining a better perspective about dark matter and dark energy.

Dark Matter

Dark matter is a hypothetical invisible mass thought to be responsible for adding gravity to galaxies and other bodies.

Since the amount of visible material in galaxies can't account for their shapes, distributions, and movements, astrophysicists argue there must be a form of mass yet to be discovered. This mass doesn't appear to affect normal matter significantly in any other way - such as by absorbing or emitting photons - making it completely 'dark'.

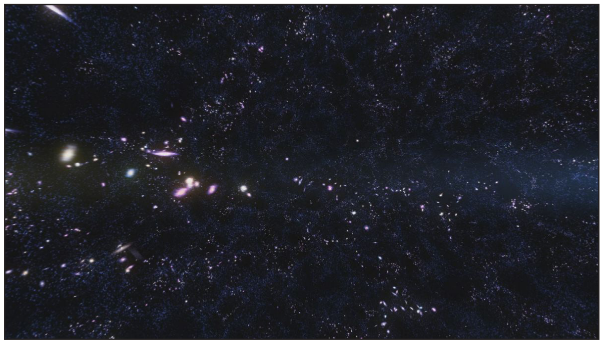
In spite of having so little to go on, researchers have continued to narrow down the characteristics of that 'missing matter' and have a few ideas what it might be. If it exists, it would make up 85 percent of the Universe's total material and roughly 25 percent of its energy.

Dark matter may be made of baryonic or non-baryonic matter. To hold the elements of the universe together, dark matter must make up approximately 80% percent of the universe. The missing matter could simply be more challenging to detect, made up of regular, baryonic matter.

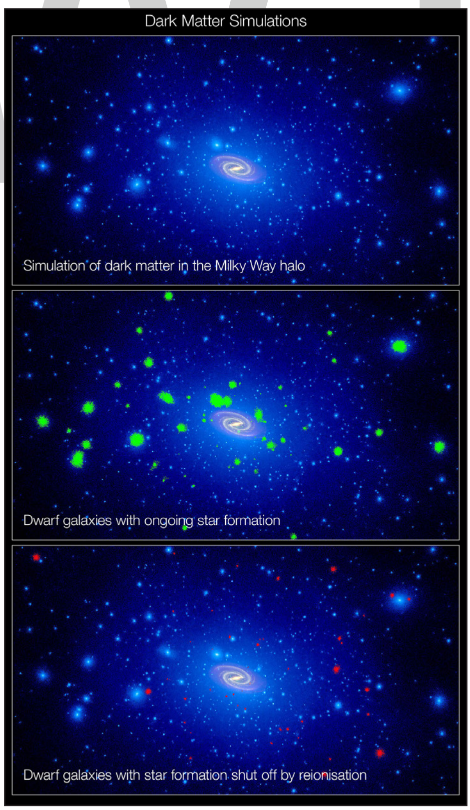
Potential candidates include dim brown dwarfs, white dwarfs and neutron stars. Supermassive black holes could also be part of the difference. But these hard-to-spot objects would have to play a more dominant role than scientists have observed to make up the missing mass, while other elements suggest that dark matter is more exotic.

Most scientists think that dark matter is composed of non-baryonic matter. The lead candidate, WIMPS (weakly interacting massive particles), have ten to a hundred times the mass of a proton, but their weak interactions with "normal" matter make them difficult to detect. Neutralinos, massive hypothetical particles heavier and slower than neutrinos, are the foremost candidate, though they have yet to be spotted.

Sterile neutrinos are another candidate. Neutrinos are particles that don't make up regular matter. A river of neutrinos streams from the sun, but because they rarely interact with normal matter, they pass through the Earth and its inhabitants. There are three known types of neutrinos; a fourth, the sterile neutrino, is proposed as a dark matter candidate. The sterile neutrino would only interact with regular matter through gravity.



Dark matter appears to be spread across the cosmos in a network-like pattern, with galaxy clusters forming at the nodes where fibers intersect. By verifying that gravity acts the same both inside and outside our solar system, researchers provide additional evidence for the existence of dark matter and dark energy.



These illustrations, taken from computer simulations, show a s warm of dark matter clumps around our Milky Way galaxy.

The smaller neutral axion and the uncharged photinos — both theoretical particles — are also potential placeholders for dark matter.

Several astronomical measurements have corroborated the existence of dark matter, leading to a world-wide effort to observe directly dark matter particle interactions with ordinary matter in extremely sensitive detectors, which would confirm its existence and shed light on its properties. However, these interactions are so feeble that they have escaped direct detection up to this point, forcing scientists to build detectors that are more and more sensitive.

Existence of Dark Matter

Scientists calculate the mass of large objects in space by studying their motion. Astronomers examining spiral galaxies in the 1970s expected to see material in the center moving faster than on the outer edges. Instead, they found the stars in both locations traveled at the same velocity, indicating the galaxies contained more mass than could be seen. Studies of the gas within elliptical galaxies also indicated a need for more mass than found in visible objects. Clusters of galaxies would fly apart if the only mass they contained were visible to conventional astronomical measurements.

Albert Einstein showed that massive objects in the universe bend and distort light, allowing them to be used as lenses. By studying how light is distorted by galaxy clusters, astronomers have been able to create a map of dark matter in the universe.

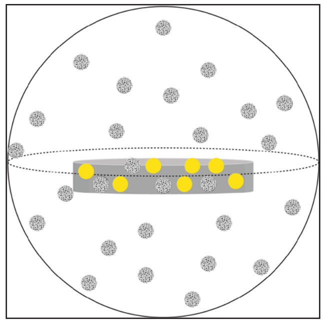
All of these methods provide a strong indication that most of the matter in the universe is something yet unseen.

A Universal Constant for Dark Matter-baryon Interplay

The effect of dark matter-baryon interplay inside an 'interacting region' would depend on the average dark matter number density $\overline{n}_D$, average baryon number density $\overline{n}_B$ and the volume of the interacting region V which depends on the optical radius r_o. Since most of the galaxies are disk-dominated at r_o, the enclosed interacting volume would be a disk-like thin cylinder with radius r_o. Consider the following product (unit: cm^{-2}):

$$K = \frac{\overline{\rho}_D \overline{\rho}_B}{m_D m_B} (\frac{v}{c})^{-4} r_o V,$$

where $\overline{\rho}_D = m_D \overline{n}_D, \overline{\rho}_B = m_B \overline{n}_B$, v is the characteristic velocity, m_D is the dark matter mass and m_B is the average mass of a baryonic particle. Here, we have assumed that all dark matter can interact with baryons. The potential interaction can be gravitational only, or including some other forms of interactions such as scattering.



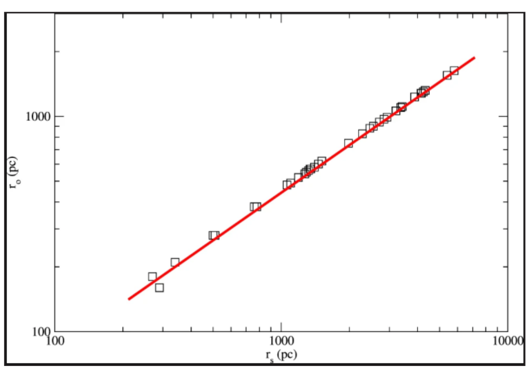
The large sphere is the dark matter core halo with radius $r = r_s$. The grey thin cylinder with base radius r_o indicates the interacting volume V (the effective volume of baryonic matter in a galaxy). The small yellow and black shaded circles represent the baryonic particles and dark matter particles respectively.

The Constant for Galaxies

For spiral galaxies, the Tully-Fisher relation relates the total baryonic mass with velocity as $M_B = (47 \pm 6) M_\odot km^{-4}\,s^{-4}\,v'^4$, where v' is the asymptotic circular velocity. Since most of the galactic rotation curves rise to a constant v' quickly within r_o, we assume that the value of v' is equal to the characteristic velocity v of a galaxy.

Besides, the optical radius r_o is closely related to the scale radius of dark matter r_s. By using the data of the local volume catalog (LVC) dwarf galaxies, a very strong correlation between r_s and r_o is found. Although the relation is not a perfect linear relation, their ratio is close to a constant $r_s = (2.59 \pm 0.55) r_o$. Since $M_B = \bar{\rho}_B V = (47 \pm 6) M_\odot km^{-4} s^{-4} v^4$, the term v in equation above is canceled. Furthermore, putting $r_s = (2.59 \pm 0.55) r_o$ into previous equation and taking $m_B = 1.2\,m_p$, where m_p is the proton mass, we can calculate the value and the uncertainty of K:

$$K = 1.7^{+0.8}_{-0.4} \times 10^{102} \left(\frac{\bar{\rho}_D r_s}{100 M_\odot\,\mathrm{pc}^{-2}} \right) \left(\frac{m_D}{\mathrm{GeV}} \right)^{-1} \mathrm{cm}^{-2}.$$



The graph of r_o versus r_s. The data (indicated by squares) are taken from. The red line indicates the tight correlation between the two quantities.

On the other hand, the product K can also be understood as follow. Since $v = \sqrt{GM_D(r)/r}$, where $M_D(r)$ is the enclosed dynamical mass of a galaxy, we have $v^4 \propto M_D^2 r_o^{-2}$. Assuming $M_D(r_o) \propto \bar{\rho}_D r_o^3$, we get:

$$K \propto \frac{M_B}{M_D(r_o)}.$$

Therefore, K is directly proportional to the ratio of the baryonic mass and the enclosed dynamical mass at radius r_o. Note that this ratio is not necessarily equal to the cosmic baryon fraction $f = 0.156 \pm 0.003$. For different sizes and types of galaxies, it is possible to have entirely different ratios at radii r_o.

Furthermore, the above relation is obtained by using the Tully-Fisher relation, which is formulated based on the data of spiral galaxies. For elliptical galaxies, many of them satisfy another relation called the Faber-Jackson relation. This relation indicates a close relationship between the luminosity L and the stellar velocity dispersion v_b of elliptical galaxies: $L \propto v_b^4$, which is similar to the form of the Tully-Fisher relation. Recent studies show that the data of some elliptical galaxies also fall on the spiral baryonic Tully-Fisher relation if one assumes a certain value of mass-to-luminosity ratio. Therefore, the above result may also be applicable for elliptical galaxies.

The Constant for Galaxy Clusters

The above deduction method can only be applied for galaxies. It is because there is no Tully-Fisher relation for galaxy clusters. Also, the product $\bar{\rho}_D r_s$ is not a constant for galaxy clusters. Fortunately, we can relate the velocity of baryons v' by the temperature T of hot gas in galaxy clusters, which can be given by Virial relation $v \approx \sqrt{3kT/m_p}$. The baryonic component in a large galaxy cluster is dominated by the spherical hot gas halo within the core radius of the hot gas r_c. Therefore, we consider the spherical interacting region inside $r = r_0 = r_c$ so that $V = 4\pi r_c^3/3$. Besides, the hot gas is in hydrostatic equilibrium so that the total mass can be determined:

$$\frac{dP}{dr} = -\frac{GM(r)\rho_B}{r^2},$$

where $P = \rho_B kT/(\mu m_p)$ is the pressure of hot gas, $\mu = 0.61$ is the molecular weight and $M(r)$ is the total enclosed mass. Recent studies show that the x-ray hydrostatic mass measurements are remarkably robust and method-independent. The overall percentage error of mass estimation is about 10–30% if we assume that T is a constant. The hot gas surface brightness profile can be determined by x-ray observations. It is usually described by a β-model:

$$S(r) = S_0[1 + (\frac{r}{r_c})^2]^{-3\beta+1/2},$$

where S_o is the central surface brightness, r_c is the core radius and β is a fitted parameter. These parameters can be used to construct the density profile of hot gas:

$$\rho_B = \rho_{B0}(1+\frac{r^2}{r_c^2})^{-3\beta/2},$$

where ρ_{Bo} is the central baryon density. Combining equations we get:

$$M(r) = \frac{3kT\beta r^3}{\mu m_p G(r_c^2 + r^2)}.$$

The total mass density can be obtained by:

$$\rho_t = \frac{1}{4\pi r^2}\frac{dM(r)}{dr} = \frac{3kT\beta}{4\pi G\mu m_p}[\frac{3r_c^2 + r^2}{(r_c^2 + r^2)^2}].$$

Taking $r \to 0$, we can get the central total mass density ρ_{to}. Since dark matter dominates the mass in a galaxy cluster, the central dark matter density is close to the central mass density:

$$\rho_{D0} \approx \rho_{t0} = \frac{9\beta kT}{4\pi\mu m_p G r_c^2}.$$

Note that for most of the galaxy clusters, the density profiles are close to the NFW profile (a cuspy profile). Here, the density ρ_{Do} represents the average dark matter density for $r \leq r_c$ $(\mathrm{i.e.}\bar{\rho}_D \approx \rho_{D0})$. Using above equation the value of ρ_{Do} for each galaxy cluster can be calculated by the parameters T, β and r_c, including their uncertainties.

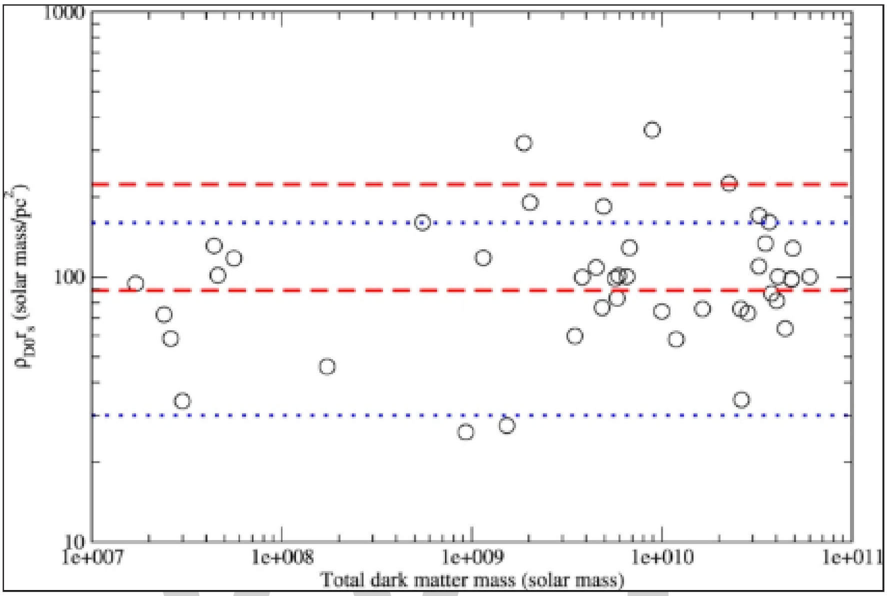
Putting all the above relations to $K = \frac{\bar{\rho}_D\bar{\rho}_B}{m_D m_B}(\frac{v}{c})^{-4} r_o V$, the term v and r_c would be cancelled naturally. Therefore, we get:

$$K = 1.4\times10^{103}(\frac{\bar{\rho}_B}{\bar{\rho}_D})(\frac{m_D}{GeV})^{-1}(\frac{\beta}{0.65})\mathrm{cm}^{-2}.$$

For Galaxies

In many dark matter-dominated galaxies, the mass density of dark matter for $r \leq r_s$ is close to constant (cored structure: $\bar{\rho}_D \approx \rho_{D0}$). Also, recent analyses suggest that the central dark matter column density (the product of the central dark matter mass density and the scale radius) is almost a constant. Early analysis using ~1000 galaxies gives $\rho_{D0} r_s = 141^{+82}_{-52} M_\odot \mathrm{pc}^{-2}$ while later analysis using Milky Way spheroidal dwarf galaxies gives $\rho_{D0} r_s = 75^{+85}_{-45} M_\odot \mathrm{pc}^{-2}$. After that, using a sample of the LVC dwarf galaxies, the resultant central surface mass density gives $\rho_{D0} r_s \sim 100 M_\odot \mathrm{pc}^{-2}$. By combining the data of the Milky Way spheroidal dwarf galaxies and the sample of disc galaxies in, we get $\bar{\rho}_D \propto r_s^{-0.97\pm0.14}$ and $\bar{\rho}_D r_s = 95^{+70}_{-40} M_\odot \mathrm{pc}^{-2}$. Therefore, using this result, we get:

$$K = 1.7^{+2.4}_{-1.1} \times 10^{102} \left(\frac{m_D}{\mathrm{GeV}}\right)^{-1} \mathrm{cm}^{-2}.$$

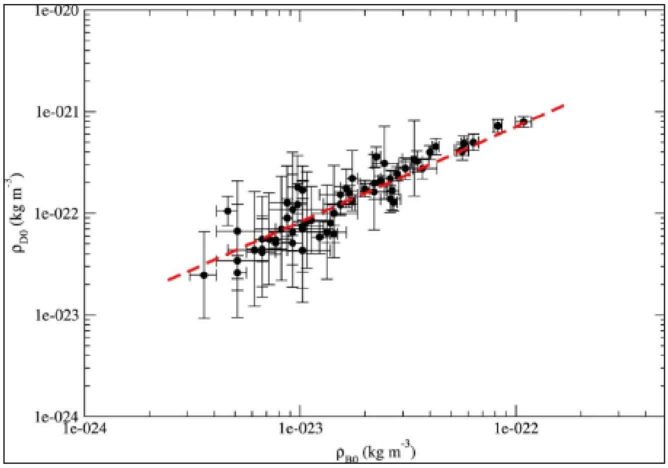


The central dark matter column density $(\mathrm{in}\, M_\odot pc^{-2})$ as a function of total galactic mass of dark matter. The regions bounded by the red dashed lines and bounded by the blue dotted lines are the ranges of the central dark matter column density obtained in and respectively.

The narrow range of K (or the ratio of baryonic mass to the enclosed dynamical mass) for galaxies has been known for a decade. The almost constant central dark matter column density for a wide range of galaxies and the constant luminous-to-dark matter ratio within one halo scale-length suggest the narrow range of K. Therefore, the almost constant value of K for galaxies is expected.

For Galaxy Clusters

By using the x-ray data of 64 large galaxy clusters (with $r_c \geq 100\,\mathrm{kpc}$), we get $\rho_{D0} \propto \rho_{B0}^{1.00\pm0.06}$, where $\rho_{B0} \approx \bar{\rho}_B$ is the central mass density of hot gas. It means that the ratio $\rho_{D0}/\rho_{B0} \approx 8.70 \pm 3.42$ is roughly a constant. This correlation has not been discovered. Although the average dark matter density to average baryonic density ratio for a galaxy cluster should be close to $f^{-1} \approx 6.41$, the ratio at the central region for all galaxy clusters may not be a constant and equal to f^{-1} because the density profiles for dark matter and hot gas are entirely different. This approximately constant ratio of ρ_{Do} to ρ_{Bo} at r_c may reveal some global interplay between dark matter and baryons.

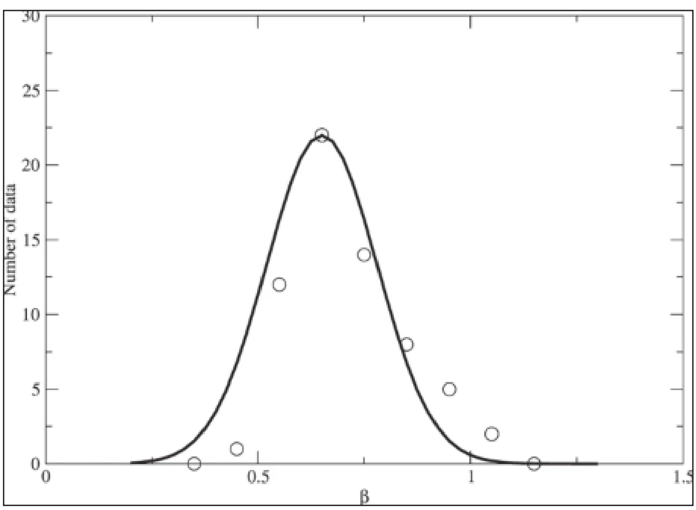


The graph of ρ_{Do} versus ρ_{Bo} for 64 galaxy clusters. The red dashed line indicates the slope of the fit: 1.00 ± 0.06.

Although different galaxy clusters have different values of β, we approximate the distribution of β by a Gaussian function with an average value $\beta = 0.65$ and a dispersion of 0.13. Putting the relation of ρ_{Do} and ρ_{Bo} to $K = 1.4\times10^{103}(\frac{\bar{\rho}_B}{\bar{\rho}_D})(\frac{m_D}{GeV})^{-1}(\frac{\beta}{0.65})\text{cm}^{-2}$ and using $\beta = 0.65 \pm 0.13$, we get a scale invariant constant for galaxy clusters:

$$K \approx 1.6^{+2.3}_{-0.9}\times10^{102}(\frac{m_D}{\text{GeV}})^{-1}\text{cm}^{-2},$$

where $\bar{\rho}_D / \bar{\rho}_B \approx \rho_{D0} / \rho_{B0}$ We can see that the values of K for galaxies and galaxy clusters give excellent agreement with each other, within a factor of 2–3. This constant can be regarded as a universal constant for dark matter-baryon interplay because it is independent of scale.

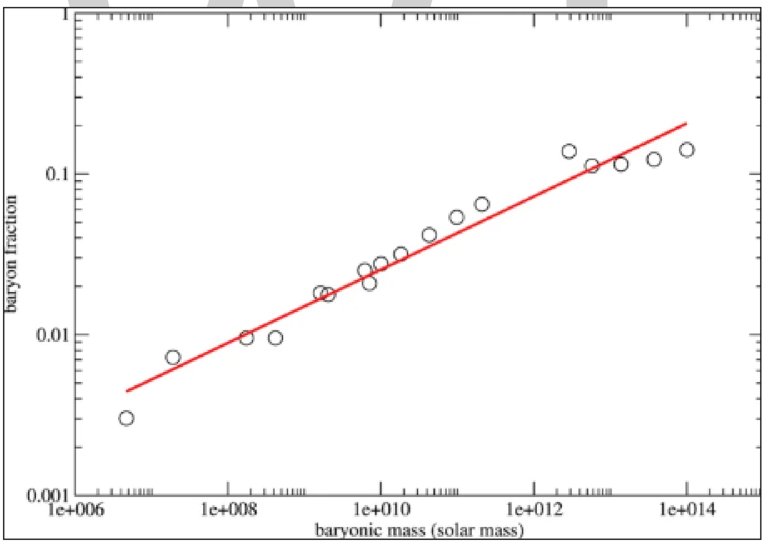


The distribution of β for the 64 galaxy clusters. The black line indicates the Gaussian fit with $\beta = 0.65 \pm 0.13$.

Here, note that the x-ray data used have assumed the hubble parameter $h=0.5$. To match the current observed hubble parameter h = 0.68, we have re-scaled the parameters r_c and ρ_{Bo}. Besides, using the data of large galaxy clusters can minimize the errors because the uncertainties of the central hot gas temperature significantly affect the mass determination of small galaxy clusters. Therefore, we only choose the large galaxy clusters (with $r_c \geq 100$ kpc) to perform the analysis.

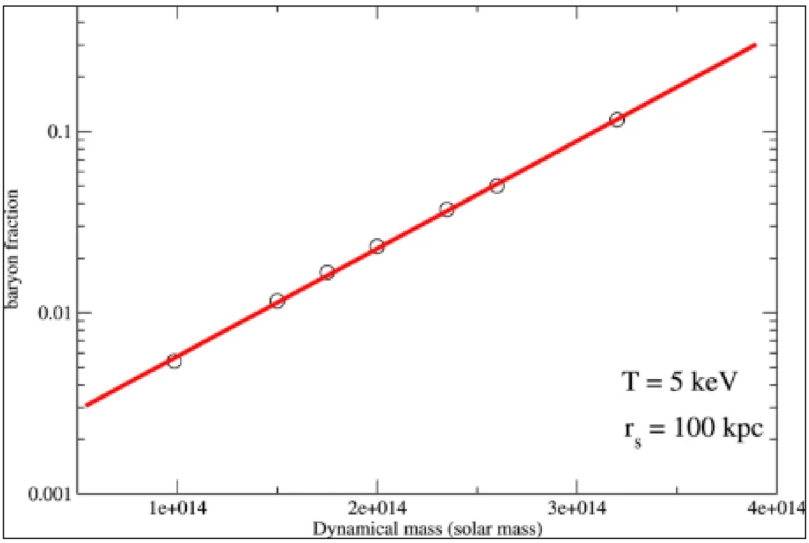
The Relation between the Constant and the Cosmic Baryon Fraction

The value of K is directly proportional to the ratio $M_B/M_D(r_o)$ (for galaxies) or ρ_{Bo}/ρ_{Do} (for galaxy clusters). One may claim that it is not surprising to have a similar constant value of K for galaxies and galaxy clusters because $M_B/M_D(r_o)$ or ρ_{Bo}/ρ_{Do} should be close to the cosmic baryon fraction $f=0.156\pm0.003$, which should be almost a constant for various structures. However, previous studies show that the baryon fractions in various structures are entirely different and the ratios quite depend on the positions of the structures. In figure below, we follow the study in and show the corresponding baryon fractions for different structures (from dwarf galaxies to galaxy clusters). We can see that the ratio can be ranging from 10^{-3} to 10^{-1}, which spans 3 orders of magnitude.



The graph of baryon fraction versus total baryonic mass for different structures. The red line indicates the correlation $f \propto M_B^{0.23}$.

If we assume the NFW profile for dark matter distribution in galaxy clusters and fix the baryon fraction $f=M_B/(M_B+M_D)=0.156$ at a large radius $r=1$ Mpc, we can predict the baryon fraction at the characteristic radius r_s for various total dynamical mass using hydrostatic equilibrium. In figure we can see that the baryon fraction at $r_s=100$ kpc ranges from 10^{-3} to 10^{-1} (not close to $f=0.156\pm0.003$) for just a short range of dynamical mass if we consider a fix hot gas temperature ($T=5$ keV) and scale radius of dark matter $r_s=100$ kpc. For various temperature of hot gas and scale radii of different galaxy clusters, the variation of the baryon fraction can be much larger. This large variation can also be seen in galaxies by examining the radial acceleration relation. Therefore, theoretically, the baryon-dark matter density ratio is not necessarily a constant for various structures, but possibly spans at least 2 orders of magnitude. Furthermore, the characteristic radii of the galaxies (~1–10 kpc) and galaxy clusters considered (~100–500) kpc are different. There is no reason why the baryon fractions for galaxies and galaxy clusters are close to each other at different characteristic radii. Nevertheless, our results show that this ratio is almost constant at the characteristic radii r_o or r_c. Based on the result of figure it seems that gravitational interaction between dark matter and baryonic matter alone is not enough to provide a satisfactory explanation for the constant value of *K*. This may reveal the existence of some interplay or self-organizing processes between baryons and dark matter particles within the 'interacting region'.



The graph of baryon fraction within r_s versus total dynamical mass for galaxy clusters, assuming NFW density profile for dark matter and in hydrostatic equilibrium with hot gas. Here, we fix the hot gas temperature and scale radius of dark matter as 5 keV and $r_s = 100$ kpc respectively.

Cold Dark Matter

In cosmology and physics, cold dark matter (CDM) is a hypothetical type of dark matter. Observations indicate that approximately 85% of the matter in the universe is dark matter, with only a small fraction being the ordinary baryonic matter that composes stars, planets, and living organisms. *Cold* refers to the fact that the dark matter moves slowly compared to the speed of light, while *dark* indicates that it interacts very weakly with ordinary matter and electromagnetic radiation.

The physical nature of CDM is currently unknown, and there are a wide variety of possibilities. Among them are a new type of weakly interacting massive particle, primordial black holes, and axions.

Structure Formation

In the cold dark matter theory, structure grows hierarchically, with small objects collapsing under their self-gravity first and merging in a continuous hierarchy to form larger and more massive objects. Predictions of the cold dark matter paradigm are in general agreement with observations of cosmological large-scale structure.

In the hot dark matter paradigm, popular in the early 1980s and less so now, structure does not form hierarchically (*bottom-up*), but forms by fragmentation (*top-down*), with the largest superclusters forming first in flat pancake-like sheets and subsequently fragmenting into smaller pieces like our galaxy the Milky Way.

Since the late 1980s or 1990s, most cosmologists favor the cold dark matter theory (specifically the modern Lambda-CDM model) as a description of how the universe went from a smooth initial state at early times (as shown by the cosmic microwave background radiation) to the lumpy distribution of galaxies and their clusters we see today—the large-scale structure of the universe. Dwarf galaxies are crucial to this theory, having been created by small-scale density fluctuations in the early universe; they have now become natural building blocks that form larger structures.

Composition

Dark matter is detected through its gravitational interactions with ordinary matter and radiation. As such, it is very difficult to determine what the constituents of cold dark matter are. The candidates fall roughly into three categories:

- Axions, very light particles with a specific type of self-interaction that makes them a suitable CDM candidate. Axions have the theoretical advantage that their existence solves the strong CP problem in quantum chromodynamics, but axion particles have only been theorized and never detected.
- Massive compact halo objects (MACHOs), large, condensed objects such as black holes, neutron stars, white dwarfs, very faint stars, or non-luminous objects like planets. The search for these objects consists of using gravitational lensing to detect the effects of these objects on background galaxies. Most experts believe that the constraints from those searches rule out MACHOs as a viable dark matter candidate.
- Weakly interacting massive particles (WIMPs). There is no currently known particle with the required properties, but many extensions of the standard model of particle physics predict such particles. The search for WIMPs involves attempts at direct detection by highly sensitive detectors, as well as attempts at production of WIMPs by particle accelerators. WIMPs are generally regarded as the most promising candidates for the composition of dark matter. The DAMA/NaI experiment and its successor DAMA/LIBRA have claimed to have directly

detected dark matter particles passing through the Earth, but many scientists remain skeptical because no results from similar experiments seem compatible with the DAMA results.

Challenges

Several discrepancies between the predictions of the particle cold dark matter paradigm and observations of galaxies and their clustering have arisen:

- The cuspy halo problem: The density distributions of dark matter halos in cold dark matter simulations (at least those that do not include the impact of baryonic feedback) are much more peaked than what is observed in galaxies by investigating their rotation curves.
- The missing satellites problem: Cold dark matter simulations predict large numbers of small dark matter halos, more numerous than the number of small dwarf galaxies that are observed around galaxies like the Milky Way.
- The disk of satellites problem: Dwarf galaxies around the Milky Way and Andromeda galaxies are observed to be orbiting in thin, planar structures whereas the simulations predict that they should be distributed randomly about their parent galaxies.
- Galaxy morphology problem: If galaxies grew hierarchically, then massive galaxies required many mergers. Major mergers inevitably create a classical bulge. On the contrary, about 80% of observed galaxies give evidence of no such bulges, and giant pure-disc galaxies are commonplace. That bulgeless fraction was nearly constant for 8 billion years.

Some of these problems have proposed solutions, but it remains unclear whether they can be solved without abandoning the CDM paradigm.

Warm Dark Matter

Warm dark matter (WDM) is a hypothesized form of dark matter that has properties intermediate between those of hot dark matter and cold dark matter, causing structure formation to occur bottom-up from above their free-streaming scale, and top-down below their free streaming scale. The most common WDM candidates are sterile neutrinos and gravitinos. The WIMPs (weakly interacting massive particles), when produced non-thermally could be candidates for warm dark matter. In general, however the thermally produced WIMPs are cold dark matter candidates.

keVins and GeVins

One possible WDM candidate particle with a mass of a few keV comes from introducing two new, zero charge, zero lepton number fermions to the Standard Model of Particle

Physics: “keV-mass inert fermions” (keVins) and “GeV-mass inert fermions” (GeVins). keVins are overproduced if they reach thermal equilibrium in the early universe, but in some scenarios the entropy production from the decays of unstable heavier particles can suppresses their abundance to the correct value. These particles are considered “inert” because they only have suppressed interactions with the Z boson. Sterile neutrinos with masses of a few keV are possible candidates for keVins. At temperatures below the electroweak scale their only interactions with standard model particles are weak interactions due to their mixing with ordinary neutrinos. Due to the smallness of the mixing angle they are not overproduced because they freeze out before reaching thermal equilibrium. Their properties are consistent with astrophysical bounds coming from structure formation and the Pauli principle if their mass is larger than 1-8 keV.

In February 2014, different analyses have extracted from the spectrum of X-ray emissions observed by XMM-Newton, a monochromatic signal around 3.5 keV. This signal is coming from different galaxy clusters (like Perseus and Centaurus) and several scenarios of warm dark matter can justify such a line. We can cite, for example, a 3.5 keV candidate annihilating into 2 photons, or a 7 keV dark matter particle decaying into a photon and a neutrino.

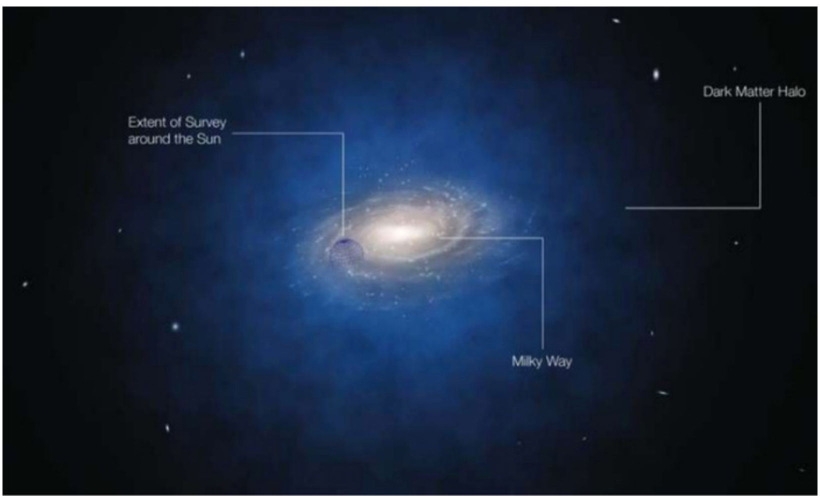
In November 2019, analysis of the interaction of various galactic halo matter on densities and distribution of stellar streams, coming off the satellites of the Milky Way, they were able to constrain minimums of mass for density perturbations by warm dark matter Kevins in the GD-1 and Pal 5 streams. This lower limit on the mass of warm dark matter thermal relics mWDM > 4.6 keV; or adding dwarf satellite counts mWDM > 6.3 keV.

Hot Dark Matter

Dark matter is a form of matter that neither emits nor absorbs light. Within physics, this behavior is characterized by dark matter not interacting with electromagnetic radiation, hence making it dark and rendering it undetectable via conventional instruments in physics. Data from galaxy rotation curves indicate that approximately 80% of the mass of a galaxy cannot be seen, forcing researchers to innovate ways that indirectly detect it through dark matter’s effects on gravitational fluctuations. There exists no consensus in the theoretical physics community as to whether dark matter is divisible into various ‘types’, but there exists evidence for differentiating dark matter into “hot” (HDM) and “cold” (CDM) types– some even suggesting a middle-ground of “warm” dark matter (WDM). The terminology is not meant to invoke any association with temperature, but instead refer to the size of the purported dark matter particles (WIMPs). In turn, the size of the particles determines the velocities at which they travel at in an inverse relationship: HDM travels faster than CDM because the HDM particles are theorized to be of lower mass.

Role in Galaxy Formation

In terms of its application, the distribution of hot dark matter could also help explain how clusters and superclusters of galaxies formed after the Big Bang. Theorists claim that there exist two classes of dark matter: 1) those that "congregate around individual members of a cluster of visible galaxies" and 2) those that encompass "the clusters as a whole." Because cold dark matter possesses a lower velocity, it could be the source of "smaller, galaxy-sized lumps," as shown in the image. Hot dark matter, then, should correspond to the formation of larger mass aggregates that surround whole galaxy clusters. However, data from the cosmic microwave background radiation, as measured by the COBE satellite, is highly uniform, and such high-velocity hot dark matter particles cannot form clumps as small as galaxies beginning from such a smooth initial state, highlighting a discrepancy in what dark matter theory and the actual data are saying. Theoretically, in order to explain relatively small-scale structures in the observable Universe, it is necessary to invoke cold dark matter or WDM. In other words, Hot dark matter being the sole substance in explaining cosmic galaxy formation is no longer viable, placing hot dark matter under the larger umbrella of mixed dark matter (MDM) theory.



Artist's impression of dark matter surrounding the Milky Way.

Neutrinos

An example of a hot dark matter particle is the neutrino. Neutrinos have very small masses, and do not take part in two of the four fundamental forces, the electromagnetic interaction and the strong interaction. They interact by the weak interaction, and gravity, but due to the feeble strength of these forces, they are difficult to detect. A number of projects, such as the Super-Kamiokande neutrino observatory, in Gifu, Japan are currently studying these neutrinos.

Dark Radiation

Dark radiation (also dark electromagnetism) is a postulated type of radiation that mediates interactions of dark matter.

By analogy to the way photons mediate electromagnetic interactions between particles in the Standard Model (called *baryonic matter* in cosmology), dark radiation is proposed to mediate interactions between dark matter particles. Similar to dark matter particles, the hypothetical dark radiation does not interact with Standard Model particles.

There has been no notable evidence for the existence of such radiation, but since baryonic matter contains multiple interacting particle types, it is reasonable to suppose that dark matter does also. Moreover, it has been pointed out recently that the cosmic microwave background data seems to suggest that the number of effective neutrino degrees of freedom is more than 3.046, which is slightly more than the standard case for 3 types of neutrino. This extra degree of freedom could arise from having a non-trivial amount of dark radiation in the universe. One possible candidate for dark radiation is the sterile neutrino.

Self-interacting Dark Matter

In astrophysics and particle physics, self-interacting dark matter (SIDM) assumes dark matter has self-interactions, in contrast to the collisionless dark matter assumed by the Lambda-CDM model. SIDM was postulated in 2000 to resolve a number of conflicts between observations and N-body simulations (of cold collisionless dark matter only) on the galactic scale and smaller. It was also used to explain the 2015 observations of ESO 146-5 the core of the Abell 3827 galaxy cluster. However, the latter finding has since been discounted based on further observations and modelling of the cluster.

If the self-interacting dark matter is in the hydrostatic equilibrium, its pressure and density follow:

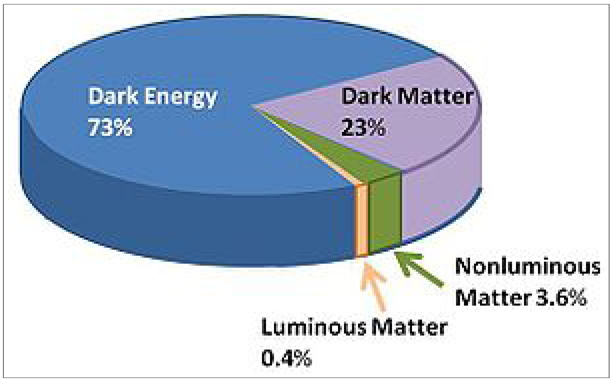
$$\nabla P_\chi / \rho_\chi = \nabla \Phi_{\text{tot}} = \nabla(\Phi_\chi + \Phi_b)$$

where $\Phi\chi$ and Φ_b are gravitational potential of the dark matter and baryon respectively. The equation naturally correlates the dark matter distribution to that of the baryonic matter distribution. With this correlation, the self-interacting dark matter can explain phenomena such as the Tully-Fisher relation.

Self-interacting dark matter has also been postulated as an explanation for the DAMA annual modulation signal.

Scalar Field Dark Matter

In astrophysics and cosmology scalar field dark matter is a classical, minimally coupled, scalar field postulated to account for the inferred dark matter.



Pie chart showing the fractions of energy in the universe contributed by different sources. Ordinary matter is divided into luminous matter (the stars and luminous gases and 0.005% radiation) and nonluminous matter (intergalactic gas and about 0.1% neutrinos and 0.04% supermassive black holes). Ordinary matter is uncommon.

The universe may be accelerating, fueled perhaps by a cosmological constant or some other field possessing long range 'repulsive' effects. A model must predict the correct form for the large scale clustering spectrum, account for cosmic microwave background anisotropies on large and intermediate angular scales, and provide agreement with the luminosity distance relation obtained from observations of high redshift supernovae. The modeled evolution of the universe includes a large amount of unknown matter and energy in order to agree with such observations. This energy density has two components: cold dark matter and dark energy. Each contributes to the theory of the origination of galaxies and the expansion of the universe. The universe must have a critical density, a density not explained by baryonic matter (ordinary matter) alone.

Scalar Field

The dark matter can be modeled as a scalar field using two fitted parameters, mass and self-interaction. In this picture the dark matter consists of an ultralight particle with a mass of $\sim 10^{-22}$ eV when there is no self-interaction. If there is a self-interaction a wider mass range is allowed. The uncertainty in position of a particle is larger than its Compton wavelength, and for some reasonable estimates of particle mass and density of dark matter there is no point talking about the individual particles' positions and momenta. Ultra-light dark matter would be more like a wave than a particle, and the galactic halos are giant systems of condensed bose liquid, possibly superfluid. The dark matter can be described as a Bose–Einstein condensate of the ultralight quanta of the field and as boson stars. The enormous Compton wavelength of these particles prevents structure formation on small, subgalactic scales, which is a major problem in traditional cold dark matter models. The collapse of initial over-densities is studied in the references.

This dark matter model is also known as BEC dark matter or wave dark matter. Fuzzy dark matter and ultra-light axion are examples of scalar field dark matter.

Light Dark Matter

In astronomy and cosmology, light dark matter are dark matter weakly interacting massive particles (WIMPS) candidates with masses less than 1 GeV. These particles are heavier than warm dark matter and hot dark matter, but are lighter than the traditional forms[which?] of cold dark matter. The Lee-Weinberg bound limits the mass of the favored dark matter candidate, WIMPs, that interact via the weak interaction to ≈ 2 GeV. This bound arises as follows. The lower the mass of WIMPs is, the lower the annihilation cross section, which is of the order $\approx m^2 / M^4$, where m is the WIMP mass and M the mass of the Z-boson. This means that low mass WIMPs, which would be abundantly produced in the early universe, freeze out (i.e. stop interacting) much earlier and thus at a higher temperature, than higher mass WIMPs. This leads to a higher relic WIMP density. If the mass is lower than ~ 2 GeV the WIMP relic density would overclose the universe.

Some of the few loopholes allowing one to avoid the Lee-Weinberg bound without introducing new forces below the electroweak scale have been ruled out by accelerator experiments (i.e. CERN, Tevatron), and in decays of B mesons.

A viable way of building light dark matter models is thus by postulating new light bosons. This increases the annihilation cross section and reduces the coupling of dark matter particles to the Standard Model making them consistent with accelerator experiments.

Motivation

In recent years, light dark matter has become popular due in part to the many benefits of the theory. Sub-GeV dark matter has been used to explain the positron excess in the galactic center observed by INTEGRAL, excess gamma rays from the galactic center and extragalactic sources. It has also been suggested that light dark matter may explain a small discrepancy in the measured value of the fine structure constant in different experiments.

Mixed Dark Matter

Mixed dark matter (MDM) is a dark matter (DM) model proposed during the late 1990s. It is also called hot + cold dark matter. The most abundant form of dark matter is cold dark matter, almost one fourth of the energy contents of the Universe. Neutrinos are the only known particles whose Big-Bang thermal relic should compose at least a fraction of Hot dark matter (HDM), albeit other candidates are speculated to exist. In the early 1990s, the power spectrum of fluctuations in the galaxy clustering did not agree entirely with the predictions for a standard cosmology built around pure cold DM. Mixed dark matter with a composition of about 80% cold and 20% hot (neutrinos) was investigated and found to agree better with observations. This large amount of HDM was made obsolete by the discovery in 1998 of the acceleration of universal expansion, which eventually led to the dark energy + dark matter paradigm of this decade.

The cosmological effects of cold DM are almost opposite to the hot DM effects. Given that cold DM promotes the growth of large scale structures, it is often believed to be composed of Weakly interacting massive particles (WIMPs). Conversely hot DM suffers of free-streaming for most of the history of the Universe, washing-out the formation of small scales. In other words, the mass of hot DM particles is too small to produce the observed gravitationally bounded objects in the Universe. For that reason, the hot DM abundance is constrained by Cosmology to less than one percent of the Universe contents.

The Mixed Dark Matter scenario recovered relevance when DM was proposed to be a thermal relic of a Bose–Einstein condensate made of very light bosonic particles, as light as neutrinos or even lighter like the Axion. This cosmological model predicts that cold DM is made of many condensed particles, while a small fraction of these particles resides in excited energetic states contributing to hot DM.

Dark Energy

In physical cosmology, dark energy is a hypothetical form of energy that permeates all of space and tends to increase the rate of expansion of the universe. It is called "dark" because it has no electric charge and does not interact with electromagnetic radiation, such as light.

If this concept is correct, dark energy will join the other main components so far established in the universe: (a) Regular ("baryonic") matter, made of electrons and quarks; (b) dark matter, which outweighs regular matter but is of unknown composition; and (c) relic photons (the cosmic microwave background radiation) and neutrinos from the Big Bang. (While outnumbering both regular and dark matter, these photons are thought to be so attenuated by time that their role is now minimal).

All these components of the universe have positive mass/energy. They have an attractive gravitational interaction and are slowing down the expansion of the universe. The putative dark energy, however, has a negative mass/energy and a repelling gravitational interaction. The effect of dark energy is opposite to the positive energy components. It is speeding up the expansion of the universe.

Postulating the existence of dark energy is the most popular way to explain recent observations that the universe appears to be expanding at an accelerating rate. In the standard model of cosmology, dark energy currently accounts for almost three-quarters of the total mass-energy of the universe.

Dark energy has been used as a crucial ingredient in a recent attempt to formulate a cyclic model for the universe.

Two Proposed Forms for Dark Energy

Two proposed forms for dark energy are the cosmological constant, a constant energy density filling space homogeneously, and scalar fields such as quintessence or moduli, dynamic fields whose energy density can vary in time and space. In fact, contributions from scalar fields that are constant in space are usually also included in the cosmological constant. The cosmological constant is thought to arise from the vacuum energy. Scalar fields which do change in space are hard to distinguish from a cosmological constant, because the change may be extremely slow.

High-precision measurements of the expansion of the universe are required to understand how the speed of the expansion changes over time. The rate of expansion is parameterized by the cosmological equation of state. Measuring the equation of the state of dark energy is one of the biggest efforts in observational cosmology today.

Adding the cosmological constant to cosmology's standard Friedmann-Robertson-Walker metric (FLRW metric) leads to the Lambda-CDM model, which has been referred to as the "standard model" of cosmology because of its precise agreement with observations.

Negative Pressure

Strangely, dark energy causes expansion because it has strong negative pressure. A substance has positive pressure when it pushes outward on its surroundings. This is the usual situation for fluids. Negative pressure, or tension, exists when the substance instead pulls on its surroundings. A common example of negative pressure occurs when a solid is stretched to support a hanging weight.

According to the Friedmann-Lemaître-Robertson-Walker metric, which is an application of General Relativity to cosmology, the pressure within a substance contributes to its gravitational attraction for other things just as its mass density does. Negative pressure causes a gravitational repulsion.

The gravitational repulsive effect of dark energy's negative pressure is greater than the gravitational attraction caused by the energy itself. At the cosmological scale, it also overwhelms all other forms of gravitational attraction, resulting in the accelerating expansion of the universe.

One might wonder, how can pushing cause attraction? How can pulling cause repulsion? This sounds like a contradiction. The solution is:

- The pushing of positive pressure (and the pulling of negative pressure) are non-gravitational forces which just move substances around within space without changing space itself.
- But the *gravitational* attraction (or repulsion) they cause operates on space

itself, decreasing (or increasing) the amount of space between things. It is this which determines the size of the universe.

- There is no necessity that these two effects should act in the same direction. In fact, they act in opposite directions.

Evidence for Dark Energy

In 1998, observations of type Ia supernovae ("one-A") by the Supernova Cosmology Project at the Lawrence Berkeley National Laboratory and the High-z Supernova Search Team suggested that the expansion of the universe is accelerating. Since then, these observations have been corroborated by several independent sources. Measurements of the cosmic microwave background, gravitational lensing, and the large scale structure of the cosmos as well as improved measurements of supernovae have been consistent with the Lambda-CDM model.

The type Ia supernovae provide the most direct evidence for dark energy. Measuring the scale factor at the time that light was emitted from an object is accomplished easily by measuring the redshift of the receding object. Finding the distance to an object is a more difficult problem, however. It is necessary to find standard candles: Objects for which the actual brightness, what astronomers call the absolute magnitude, is known, so that it is possible to relate the observed brightness, or apparent magnitude, to the distance. Without standard candles, it is impossible to measure the redshift-distance relation of Hubble's law. Type Ia supernovae are the best known standard candles for cosmological observation because they are very bright and thus visible across billions of light years. The consistency in absolute magnitude for type Ia supernovae is explained by the favored model of an old white dwarf star which gains mass from a companion star and grows until it reaches the precisely defined Chandrasekhar limit. At this mass, the white dwarf is unstable to thermonuclear runaway and explodes as a type Ia supernova with a characteristic brightness. The observed brightness of the supernovae are plotted against their redshifts, and this is used to measure the expansion history of the universe. These observations indicate that the expansion of the universe is not decelerating, which would be expected for a matter-dominated universe, but rather is mysteriously accelerating. These observations are explained by postulating a kind of energy with negative pressure: Dark energy.

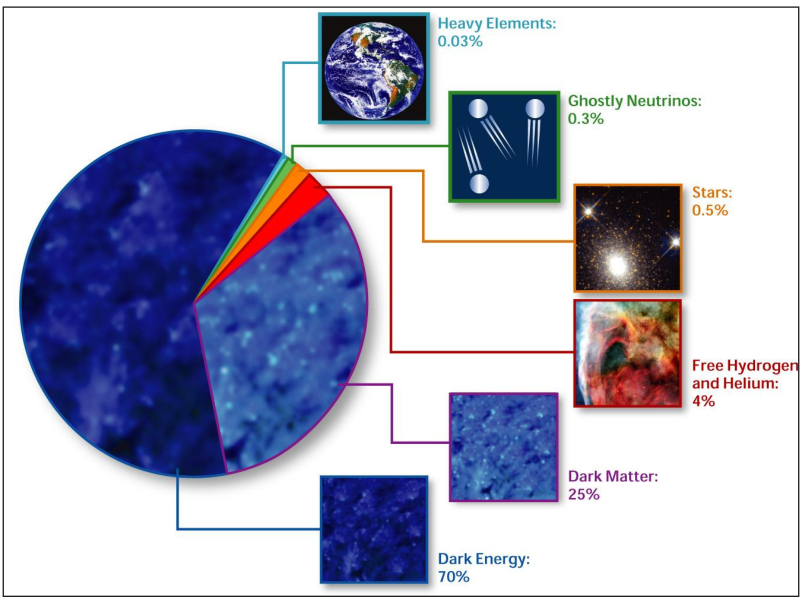
The existence of dark energy, in whatever form, is needed to reconcile the measured geometry of space with the total amount of matter in the universe. Measurements of the cosmic microwave background (CMB), most recently by the WMAP satellite, indicate that the universe is very close to flat. For the shape of the universe to be flat, the mass/energy density of the Universe must be equal to a certain critical density. The total amount of matter in the Universe (including baryons and dark matter), as measured by the CMB, accounts for only about 30 percent of the critical density. This implies the existence of an additional form of energy to account for the remaining 70 percent.

The theory of large scale structure, which governs the formation of structure in the universe (stars, quasars, galaxies and galaxy clusters), also suggests that the density of matter in the universe is only 30 percent of the critical density.

The most recent WMAP observations are consistent with a Universe made up of 74 percent dark energy, 22 percent dark matter, and 4 percent ordinary matter.

Nature of Dark Energy

The exact nature of this dark energy is a matter of speculation. It is known to be very homogeneous, not very dense and is not known to interact through any of the fundamental forces other than gravity. Since it is not very dense—roughly 10^{-29} grams per cubic centimeter—it is hard to imagine experiments to detect it in the laboratory. Dark energy can only have such a profound impact on the universe, making up 70 percent of all energy, because it uniformly fills otherwise empty space. The two leading models are quintessence and the cosmological constant.



As this NASA chart indicates, roughly 70 percent or more of the universe consists of dark energy, about which we know next to nothing.

Cosmological Constant

The simplest explanation for dark energy is that it is simply the "cost of having space:" That is, a volume of space has some intrinsic, fundamental energy. This is the cosmological constant, sometimes called Lambda (hence Lambda-CDM model) after the Greek letter Λ, the symbol used to mathematically represent this quantity. Since energy and mass are related by $E = mc^2$, Einstein's theory of general relativity predicts that it will have a gravitational effect. It is sometimes called a vacuum energy because it is the energy density of empty vacuum. In fact, most theories of particle physics predict vacuum fluctuations that would give the vacuum exactly this sort of energy. The cosmological constant is estimated by cosmologists to be on the order of 10^{-29}g/cm^3, or about 10^{-120} in reduced Planck units.

The cosmological constant has negative pressure equal to its energy density and so causes the expansion of the universe to accelerate. The reason why a cosmological constant has negative pressure can be seen from classical thermodynamics; Energy must be lost from inside a container to do work on the container. A change in volume dV requires work done equal to a change of energy $-p\ dV$, where p is the pressure. But the amount of energy in a box of vacuum energy actually increases when the volume increases (dV is positive), because the energy is equal to ρV, where ρ (rho) is the energy density of the cosmological constant. Therefore, p is negative and, in fact, $p = -\rho$.

A major outstanding problem is that most quantum field theories predict a huge cosmological constant from the energy of the quantum vacuum, up to 120 orders of magnitude too large. This would need to be canceled almost, but not exactly, by an equally large term of the opposite sign. Some supersymmetric theories require a cosmological constant that is exactly zero, which does not help. The present scientific consensus amounts to extrapolating the empirical evidence where it is relevant to predictions, and fine-tuning theories until a more elegant solution is found. Philosophically, the most elegant solution may be to say that if things were different, humans would not be around to observe anything—the anthropic principle. Technically, this amounts to checking theories against macroscopic observations. Unfortunately, as the known error margin in the constant predicts the fate of the universe more than its present state, many such “deeper” answers remain unknown.

Another problem arises with inclusion of the cosmic constant in the standard model, which is appearance of solutions with regions of discontinuities at low matter density. The discontinuity also affects the past sign of the vacuum energy, changing from the current negative pressure to attractive, as one looks back towards the early Universe. This finding should be considered a shortcoming of the standard model, but only when a term for vacuum energy is included.

In spite of its problems, the cosmological constant is in many respects the most economical solution to the problem of cosmic acceleration. One number successfully explains a multitude of observations. Thus, the current standard model of cosmology, the Lambda-CDM model, includes the cosmological constant as an essential feature.

Quintessence

Dark energy may become dark matter when buffeted by baryonic particles, thus leading to particle-like excitations in some type of dynamical field, referred to as quintessence. Quintessence differs from the cosmological constant in that it can vary in space and time. In order for it not to clump and form structure like matter, it must be very light so that it has a large Compton wavelength.

No evidence of quintessence is yet available, but it has not been ruled out either. It generally predicts a slightly slower acceleration of the expansion of the universe than the cosmological constant. Some scientists think that the best evidence for quintessence would come from violations of Einstein's equivalence principle and variation of the fundamental constants in space or time. Scalar fields are predicted by the standard model and string theory, but an analogous problem to the cosmological constant problem (or the problem of constructing models of cosmic inflation) occurs: Renormalization theory predicts that scalar fields should acquire large masses.

The cosmic coincidence problem asks why the cosmic acceleration began when it did. If cosmic acceleration began earlier in the universe, structures such as galaxies would never have had time to form and life, at least as it is known, would never have had a chance to exist. Proponents of the anthropic principle view this as support for their arguments. However, many models of quintessence have a so-called tracker behavior, which solves this problem. In these models, the quintessence field has a density which closely tracks (but is less than) the radiation density until matter-radiation equality, which triggers quintessence to start behaving as dark energy, eventually dominating the universe. This naturally sets the low energy scale of the dark energy.

Some special cases of quintessence are phantom energy, in which the energy density of quintessence actually increases with time, and k-essence (short for kinetic quintessence) which has a non-standard form of kinetic energy. They can have unusual properties: Phantom energy, for example, can cause a Big Rip.

Alternative Ideas

Some theorists think that dark energy and cosmic acceleration are a failure of general relativity on very large scales, larger than superclusters. It is a tremendous extrapolation to think that the law of gravity, which works so well in the solar system, should work without correction on the scale of the universe. Most attempts at modifying general relativity, however, have turned out to be either equivalent to theories of quintessence, or inconsistent with observations.

Alternative ideas for dark energy have come from string theory, brane cosmology, and the holographic principle, but have not yet proved as compelling as quintessence and the cosmological constant.

Yet another, "radically conservative" class of proposals aims to explain the observational data by a more refined use of established theories rather than through the introduction of dark energy, focusing, for example, on the gravitational effects of density inhomogeneities (assumed negligible in the standard Friedmann-Robertson-Walker approximation and confirmed negligible by studies of the anisotropies of the cosmic microwave background and statistics of large-scale structure) or on consequences of electroweak symmetry breaking in the early universe.

Implications for the Fate of the Universe

Cosmologists estimate that the acceleration began roughly 9 billion years ago. Before that, it is thought that the expansion was decelerating, due to the attractive influence of dark matter and baryons. The density of dark matter in an expanding universe decreases more quickly than dark energy, and eventually the dark energy dominates. Specifically, when the volume of the universe doubles, the density of dark matter is halved but the density of dark energy is nearly unchanged (it is exactly constant in the case of a cosmological constant).

If the acceleration continues indefinitely, the ultimate result will be that galaxies outside the local supercluster will move beyond the cosmic horizon: They will no longer be visible, because their line-of-sight velocity becomes greater than the speed of light. This is not a violation of special relativity, and the effect cannot be used to send a signal between them. (Actually there is no way to even define "relative speed" in a curved spacetime. Relative speed and velocity can only be meaningfully defined in flat spacetime or in sufficiently small, infinitesimal regions of curved spacetime). Rather, it prevents any communication between them and the objects pass out of contact.

The night sky would remain the same however, to the naked eye. The Earth, the Milky Way, and the Virgo supercluster would remain virtually undisturbed while the rest of the universe recedes. In this scenario, the local supercluster would ultimately suffer heat death, just as was thought for the flat, matter-dominated universe, before measurements of cosmic acceleration.

There are some very speculative ideas about the future of the universe. One suggests that phantom energy causes *divergent* expansion, which would imply that the effective force of dark energy continues growing until it dominates all other forces in the universe. Under this scenario, dark energy would ultimately tear apart all gravitationally bound structures, including galaxies and solar systems, and eventually overcome the electrical and nuclear forces to tear apart atoms themselves, ending the universe in a "Big Rip." On the other hand, dark energy might dissipate with time, or even become attractive. Such uncertainties leave open the possibility that gravity might yet rule the day and lead to a universe that contracts in on itself in a "Big Crunch." Some scenarios, such as the cyclic model suggest this could be the case. While these ideas are not supported by observations, they are not ruled out. Measurements of acceleration are crucial to determining the ultimate fate of the universe in big bang theory.

References

- Dark-matter: sciencealert.com, Retrieved 25 July, 2019
- King, Stephen F; Merle, Alexander (2012-08-16). "Warm Dark Matter from keVins". Journal of Cosmology and Astroparticle Physics. IOP Publishing. 2012 (08): 016–016. arXiv:1205.0551. doi:10.1088/1475-7516/2012/08/016. ISSN 1475-7516

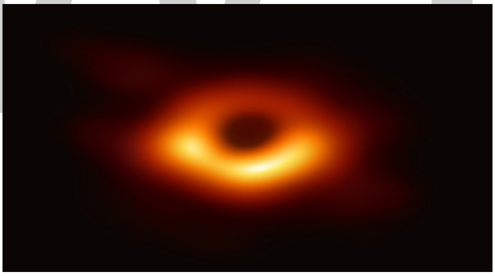
- Dark-matter- 20930: space.com, Retrieved 08 August, 2019
- Hannestad, Steen; Mirizzi, Alessandro; Raffelt, Georg G.; Wong, Yvonne Y. Y. (2010-08-02). "Neutrino and axion hot dark matter bounds after WMAP-7". Journal of Cosmology and Astroparticle Physics. 2010 (08): 001–001. arXiv:1004.0695. Bibcode:2010JCAP...08..001H. doi:10.1088/1475-7516/2010/08/001. ISSN 1475-7516
- Dark-energy: newworldencyclopedia.org, Retrieved 10 July, 2019
- Ackerman, Lotty; et al. (2008). "Dark Matter and Dark Radiation". arXiv:0810.5126. Bibcode:2009PhRvD..79b3519A. doi:10.1103/PhysRevD.79.023519
- Mitra, Saibal (15 June 2005). "Has DAMA detected self-interacting dark matter?". Physical Review D. 71 (12): 121302. arXiv:astro-ph/0409121. Bibcode:2005PhRvD..71l1302M. doi:10.1103/PhysRevD.71.121302

Chapter 4

The Black Holes and Wormholes

Black hole is a spacetime region that exhibits a strong gravitational acceleration that no particles and electromagnetic radiations can escape from it. A wormhole is a speculative structure that links disparate points in spacetime. It is based on a special solution of the Einstein field equations. This chapter discusses the types of blackholes and wormholes in detail.

Black Hole is cosmic body of extremely intense gravity from which nothing, not even light, can escape. A black hole can be formed by the death of a massive star. When such a star has exhausted the internal thermonuclear fuels in its core at the end of its life, the core becomes unstable and gravitationally collapses inward upon itself, and the star's outer layers are blown away. The crushing weight of constituent matter falling in from all sides compresses the dying star to a point of zero volume and infinite density called the singularity.



Black hole in M87.

Black hole at the centre of the massive galaxy M87, about 55 million light-years from Earth, as imaged by the Event Horizon Telescope (EHT). The black hole is 6.5 billion times more massive than the Sun. This image was the first direct visual evidence of a supermassive black hole and its shadow. The ring is brighter on one side because the black hole is rotating, and thus material on the side of the black hole turning toward Earth has its emission boosted by the Doppler effect. The shadow of the black hole is about five and a half times larger than the event horizon, the boundary marking the black hole's limits, where the escape velocity is equal to the speed of light.

Details of the structure of a black hole are calculated from Albert Einstein's general theory of relativity. The singularity constitutes the centre of a black hole and is hidden by the object's "surface," the event horizon. Inside the event horizon the escape velocity (i.e., the velocity required for matter to escape from the gravitational field of a cosmic object) exceeds the speed of light, so that not even rays of light can escape into space. The radius of the event horizon is called the Schwarzschild radius, after the German astronomer Karl Schwarzschild, who in 1916 predicted the existence of collapsed stellar bodies that emit no radiation. The size of the Schwarzschild radius is proportional to the mass of the collapsing star. For a black hole with a mass 10 times as great as that of the Sun, the radius would be 30 km (18.6 miles).

Only the most massive stars—those of more than three solar masses—become black holes at the end of their lives. Stars with a smaller amount of mass evolve into less compressed bodies, either white dwarfs or neutron stars.

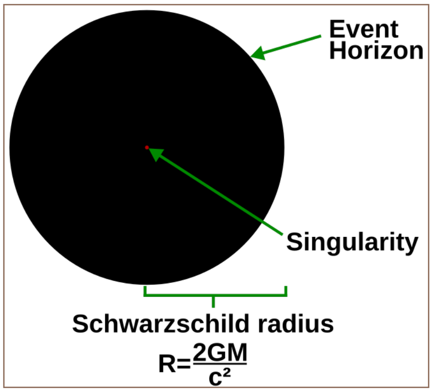
Black holes cannot be observed directly on account of both their small size and the fact that they emit no light. They can be "observed," however, by the effects of their enormous gravitational fields on nearby matter. For example, if a black hole is a member of a binary star system, matter flowing into it from its companion becomes intensely heated and then radiates X-rays copiously before entering the event horizon of the black hole and disappearing forever. One of the component stars of the binary X-ray system Cygnus X-1 is a black hole. Discovered in 1971 in the constellation Cygnus, this binary consists of a blue supergiant and an invisible companion 8.7 times the mass of the Sun that revolve about one another in a period of 5.6 days.

Some black holes apparently have nonstellar origins. Various astronomers have speculated that large volumes of interstellar gas collect and collapse into supermassive black holes at the centres of quasars and galaxies. A mass of gas falling rapidly into a black hole is estimated to give off more than 100 times as much energy as is released by the identical amount of mass through nuclear fusion. Accordingly, the collapse of millions or billions of solar masses of interstellar gas under gravitational force into a large black hole would account for the enormous energy output of quasars and certain galactic systems.

One such supermassive black hole, Sagittarius A*, exists at the centre of the Milky Way Galaxy. In 2005, infrared observations of stars orbiting the position of Sagittarius A* demonstrated the presence of a black hole with a mass equivalent to 4,310,000 Suns. Supermassive black holes have been detected in other galaxies as well. In 2017 the Event Horizon Telescope obtained an image of the supermassive black hole at the centre of the M87 galaxy. That black hole has a mass equal to six and a half billion Suns but is only 38 billion km (24 billion miles) across. It was the first black hole to be imaged directly. The existence of even larger black holes, each with a mass equal to 10 billion Suns, can be inferred from the energetic effects on gas swirling at extremely high velocities around the centre of NGC 3842 and NGC 4889, galaxies near the Milky Way.

The existence of another kind of nonstellar black hole was proposed by the British astrophysicist Stephen Hawking. According to Hawking's theory, numerous tiny primordial black holes, possibly with a mass equal to or less than that of an asteroid, might have been created during the big bang, a state of extremely high temperatures and density in which the universe is thought to have originated 13.8 billion years ago. These so-called mini black holes, like the more massive variety, lose mass over time through Hawking radiation and disappear. If certain theories of the universe that require extra dimensions are correct, the Large Hadron Collider could produce significant numbers of mini black holes.

Properties and Structure

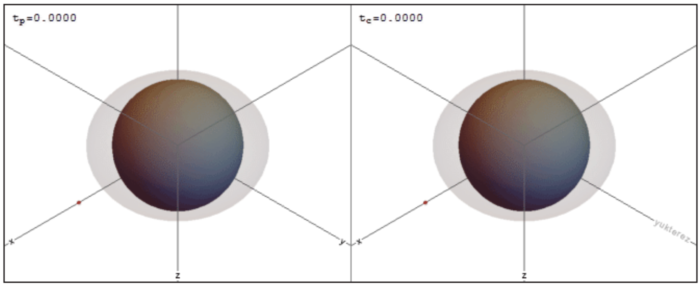


Simple illustration of a non-spinning black hole.

The no-hair conjecture postulates that, once it achieves a stable condition after formation, a black hole has only three independent physical properties: Mass, charge, and angular momentum; the black hole is otherwise featureless. If the conjecture is true, any two black holes that share the same values for these properties, or parameters, are indistinguishable from one another. The degree to which the conjecture is true for real black holes under the laws of modern physics, is currently an unsolved problem.

These properties are special because they are visible from outside a black hole. For example, a charged black hole repels other like charges just like any other charged object. Similarly, the total mass inside a sphere containing a black hole can be found by using the gravitational analog of Gauss's law, the ADM mass, far away from the black hole. Likewise, the angular momentum can be measured from far away using frame dragging by the gravitomagnetic field.

When an object falls into a black hole, any information about the shape of the object or distribution of charge on it is evenly distributed along the horizon of the black hole, and is lost to outside observers. The behavior of the horizon in this situation is a dissipative system that is closely analogous to that of a conductive stretchy membrane with friction and electrical resistance—the membrane paradigm. This is different from other field theories such as electromagnetism, which do not have any friction or resistivity at the microscopic level, because they are time-reversible. Because a black hole eventually achieves a stable state with only three parameters, there is no way to avoid losing information about the initial conditions: the gravitational and electric fields of a black hole give very little information about what went in. The information that is lost includes every quantity that cannot be measured far away from the black hole horizon, including approximately conserved quantum numbers such as the total baryon number and lepton number. This behavior is so puzzling that it has been called the black hole information loss paradox.



Gravitational time dilation around a black hole.

Physical Properties

The simplest static black holes have mass but neither electric charge nor angular momentum. These black holes are often referred to as Schwarzschild black holes after Karl Schwarzschild who discovered this solution in 1916. According to Birkhoff's theorem, it is the only vacuum solution that is spherically symmetric. This means there is no observable difference at a distance between the gravitational field of such a black hole and that of any other spherical object of the same mass. The popular notion of a black hole "sucking in everything" in its surroundings is therefore correct only near a black hole's horizon; far away, the external gravitational field is identical to that of any other body of the same mass.

Solutions describing more general black holes also exist. Non-rotating charged black holes are described by the Reissner–Nordström metric, while the Kerr metric describes a non-charged rotating black hole. The most general stationary black hole solution known is the Kerr–Newman metric, which describes a black hole with both charge and angular momentum.

While the mass of a black hole can take any positive value, the charge and angular momentum are constrained by the mass. In Planck units, the total electric charge Q and the total angular momentum J are expected to satisfy:

$$Q^2 + \left(\tfrac{J}{M}\right)^2 \leq M^2$$

for a black hole of mass M. Black holes with the minimum possible mass satisfying this inequality are called extremal. Solutions of Einstein's equations that violate this inequality exist, but they do not possess an event horizon. These solutions have so-called naked singularities that can be observed from the outside, and hence are deemed *unphysical*. The cosmic censorship hypothesis rules out the formation of such singularities, when they are created through the gravitational collapse of realistic matter. This is supported by numerical simulations.

Due to the relatively large strength of the electromagnetic force, black holes forming from the collapse of stars are expected to retain the nearly neutral charge of the star.

Rotation, however, is expected to be a universal feature of compact astrophysical objects. The black-hole candidate binary X-ray source GRS 1915+105 appears to have an angular momentum near the maximum allowed value. That uncharged limit is:

$$J \le \frac{GM^2}{c},$$

allowing definition of a dimensionless spin parameter such that:

$$0 \le \frac{cJ}{GM^2} \le 1.$$

Black holes are commonly classified according to their mass, independent of angular momentum, J. The size of a black hole, as determined by the radius of the event horizon, or Schwarzschild radius, is proportional to the mass, M, through:

$$r_s = \frac{2GM}{c^2} \approx 2.95 \frac{M}{M_{Sun}} \text{km},$$

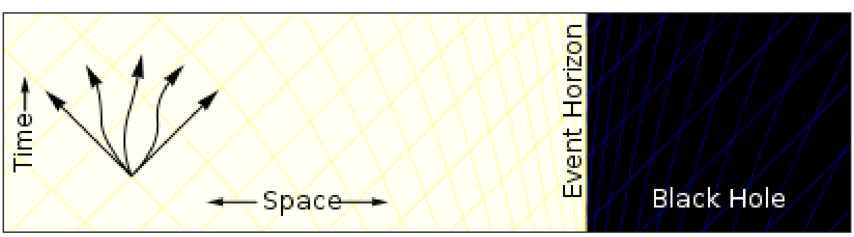
where r_s is the Schwarzschild radius and M_{Sun} is the mass of the Sun. For a black hole with nonzero spin and/or electric charge, the radius is smaller, until an extremal black hole could have an event horizon close to:

$$r_+ = \frac{GM}{c^2}.$$

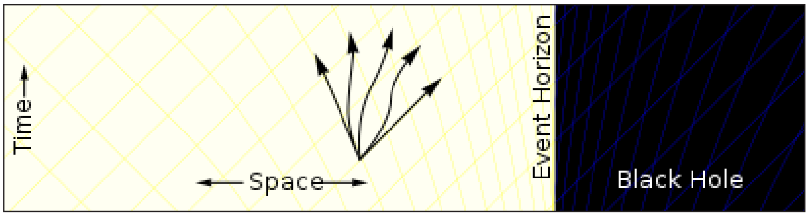
Table: Black hole classifications.

Class	Approx. mass	Approx. radius
Supermassive black hole	10^5–10^{10} M_{Sun}	0.001–400 AU
Intermediate-mass black hole	10^3 M_{Sun}	10^3 km ≈ R_{Earth}
Stellar black hole	10 M_{Sun}	30 km
Micro black hole	up to M_{Moon}	up to 0.1 mm

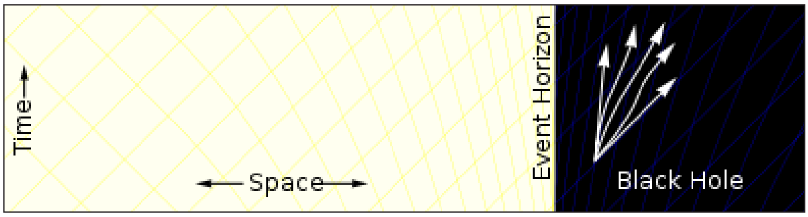
Event Horizon



Far away from the black hole, a particle can move in any direction, as illustrated by the set of arrows. It is restricted only by the speed of light.



Closer to the black hole, spacetime starts to deform. There are more paths going towards the black hole than paths moving away.



Inside of the event horizon, all paths bring the particle closer to the center of the black hole. It is no longer possible for the particle to escape.

The defining feature of a black hole is the appearance of an event horizon—a boundary in spacetime through which matter and light can pass only inward towards the mass of the black hole. Nothing, not even light, can escape from inside the event horizon. The event horizon is referred to as such because if an event occurs within the boundary, information from that event cannot reach an outside observer, making it impossible to determine whether such an event occurred.

As predicted by general relativity, the presence of a mass deforms spacetime in such a way that the paths taken by particles bend towards the mass. At the event horizon of a black hole, this deformation becomes so strong that there are no paths that lead away from the black hole.

To a distant observer, clocks near a black hole would appear to tick more slowly than those further away from the black hole. Due to this effect, known as gravitational time dilation, an object falling into a black hole appears to slow as it approaches the event horizon, taking an infinite time to reach it. At the same time, all processes on this object slow down, from the view point of a fixed outside observer, causing any light emitted by the object to appear redder and dimmer, an effect known as gravitational redshift. Eventually, the falling object fades away until it can no longer be seen. Typically this process happens very rapidly with an object disappearing from view within less than a second.

On the other hand, indestructible observers falling into a black hole do not notice any of these effects as they cross the event horizon. According to their own clocks, which appear to them to tick normally, they cross the event horizon after a finite time without noting any singular behaviour; in classical general relativity, it is impossible to determine the location of the event horizon from local observations, due to Einstein's equivalence principle.

The shape of the event horizon of a black hole is always approximately spherical. For non-rotating (static) black holes the geometry of the event horizon is precisely spherical, while for rotating black holes the event horizon is oblate.

Singularity

At the center of a black hole, as described by general relativity, may lie a gravitational singularity, a region where the spacetime curvature becomes infinite. For a non-rotating black hole, this region takes the shape of a single point and for a rotating black hole, it is smeared out to form a ring singularity that lies in the plane of rotation. In both cases, the singular region has zero volume. It can also be shown that the singular region contains all the mass of the black hole solution. The singular region can thus be thought of as having infinite density.

Observers falling into a Schwarzschild black hole (*i.e.*, non-rotating and not charged) cannot avoid being carried into the singularity, once they cross the event horizon. They can prolong the experience by accelerating away to slow their descent, but only up to a limit. When they reach the singularity, they are crushed to infinite density and their mass is added to the total of the black hole. Before that happens, they will have been torn apart by the growing tidal forces in a process sometimes referred to as spaghettification or the "noodle effect".

In the case of a charged (Reissner–Nordström) or rotating (Kerr) black hole, it is possible to avoid the singularity. Extending these solutions as far as possible reveals the hypothetical possibility of exiting the black hole into a different spacetime with the black hole acting as a wormhole. The possibility of traveling to another universe is, however, only theoretical since any perturbation would destroy this possibility. It also appears to be possible to follow closed timelike curves (returning to one's own past) around the Kerr singularity, which leads to problems with causality like the grandfather paradox. It is expected that none of these peculiar effects would survive in a proper quantum treatment of rotating and charged black holes.

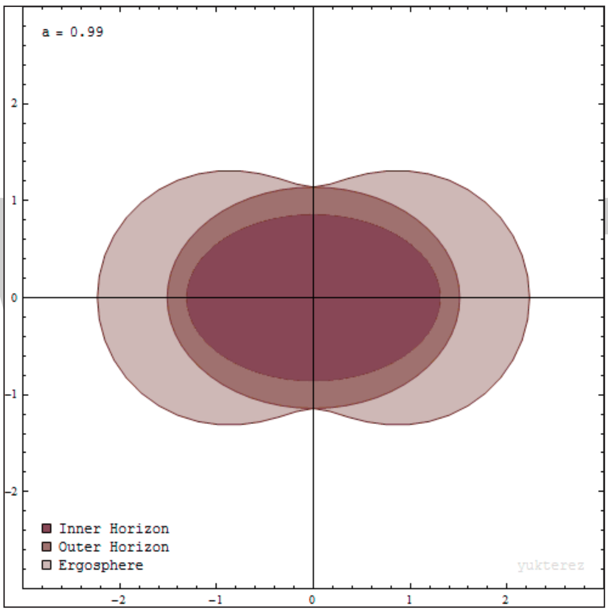
The appearance of singularities in general relativity is commonly perceived as signaling the breakdown of the theory. This breakdown, however, is expected; it occurs in a situation where quantum effects should describe these actions, due to the extremely high density and therefore particle interactions. To date, it has not been possible to combine quantum and gravitational effects into a single theory, although there exist attempts to formulate such a theory of quantum gravity. It is generally expected that such a theory will not feature any singularities.

Photon Sphere

The photon sphere is a spherical boundary of zero thickness in which photons that move on tangents to that sphere would be trapped in a circular orbit about the black hole. For non-rotating black holes, the photon sphere has a radius 1.5 times the Schwarzschild radius. Their orbits would be dynamically unstable, hence any small perturbation, such as a particle of infalling matter, would cause an instability that would grow over time, either setting the photon on an outward trajectory causing it to escape the black hole, or on an inward spiral where it would eventually cross the event horizon.

While light can still escape from the photon sphere, any light that crosses the photon sphere on an inbound trajectory will be captured by the black hole. Hence any light that reaches an outside observer from the photon sphere must have been emitted by objects between the photon sphere and the event horizon.

Ergosphere



The ergosphere is a pumpkin-shaped region outside of the event horizon, where objects cannot remain stationary.

Rotating black holes are surrounded by a region of spacetime in which it is impossible to stand still, called the ergosphere. This is the result of a process known as frame-dragging; general relativity predicts that any rotating mass will tend to slightly “drag” along the spacetime immediately surrounding it. Any object near the rotating mass will tend to start moving in the direction of rotation. For a rotating black hole, this effect is so strong near the event horizon that an object would have to move faster than the speed of light in the opposite direction to just stand still.

The ergosphere of a black hole is a volume whose inner boundary is the black hole’s oblate spheroid event horizon and a pumpkin-shaped outer boundary, which coincides with the event horizon at the poles but noticeably wider around the equator. The outer boundary is sometimes called the *ergosurface*.

Objects and radiation can escape normally from the ergosphere. Through the Penrose process, objects can emerge from the ergosphere with more energy than they entered. This energy is taken from the rotational energy of the black hole causing the latter to slow. A variation of the Penrose process in the presence of strong magnetic fields, the Blandford–Znajek process is considered a likely mechanism for the enormous luminosity and relativistic jets of quasars and other active galactic nuclei.

Innermost Stable Circular Orbit

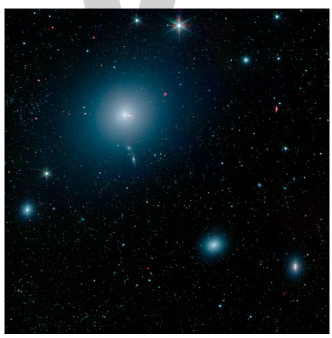
In Newtonian gravity, test particles can stably orbit at arbitrary distances from a central object. In general relativity, however, there exists an innermost stable circular orbit (often called the ISCO), inside of which, any infinitesimal perturbations to a circular orbit will lead to inspiral into the black hole. The location of the ISCO depends on the spin of the black hole, in the case of a Schwarzschild black hole (spin zero) is:

$$r_{isco} = 3r_s = \frac{6GM}{c^2},$$

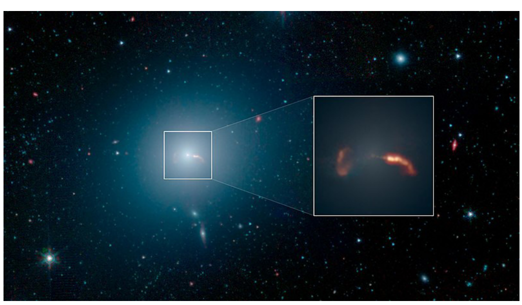
and decreases with increasing black hole spin for particles orbiting in the same direction as the spin.

Observational Evidence

Messier 87 galaxy – home of the first imaged black hole.

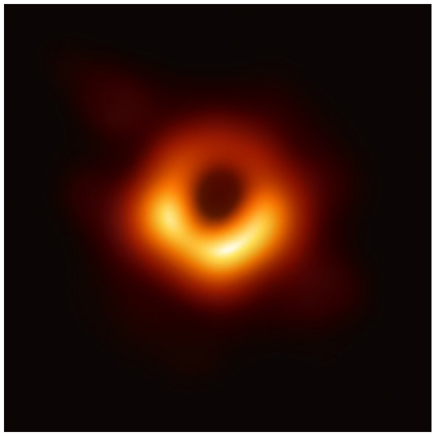


Context.



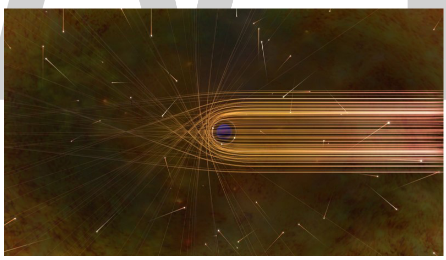
Closeup.

By nature, black holes do not themselves emit any electromagnetic radiation other than the hypothetical Hawking radiation, so astrophysicists searching for black holes must generally rely on indirect observations. For example, a black hole's existence can sometimes be inferred by observing its gravitational influence upon its surroundings.



Supermassive black hole.

On 10 April 2019 an image was released of a black hole, which is seen in magnified fashion because the light paths near the event horizon are highly bent. The dark shadow in the middle results from light paths absorbed by the black hole. The image is in false color, as the detected light halo in this image is not in the visible spectrum, but radio waves.



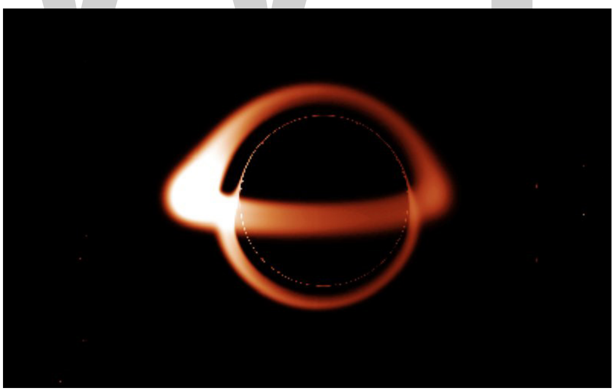
This artist's impression depicts the paths of photons in the vicinity of a black hole. The gravitational bending and capture of light by the event horizon is the cause of the shadow captured by the Event Horizon Telescope.

The Event Horizon Telescope (EHT), run by MIT's Haystack Observatory, is an active program that directly observes the immediate environment of the event horizon of black holes, such as the black hole at the centre of the Milky Way. In April 2017, EHT began observation of the black hole in the center of Messier 87. "In all, eight radio observatories on six mountains and four continents observed the galaxy in Virgo on and off for 10 days in April 2017" to provide the data yielding the image two years later in April 2019. After two years of data processing, EHT released the first direct image of a black hole, specifically the supermassive black hole that lies in the center of the aforementioned galaxy. What is visible is not the black hole, which shows as black because of the loss of all light within this dark region, rather it is the gases at the edge of the event horizon, which are displayed as orange or red, that define the black hole.

The brightening of this material in the 'bottom' half of the processed EHT image is thought to be caused by Doppler beaming, whereby material approaching the viewer at relativistic speeds is perceived as brighter than material moving away. In the case of a black hole this phenomenon implies that the visible material is rotating at relativistic speeds (>1,000 km/s), the only speeds at which it is possible to centrifugally balance the immense gravitational attraction of the singularity, and thereby remain in orbit above the event horizon. This configuration of bright material implies that the EHT observed M87 from a perspective catching the black hole's accretion disc nearly edge-on, as the whole system rotated clockwise. However, the extreme gravitational lensing associated with black holes produces the illusion of a perspective that sees the accretion disc from above. In reality, most of the ring in the EHT image was created when the light emitted by the far side of the accretion disc bent around the black hole's gravity well and escaped such that most of the possible perspectives on M87 can see the entire disc, even that directly behind the "shadow".

Prior to this, in 2015, the EHT detected magnetic fields just outside the event horizon of Sagittarius A*, and even discerned some of their properties. The field lines that pass through the accretion disc were found to be a complex mixture of ordered and tangled. The existence of magnetic fields had been predicted by theoretical studies of black holes.

Detection of Gravitational Waves from Merging Black Holes



Predicted appearance of non-rotating black hole with toroidal ring of ionised matter, such as has been proposed as a model for Sagittarius A*. The asymmetry is due to the Doppler effect resulting from the enormous orbital speed needed for centrifugal balance of the very strong gravitational attraction of the hole.

On 14 September 2015 the LIGO gravitational wave observatory made the first-ever successful direct observation of gravitational waves. The signal was consistent with theoretical predictions for the gravitational waves produced by the merger of two black holes: one with about 36 solar masses, and the other around 29 solar masses.

This observation provides the most concrete evidence for the existence of black holes to date. For instance, the gravitational wave signal suggests that the separation of the two objects prior to the merger was just 350 km (or roughly four times the Schwarzschild radius corresponding to the inferred masses). The objects must therefore have been extremely compact, leaving black holes as the most plausible interpretation.

More importantly, the signal observed by LIGO also included the start of the post-merger ringdown, the signal produced as the newly formed compact object settles down to a stationary state. Arguably, the ringdown is the most direct way of observing a black hole. From the LIGO signal it is possible to extract the frequency and damping time of the dominant mode of the ringdown. From these it is possible to infer the mass and angular momentum of the final object, which match independent predictions from numerical simulations of the merger. The frequency and decay time of the dominant mode are determined by the geometry of the photon sphere. Hence, observation of this mode confirms the presence of a photon sphere, however it cannot exclude possible exotic alternatives to black holes that are compact enough to have a photon sphere.

The observation also provides the first observational evidence for the existence of stellar-mass black hole binaries. Furthermore, it is the first observational evidence of stellar-mass black holes weighing 25 solar masses or more.

On 15 June 2016, a second detection of a gravitational wave event from colliding black holes was announced, and other gravitational wave events have since been observed.

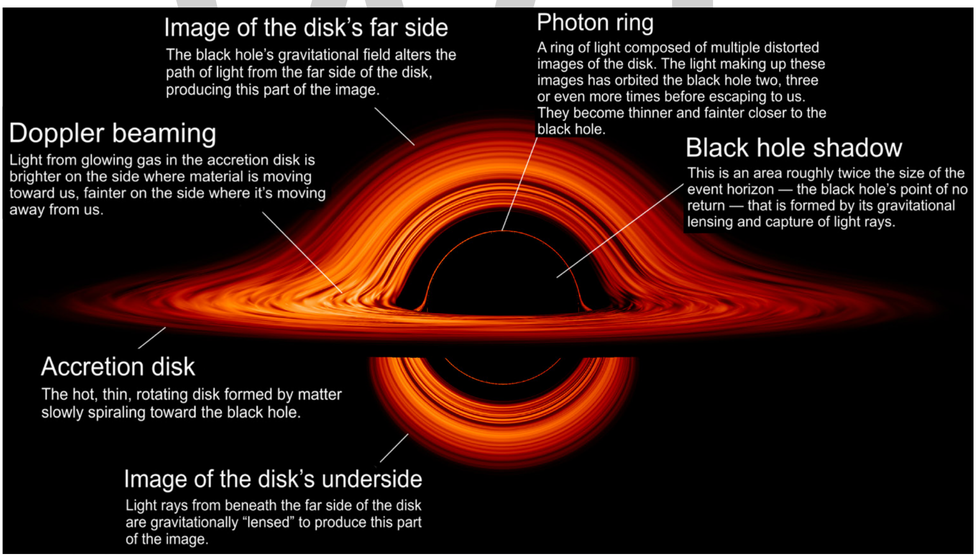
Proper Motions of Stars Orbiting Sagittarius A*

The proper motions of stars near the center of our own Milky Way provide strong observational evidence that these stars are orbiting a supermassive black hole. Since 1995, astronomers have tracked the motions of 90 stars orbiting an invisible object coincident with the radio source Sagittarius A*. By fitting their motions to Keplerian orbits, the astronomers were able to infer, in 1998, that a 2.6 million $M_{\odot}$ object must be contained in a volume with a radius of 0.02 light-years to cause the motions of those stars. Since then, one of the stars—called S2—has completed a full orbit. From the orbital data, astronomers were able to refine the calculations of the mass to 4.3 million $M_{\odot}$ and a radius of less than 0.002 light years for the object causing the orbital motion of those stars. The upper limit on the object's size is still too large to test whether it is smaller than its Schwarzschild radius; nevertheless, these observations strongly suggest that the central object is a supermassive black hole as there are no other plausible scenarios for confining so much invisible mass into such a small volume. Additionally, there is some observational evidence that this object might possess an event horizon, a feature unique to black holes.

Accretion of Matter

Black hole with corona, X-ray source (artist's concept).

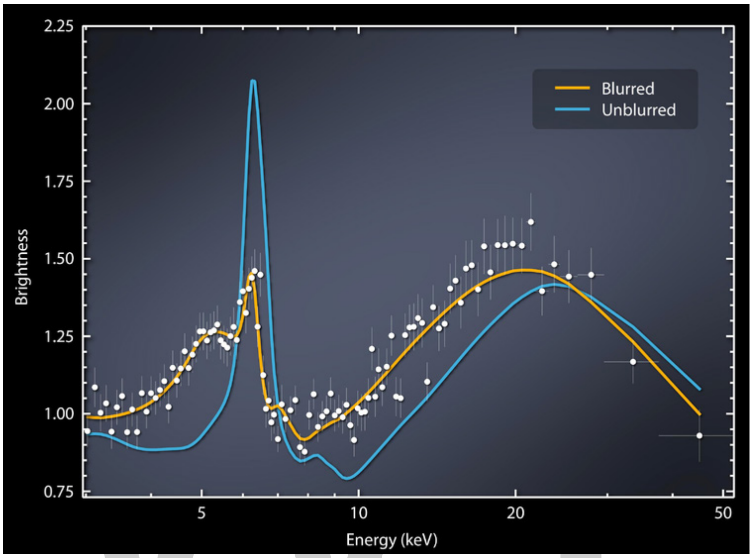
Due to conservation of angular momentum, gas falling into the gravitational well created by a massive object will typically form a disk-like structure around the object. Artists' impressions such as the accompanying representation of a black hole with corona commonly depict the black hole as if it were a flat-space body hiding the part of the disk just behind it, but in reality gravitational lensing would greatly distort the image of the accretion disk.



NASA simulated view from outside the horizon of a Schwarzschild black hole lit by a thin accretion disk.

Within such a disk, friction would cause angular momentum to be transported outward, allowing matter to fall further inward, thus releasing potential energy and increasing the temperature of the gas.

When the accreting object is a neutron star or a black hole, the gas in the inner accretion disk orbits at very high speeds because of its proximity to the compact object. The resulting friction is so significant that it heats the inner disk to temperatures at which it emits vast amounts of electromagnetic radiation (mainly X-rays). These bright X-ray sources may be detected by telescopes. This process of accretion is one of the most efficient energy-producing processes known; up to 40% of the rest mass of the accreted material can be emitted as radiation. (In nuclear fusion only about 0.7% of the rest mass will be emitted as energy.) In many cases, accretion disks are accompanied by relativistic jets that are emitted along the poles, which carry away much of the energy. The mechanism for the creation of these jets is currently not well understood, in part due to insufficient data.



Blurring of X-rays near black hole.

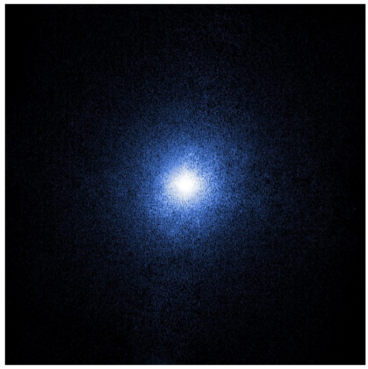
As such, many of the universe's more energetic phenomena have been attributed to the accretion of matter on black holes. In particular, active galactic nuclei and quasars are believed to be the accretion disks of supermassive black holes. Similarly, X-ray binaries are generally accepted to be binary star systems in which one of the two stars is a compact object accreting matter from its companion. It has also been suggested that some ultraluminous X-ray sources may be the accretion disks of intermediate-mass black holes.

In November 2011 the first direct observation of a quasar accretion disk around a supermassive black hole was reported.

X-ray Binaries

X-ray binaries are binary star systems that emit a majority of their radiation in the X-ray part of the spectrum. These X-ray emissions are generally thought to result when one of the stars (compact object) accretes matter from another (regular) star. The presence of an ordinary star in such a system provides an opportunity for studying the central object and to determine if it might be a black hole.

If such a system emits signals that can be directly traced back to the compact object, it cannot be a black hole. The absence of such a signal does, however, not exclude the possibility that the compact object is a neutron star. By studying the companion star it is often possible to obtain the orbital parameters of the system and to obtain an estimate for the mass of the compact object. If this is much larger than the Tolman–Oppenheimer–Volkoff limit (the maximum mass a star can have without collapsing) then the object cannot be a neutron star and is generally expected to be a black hole.



A Chandra X-Ray Observatory image of Cygnus X-1, which was the first strong black hole candidate discovered.

The first strong candidate for a black hole, Cygnus X-1, was discovered in this way by Charles Thomas Bolton, Louise Webster and Paul Murdin in 1972. Some doubt, however, remained due to the uncertainties that result from the companion star being much heavier than the candidate black hole. Currently, better candidates for black holes are found in a class of X-ray binaries called soft X-ray transients. In this class of system, the companion star is of relatively low mass allowing for more accurate estimates of the black hole mass. Moreover, these systems actively emit X-rays for only several months once every 10–50 years. During the period of low X-ray emission (called quiescence), the accretion disk is extremely faint allowing detailed observation of the companion star during this period. One of the best such candidates is V404 Cygni.

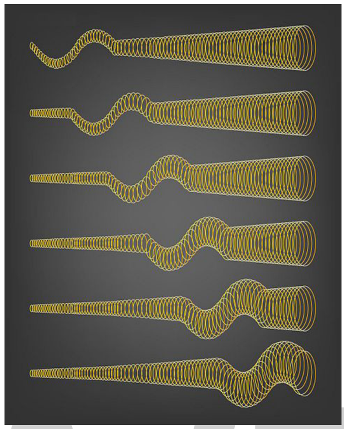
Quiescence and Advection-dominated Accretion Flow

The faintness of the accretion disk of an X-ray binary during quiescence is suspected to be caused by the flow of mass entering a mode called an advection-dominated accretion flow (ADAF). In this mode, almost all the energy generated by friction in the disk is swept along with the flow instead of radiated away. If this model is correct, then it forms strong qualitative evidence for the presence of an event horizon, since if the object at the center of the disk had a solid surface, it would emit large amounts of radiation as the highly energetic gas hits the surface, an effect that is observed for neutron stars in a similar state.

Quasi-periodic Oscillations

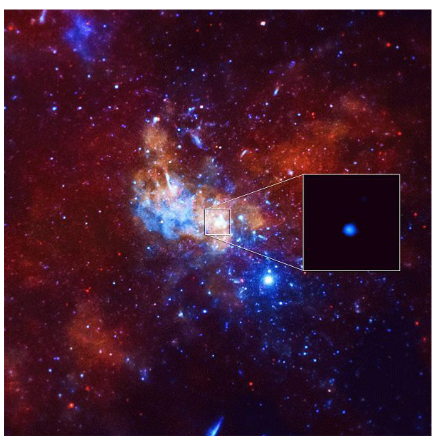
The X-ray emissions from accretion disks sometimes flicker at certain frequencies. These signals are called quasi-periodic oscillations and are thought to be caused by material moving along the inner edge of the accretion disk (the innermost stable circular orbit). As such their frequency is linked to the mass of the compact object. They can thus be used as an alternative way to determine the mass of candidate black holes.

Galactic Nuclei



Magnetic waves, called Alfvén S-waves, flow from the base of black hole jets.

Astronomers use the term "active galaxy" to describe galaxies with unusual characteristics, such as unusual spectral line emission and very strong radio emission. Theoretical and observational studies have shown that the activity in these active galactic nuclei (AGN) may be explained by the presence of supermassive black holes, which can be millions of times more massive than stellar ones. The models of these AGN consist of a central black hole that may be millions or billions of times more massive than the Sun; a disk of gas and dust called an accretion disk; and two jets perpendicular to the accretion disk.



Detection of unusually bright X-Ray flare from Sagittarius A*, a black hole in the center of the Milky Way galaxy on 5 January 2015.

Although supermassive black holes are expected to be found in most AGN, only some galaxies' nuclei have been more carefully studied in attempts to both identify and measure the actual masses of the central supermassive black hole candidates. Some of the most notable galaxies with supermassive black hole candidates include the Andromeda Galaxy, M32, M87, NGC 3115, NGC 3377, NGC 4258, NGC 4889, NGC 1277, OJ 287, APM 08279+5255 and the Sombrero Galaxy.

It is now widely accepted that the center of nearly every galaxy, not just active ones, contains a supermassive black hole. The close observational correlation between the mass of this hole and the velocity dispersion of the host galaxy's bulge, known as the M-sigma relation, strongly suggests a connection between the formation of the black hole and the galaxy itself.

Microlensing

Another way the black hole nature of an object may be tested in the future is through observation of effects caused by a strong gravitational field in their vicinity. One such effect is gravitational lensing: The deformation of spacetime around a massive object causes light rays to be deflected much as light passing through an optic lens. Observations have been made of weak gravitational lensing, in which light rays are deflected by only a few arcseconds. However, it has never been directly observed for a black hole. One possibility for observing gravitational lensing by a black hole would be to observe stars in orbit around the black hole. There are several candidates for such an observation in orbit around Sagittarius A*.

Alternatives

The evidence for stellar black holes strongly relies on the existence of an upper limit for the mass of a neutron star. The size of this limit heavily depends on the assumptions made about the properties of dense matter. New exotic phases of matter could push up this bound. A phase of free quarks at high density might allow the existence of dense quark stars, and some supersymmetric models predict the existence of Q stars. Some extensions of the standard model posit the existence of preons as fundamental building blocks of quarks and leptons, which could hypothetically form preon stars. These hypothetical models could potentially explain a number of observations of stellar black hole candidates. However, it can be shown from arguments in general relativity that any such object will have a maximum mass.

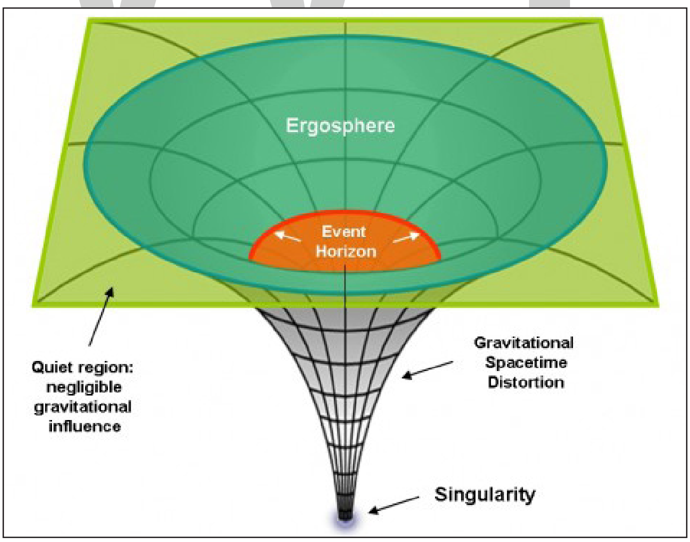
Since the average density of a black hole inside its Schwarzschild radius is inversely proportional to the square of its mass, supermassive black holes are much less dense than stellar black holes (the average density of a $10^8 M_\odot$ black hole is comparable to that of water). Consequently, the physics of matter forming a supermassive black hole is much better understood and the possible alternative explanations for supermassive black hole observations are much more mundane. For example, a supermassive black hole could be modelled by a large cluster of very dark objects. However, such alternatives are typically not stable enough to explain the supermassive black hole candidates.

The evidence for the existence of stellar and supermassive black holes implies that in order for black holes to not form, general relativity must fail as a theory of gravity, perhaps due to the onset of quantum mechanical corrections. A much anticipated feature of a theory of quantum gravity is that it will not feature singularities or event horizons and thus black holes would not be real artifacts. For example, in the fuzzball model based on string theory, the individual states of a black hole solution do not generally have an event horizon or singularity, but for a classical/semi-classical observer the statistical average of such states appears just as an ordinary black hole as deduced from general relativity.

A few theoretical objects have been conjectured to match observations of astronomical black hole candidates identically or near-identically, but which function via a different mechanism. These include the gravastar, the black star, and the dark-energy star.

Sizes of Black Holes

Black holes are singularities: points of infinitely small volume with infinite density. Such incredibly compact objects cause infinite curvature in the fabric of spacetime. Everything that falls into a black hole is sucked toward the singularity. At some distance away from the singularity, the escape velocity exceeds the speed of light, sometimes dramatically dubbed "the point of no return," although the technical term is Schwarzschild radius or event horizon.



Anatomy of a black hole.

There are a couple of different ways to conceptualize how "big" something is. The first is an object's mass (how much matter it contains) and the second is its volume (how much space it takes up). However, the radius of a black hole's event horizon is directly dependent on its mass.

Different types of black holes have very different masses. Stellar-mass black holes are typically in the range of 10 to 100 solar masses, while the supermassive black holes at the centers of galaxies can be millions or billions of solar masses. The supermassive black hole at the center of the Milky Way, Sagittarius A*, is 4.3 million solar masses. This is the only black hole whose mass has been measured directly by observing the full orbit of a circling star. Black holes grow by accreting surrounding matter and by merging with other black holes.

Artist's rendering of a supermassive black hole.

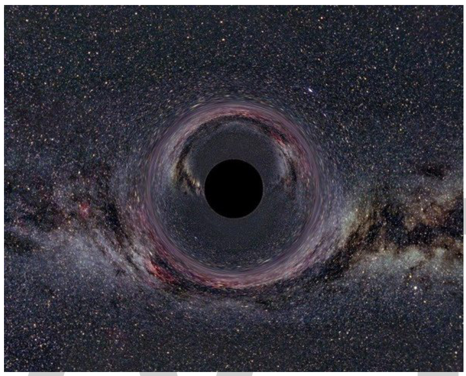
Because there is such a huge leap in sizes of black holes, between stellar-mass and supermassive black holes, it has been hypothesized that a class of *intermediate-mass* black holes also exists. The black holes would be hundreds or thousands of solar masses. There are a couple of candidate intermediate-mass black holes, such as HLX-1, which is estimated to be 20,000 solar masses.

Another hypothetical class of black holes is *primordial* black holes, which would have formed out of density fluctuations in the early universe. Generally, they would have been so tiny (the minimum mass would be the Planck mass) that they can only be properly described using quantum mechanics. But black holes evaporate through a process called Hawking Radiation. How quickly a black hole evaporates depends on its mass: the less massive a black hole, the more quickly it evaporates. For a primordial black hole to have survived to the present day, it would have to contain a few billion tons of mass, with a radius comparable to that of an atomic nucleus.

Creation of Black Holes

A slightly different kind of supernova explosion occurs when even larger, hotter stars (blue giants and blue supergiants) reach the end of their short, dramatic lives. These stars are hot enough to burn not just hydrogen and helium as fuel, but also carbon, oxygen and silicon. Eventually, the fusion in these stars forms the element iron (which is the most stable of all nuclei, and will not easily fuse into heavier elements), which effectively ends the nuclear fusion process within the star. Lacking fuel for fusion, the temperature of the star decreases and the rate of collapse due to gravity increases, until it collapses completely on itself, blowing out material in a massive supernova explosion.

If the mass of the compressed remnant of the star exceeds about 3 - 4 solar masses, then even the degeneracy pressure of neutrons is insufficient to halt the collapse and, instead of forming a neutron star, the core collapses completely into a gravitational singularity, a single point containing all the mass of the entire original star. The gravity in such a phenomenon is so strong that it overwhelms all other forces, to the extent that even light can not escape from it, hence the name black hole. Thus, the gravity of a body just a few times denser than a neutron star would result in its inevitable further collapse into a black hole.



Simulated black hole in front of the Milky Way.

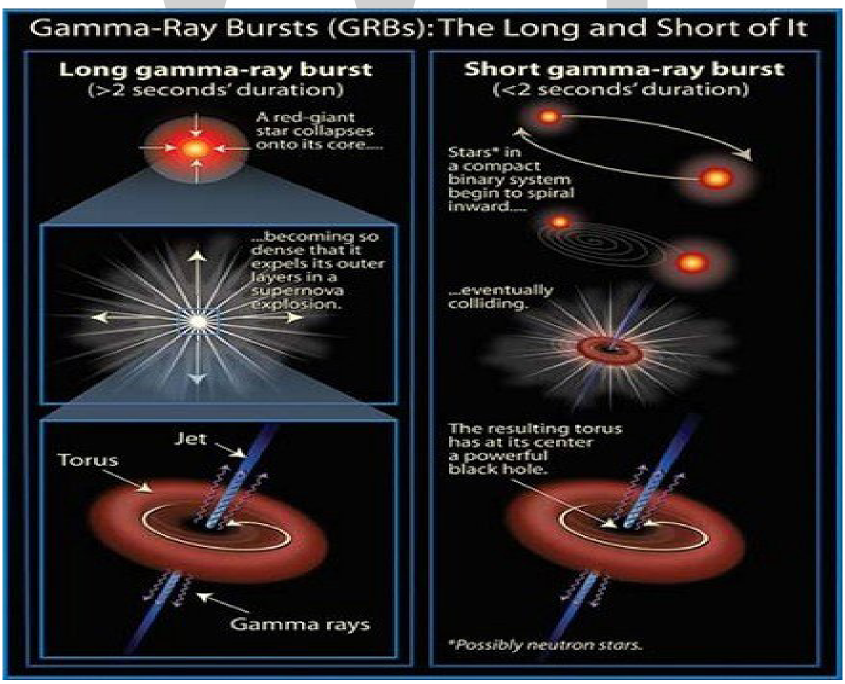
Although singularity at the center of a black hole is infinitely dense, the black hole itself is not necessarily huge, as is sometimes assumed. A black hole with the mass of our Sun, for example, would have a radius of just three kilometers (roughly two hundred million times smaller than the Sun), while one with the mass of the Earth would fit in the palm of your hand! Having said that, black holes can grow to great size over time as they assimilate more and more matter and even other black holes, and some do become extremely massive.

Contrary to popular belief, a black hole does not just “suck up” everything around it in an uncontrolled orgy of destruction: it actually exerts no more gravitational pull on the objects around it than the original star from which it was formed, and any objects orbiting the original star (and which survived the supernova blast) would now orbit a black hole instead (an object would need to approach quite close to a black hole before being sucked in). The very largest blue stars may skip even the supernova stage, so that even their outer shells become incorporated into the singularity.

By definition, we cannot observe black holes directly, but they can be detected by the gravitational effect they exert on other bodies or on light rays. This is especially easy to spot in the case of binary star systems where an ordinary star is orbiting around a black hole. In the early 1990s, Reinhard Genzel pioneered this work, using the then new technique of adaptive optics to plot and track the motions of stars near the center of our own Milky Way galaxy, to show that they must be orbiting a very massive, but invisible, object. From the immense speed with which the stars closest to the center of the galaxy are orbiting - millions of kilometers per hour - we know that there is a “supermassive black hole” (known as Sagittarius A) at the center of the Milky Way, with a mass of around 2 - 4 million times that of our Sun. In addition, in the Milky Way galaxy alone, there are many millions of black holes of at least ten solar masses each.

Supermassive black holes lurk in the centers of most galaxies, forming the hubs around which the galaxies rotate. In fact, from observations of the intense radiation of gases swirling around them at close to the speed of light, we can infer that there are much larger supermassive black holes in the centers of other galaxies, some of them weighing as much as several billion suns. The black hole at the center of a galaxy known as M87 has a mass estimated at around 20 billion solar masses, and may be as large as our entire Solar System.

It seems likely that the early universe, in which very large, short-lived stars were the norm, was scattered with many, many black holes, which gradually merged together over time, creating larger and larger black holes. Observations have shown that is not uncommon for two black holes to swirl around each other in a kind of cosmic dance as their gravitational fields interact. The ripples in space-time caused by two black holes orbiting around each other - typically in a three-leaved clover shape or more complex multi-pass configuration, rather than the simple orbit of an electron within an atom, and ever-smaller and faster as the two objects inevitably approach each other - can be recorded visually and even audibly.



Long and short gamma ray bursts.

In the case of the largest events, moments after the creation of a black hole, the heat and the hugely amplified magnetic field of the collapsing star combine to focus a pair of tight beams or jets of radiation, perpendicular to the spinning plane of the accretion disk. These beams focus vast amounts of particles and energy (of the order of a billion billion times the energy output of our Sun) away from the black hole at close to the speed of light. The shock waves of this massively energetic beam cause gamma rays to be emitted in a phenomenon known as a “gamma ray burst” or “hypernova” event (so named because its energy and brightness dwarfs even that of a supernova, by a factor of up to a hundred million times). Gamma ray bursts are by far the brightest electromagnetic events occurring in the universe, and can last from mere milliseconds to nearly an hour - a typical burst lasts a few seconds - usually followed by a longer-lived “afterglow” emitting at longer wavelengths (x-ray, ultraviolet, visible, infrared and radio waves). It is likely that collisions between neutron stars, or between a neutron star and a black hole, can also cause gamma ray bursts.

Interestingly, it appears to be easier for stars with fewer heavy elements to turn hypernova and generate gamma ray bursts. That, and the fact that larger, more short-lived stars were more common earlier in the life of the universe, mean that the phenomenon of gamma ray bursts is actually rarer today than it was. Having said that, NASA’s Swift Probe, launched in 2004 with a mission specifically to locate gamma ray bursts throughout the universe, is recording at least one such event each day, so these are not rare incidents. (It should be remembered that any supernovas or gamma ray bursts we observe today in galaxies, say, 9 billion light years away, actually occurred 9 billion years ago).

Types of Black Holes

Micro Black Hole

Micro black holes, also called quantum mechanical black holes or mini black holes, are hypothetical tiny black holes, for which quantum mechanical effects play an important role. The concept that black holes may exist that are smaller than stellar mass was introduced in 1971 by Stephen Hawking.

It is possible that such quantum primordial black holes were created in the high-density environment of the early Universe (or Big Bang), or possibly through subsequent phase transitions. They might be observed by astrophysicists through the particles they are expected to emit by Hawking radiation.

Some hypotheses involving additional space dimensions predict that micro black holes could be formed at energies as low as the TeV range, which are available in particle accelerators such as the Large Hadron Collider. Popular concerns have then been raised over end-of-the-world scenarios. However, such quantum black holes would instantly evaporate, either totally or leaving only a very weakly interacting residue. Beside the theoretical arguments, the cosmic rays hitting the Earth do not produce any damage, although they reach energies in the range of hundreds of TeV.

Stability

Hawking Radiation

In 1975, Stephen Hawking argued that, due to quantum effects, black holes "evaporate" by a process now referred to as Hawking radiation in which elementary particles (such as photons, electrons, quarks, gluons) are emitted. His calculations showed that the smaller the size of the black hole, the faster the evaporation rate, resulting in a sudden burst of particles as the micro black hole suddenly explodes.

Any primordial black hole of sufficiently low mass will evaporate to near the Planck mass within the lifetime of the Universe. In this process, these small black holes radiate away matter. A rough picture of this is that pairs of virtual particles emerge from the vacuum near the event horizon, with one member of a pair being captured, and the other escaping the vicinity of the black hole. The net result is the black hole loses mass (due to conservation of energy). According to the formulae of black hole thermodynamics, the more the black hole loses mass, the hotter it becomes, and the faster it evaporates, until it approaches the Planck mass. At this stage, a black hole would have a Hawking temperature of $\frac{Tp}{8\pi}\left(5.6\times10^{32}\,K\right)$ which means an emitted Hawking particle would have an energy comparable to the mass of the black hole. Thus, a thermodynamic description breaks down. Such a micro black hole would also have an entropy of only 4π nats, approximately the minimum possible value. At this point then, the object can no longer be described as a classical black hole, and Hawking's calculations also break down.

Conjectures for the Final State

Conjectures for the final fate of the black hole include total evaporation and production of a Planck-mass-sized black hole remnant. Such Planck-mass black holes may in effect be stable objects if the quantized gaps between their allowed energy levels bar them from emitting Hawking particles or absorbing energy gravitationally like a classical black hole. In such case, they would be weakly interacting massive particles; this could explain dark matter.

Man-made Micro Black Holes

Feasibility of Production

In familiar three-dimensional gravity, the minimum energy of a microscopic black hole is 10^{19} GeV (equivalent to 1.6 GJ or 444 kWh), which would have to be condensed into a region on the order of the Planck length. This is far beyond the limits of any current technology. It is estimated that to collide two particles to within a distance of a Planck length with currently achievable magnetic field strengths would require a ring accelerator about 1,000 light years in diameter to keep the particles on track. Stephen Hawking also said in chapter 6 of his *A Brief History of Time* that physicist John Archibald Wheeler once calculated that a very powerful hydrogen bomb using all the deuterium in all the water on Earth could also generate such a black hole, but Hawking does not provide this calculation or any reference to it to support this assertion.

However, in some scenarios involving extra dimensions of space, the Planck mass can be as low as the TeV range. The Large Hadron Collider (LHC) has a design energy of 14 TeV for proton–proton collisions and 1,150 TeV for Pb–Pb collisions. It was argued in 2001 that, in these circumstances, black hole production could be an important and observable effect at the LHC or future higher-energy colliders. Such quantum black holes should decay emitting sprays of particles that could be seen by detectors at these facilities. A paper by Choptuik and Pretorius, published in 2010 in *Physical Review Letters*, presented a computer-generated proof that micro black holes must form from two colliding particles with sufficient energy, which might be allowable at the energies of the LHC if additional dimensions are present other than the customary four (three spatial, one temporal).

Safety Arguments

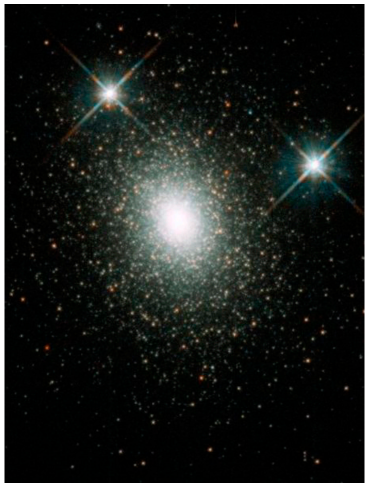
Hawking's calculation and more general quantum mechanical arguments predict that micro black holes evaporate almost instantaneously. Additional safety arguments beyond those based on Hawking radiation were given in the paper, which showed that in hypothetical scenarios with stable black holes that could damage Earth, such black holes would have been produced by cosmic rays and would have already destroyed known astronomical objects such as the Earth, Sun, neutron stars, or white dwarfs.

Black Holes in Quantum Theories of Gravity

It is possible, in some theories of quantum gravity, to calculate the quantum corrections to ordinary, classical black holes. Contrarily to conventional black holes, which are solutions of gravitational field equations of the general theory of relativity, quantum gravity black holes incorporate quantum gravity effects in the vicinity of the origin, where classically a curvature singularity occurs. According to the theory employed to model quantum gravity effects, there are different kinds of quantum gravity black holes, namely loop quantum black holes, non-commutative black holes, asymptotically safe black holes. In these approaches, black holes are singularity-free.

Virtual micro black holes have been proposed by Stephen Hawking in 1995, and by Fabio Scardigli in 1999 as part of a Grand Unified Theory which could be a quantum gravity candidate.

Intermediate-mass Black Hole



Globular cluster Mayall II (M31 G1) is a possible candidate for hosting an intermediate-mass black hole at its center.

An intermediate-mass black hole (IMBH) is a class of black hole with mass in the range 102-105 solar masses: significantly more than stellar black holes but less than the 105-109 solar mass supermassive black holes. Several IMBH candidate objects have been discovered in our galaxy and others nearby, based on indirect gas cloud velocity and accretion disk spectra observations of various evidentiary strength.

Observational Evidence

The strongest evidence for IMBHs comes from a few low-luminosity active galactic nuclei. Due to their activity, these galaxies almost certainly contain accreting black holes, and in some cases the black hole masses can be estimated using the technique of reverberation mapping. For instance, the spiral galaxy NGC 4395 at a distance of about 4 Mpc appears to contain a black hole with mass of about 3.6×10^5 solar masses.

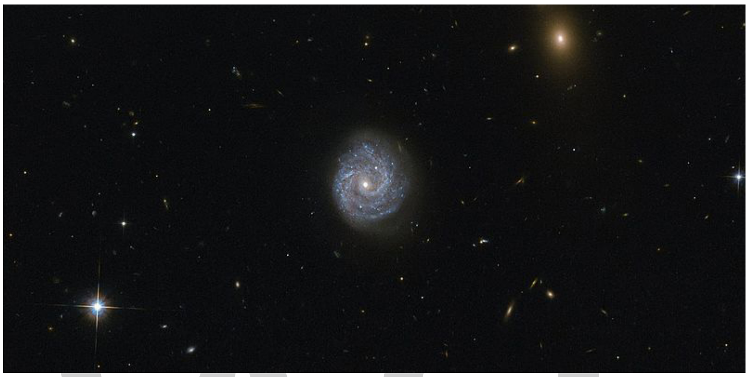
Some ultra-luminous X ray sources (ULXs) in nearby galaxies are suspected to be IMBHs, with masses of a hundred to a thousand solar masses. The ULXs are observed in star-forming regions (e.g., in starburst galaxy M82), and are seemingly associated with young star clusters which are also observed in these regions. However, only a dynamical mass measurement from the analysis of the optical spectrum of the companion star can unveil the presence of an IMBH as the compact accretor of the ULX.

A few globular clusters have been claimed to contain IMBHs, based on measurements of the velocities of stars near their centers; the figure shows one candidate object. However none of the claimed detections has stood up to scrutiny. For instance, the data for M31 G1, the object shown in the figure, can be fit equally well without a massive central object.

Additional evidence for the existence of IMBHs can be obtained from observation of gravitational radiation, emitted from a binary containing an IMBH and a compact remnant or another IMBH.

Finally, the M-sigma relation predicts the existence of black holes with masses of 10^4 to 10^6 solar masses in low-luminosity galaxies.

Potential Discoveries



RX J1140.1+0307 is a spiral galaxy, centered on a lighter, intermediate-mass black hole.

In November 2004 a team of astronomers reported the discovery of GCIRS 13E, the first intermediate-mass black hole in our galaxy, orbiting three light-years from Sagittarius A*. This medium black hole of 1,300 solar masses is within a cluster of seven stars, possibly the remnant of a massive star cluster that has been stripped down by the Galactic Center. This observation may add support to the idea that supermassive black holes grow by absorbing nearby smaller black holes and stars. However, in 2005, a German research group claimed that the presence of an IMBH near the galactic center is doubtful, based on a dynamical study of the star cluster in which the IMBH was said to reside. An IMBH near the galactic center could also be detected via its perturbations on stars orbiting around the supermassive black hole.

In January 2006 a team led by Philip Kaaret of the University of Iowa announced the discovery of a quasiperiodic oscillation from an intermediate-mass black hole candidate located using NASA's Rossi X-ray Timing Explorer. The candidate, M82 X-1, is orbited by a red giant star that is shedding its atmosphere into the black hole. Neither the existence of the oscillation nor its interpretation as the orbital period of the system are fully accepted by the rest of the scientific community. While the interpretation is quite reasonable, the periodicity claimed is based on only about 4 cycles, meaning that it is quite possible for this to be random variation. If the period is real, it could be either the orbital period, as suggested, or a super-orbital period in the accretion disk, as is seen in many other systems.

In 2009, a team of astronomers led by Sean Farrell discovered HLX-1, an intermediate-mass black hole with a smaller cluster of stars around it, in the galaxy ESO 243-49.

This evidence suggested that ESO 243-49 had a galactic collision with HLX-1's galaxy and absorbed the majority of the smaller galaxy's matter.

A team at the CSIRO radio telescope in Australia announced on 9 July 2012 that it had discovered the first intermediate-mass black hole.

In 2015 a team at Keio University in Japan found a gas cloud (CO-0.40-0.22) with very wide velocity dispersion. They performed simulations and concluded that a model with a black hole of circa 100,000 solar masses would be the best fit for the velocity distribution. However, a later work pointed out some difficulties with the association of high velocity dispersion clouds with intermediate mass black holes, and proposed that such clouds might be generated by supernovae. Radio observations with the Atacama Large Millimeter/submillimeter Array confirmed absence of an IMBH near CO-0.40-0.22, and found that the large velocity dispersion of the cloud is created by superposition of two molecular clouds with different line-of-sight velocities. The features identified as signatures of IMBH-cloud interaction in previous studies were confirmed to be artifacts created by erroneous data reduction . Further theoretical studies of the gas cloud and nearby IMBH candidates have been inconclusive but have re-opened the possibility, though no observational evidence for existence of an IMBH has been reported after that.

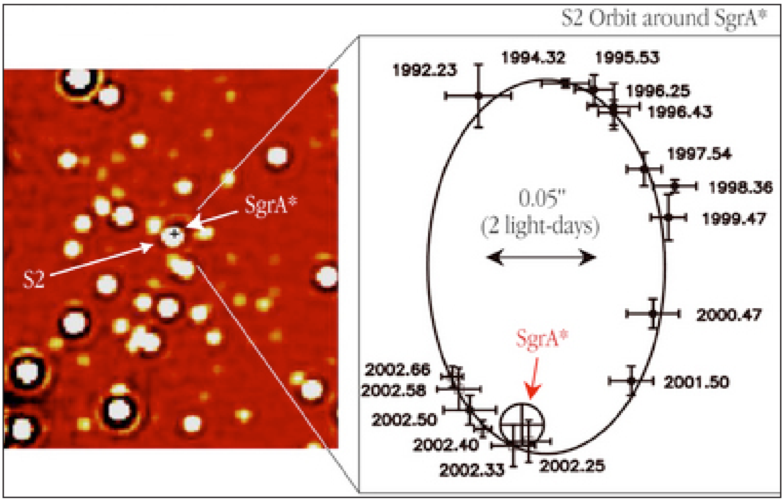
In 2017, it was announced that a black hole of a few thousand solar masses may be located in the globular cluster 47 Tucanae. This was based on the accelerations and distributions of pulsars in the cluster, however later analysis of an updated and more complete data set on these pulsars finds no positive evidence for this.

Observations in 2018 of several molecular gas streams orbiting around an invisible object near the galactic center, designated HCN-0.009-0.044, suggested it is a 32,000 solar mass black hole, and if so is the third IMBH discovered in the region.

Intermediate-mass black holes are too massive to be formed by the collapse of a single star, which is how stellar black holes are thought to form. Their environments lack the extreme conditions—i.e., high density and velocities observed at the centers of galaxies—which seemingly lead to the formation of supermassive black holes. There are three postulated formation scenarios for IMBHs. The first is the merging of stellar mass black holes and other compact objects by means of accretion. The second one is the runaway collision of massive stars in dense stellar clusters and the collapse of the collision product into an IMBH. The third is that they are primordial black holes formed in the Big Bang.

Supermassive Black Hole

As the name suggests, supermassive black holes contain between a million and a billion times more mass than a typical stellar black hole. Although there are only a handful of confirmed supermassive black holes (most are too far away to be observed), they are thought to exist at the centre of most large galaxies, including the centre of our own galaxy, the Milky Way.



Direct evidence for a supermassive black hole – a plot of the orbital motion of the star S2 around the centre of the Milky Way. From these observations, astronomers have inferred that a supermassive black hole of about 3 million solar masses lurks at the centre of our galaxy.

For many years, astronomers had only indirect evidence for supermassive black holes, the most compelling of which was the existence of quasars in remote active galaxies. Observations of the energy output and variability timescales of quasars revealed that they radiate over a trillion times as much energy as our Sun from a region about the size of the Solar System. The only mechanism capable of producing such enormous amounts of energy is the conversion of gravitational energy into light by a massive black hole.

More recently, direct evidence for the existence of supermassive black holes has come from observations of material orbiting the centres of galaxies. The high orbital velocities of these stars and gas are easily explained if they are being accelerated by a massive object with a strong gravitational field that is contained within a small region of space – i.e. a supermassive black hole.

The jet emitted by the galaxy M87 is thought to be powered by a supermassive black hole at the galaxy's centre.

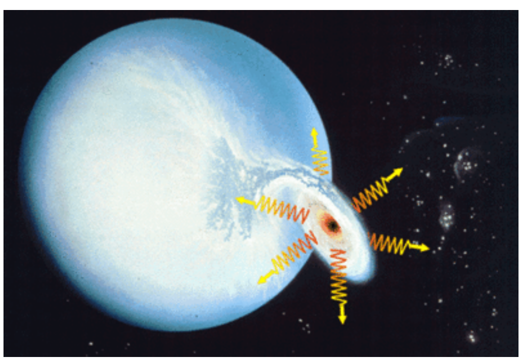
Astronomers are still not sure how these supermassive black holes form. Stellar black holes result from the collapse of massive stars, and some have suggested that supermassive black holes form out of the collapse of massive clouds of gas during the early stages of the formation of the galaxy. Another idea is that a stellar black hole consumes enormous amounts of material over millions of years, growing to supermassive black hole proportions. Yet another is that a cluster of stellar black holes form and eventually merge into a supermassive black hole.

Whatever their formation mechanism, most astronomers agree that accretion of material onto the supermassive black hole drives both active galactic nuclei and galactic jets.

Stellar Black Hole

Stellar black holes, with masses less than about 100 times that of the Sun, comprise one of the possible evolutionary endpoints of high mass stars. Once the core of the star has completely burned to iron, energy production stops and the core rapidly collapses resulting in a supernova explosion. If the core is greater than about 2-3 solar masses (the maximum mass of a neutron star), the pressure of neutrons is unable to stop the collapse and a stellar black hole is formed. These black holes are generally modelled as Kerr black holes, as it is expected that the original rotation of the massive star would be conserved during the collapse, and that black holes contain little electric charge.

Since radiation cannot escape the extreme gravitational pull of a black hole once it crosses the event horizon, it is very difficult to discover one in isolation. Stellar black holes are therefore most easily found in X-ray binary systems, where gas from a companion star is being pulled into the black hole. X-rays are produced by this gas which is heated to tens of millions of Kelvin as it spirals towards the black hole via an accretion disk. Astronomers can also measure the mass of the black hole (typically between 3 and 20 solar masses for a stellar black hole) by observing its gravitational effect on the companion star.



An artist's impression of an X-ray binary system. Gas from the companion star is pulled into an accretion disk around the stellar black hole. As the gas spirals towards the black hole, it is heated to tens of millions of degrees and emits copious amounts of X-rays.

There are currently around 20 X-ray binary systems that are thought to contain stellar black holes, though this number continues to climb as the sensitivity of instruments improves and more observations are made. Two of the best candidates (it is very difficult to unequivocally confirm a black hole) are LMC X-3 in the Large Magellanic Cloud and Cygnus X-1 (the first stellar black hole candidate) discovered in 1965.

Primordial Black Hole

Primordial black holes are a hypothetical type of black hole that formed soon after the Big Bang. In the early universe, high densities and heterogeneous conditions could have led sufficiently dense regions to undergo gravitational collapse, forming black holes. Yakov Borisovich Zel'dovich and Igor Dmitriyevich Novikov in 1966 first proposed the existence of such black holes. The theory behind their origins was first studied in depth by Stephen Hawking in 1971. Since primordial black holes did not form from stellar gravitational collapse, their masses can be far below stellar mass (c. 2×1030 kg). Hawking calculated that primordial black holes could weigh as little as 10–8 kg.

Depending on the model, primordial black holes could have initial masses ranging from 10^{-8} kg (the so-called Planck relics) to more than thousands of solar masses. However, primordial black holes originally having mass lower than 10^{11} kg would not have survived to the present due to Hawking radiation, which causes complete evaporation in a time much shorter than the age of the Universe. Primordial black holes are non-baryonic and as such are plausible dark matter candidates. Primordial black holes are also good candidates for being the seeds of the supermassive black holes at the center of massive galaxies, as well as of intermediate-mass black holes.

Primordial black holes belong to the class of massive compact halo objects (MACHOs). They are naturally a good dark matter candidate: they are (nearly) collision-less and stable (if sufficiently massive), they have non-relativistic velocities, and they form very early in the history of the Universe (typically less than one second after the Big Bang). Nevertheless, tight limits on their abundance have been set up from various astrophysical and cosmological observations, so that it is now excluded that they contribute significantly to dark matter over most of the plausible mass range.

In March 2016, one month after the announcement of the detection by Advanced LIGO/VIRGO of gravitational waves emitted by the merging of two 30 solar mass black holes (about 6×10^{31} kg), three groups of researchers proposed independently that the detected black holes had a primordial origin. Two of the groups found that the merging rates inferred by LIGO are consistent with a scenario in which all the dark matter is made of primordial black holes, if a non-negligible fraction of them are somehow clustered within halos such as faint dwarf galaxies or globular clusters, as expected by the standard theory of cosmic structure formation. The third group claimed that these merging rates are incompatible with an all-dark-matter scenario and that primordial black holes could only contribute to less than one percent of the total dark matter. The unexpected large mass of the black holes detected by LIGO has strongly revived interest in primordial black holes with masses in the range of 1 to 100 solar masses. It is however still debated whether this range is excluded or not by other observations, such as the absence of micro-lensing of stars, the cosmic microwave background anisotropies, the size of faint dwarf galaxies, and the absence of correlation between X-ray and radio sources towards the galactic center.

In May 2016, Alexander Kashlinsky suggested that the observed spatial correlations in the unresolved gamma-ray and X-ray background radiations could be due to primordial black holes with similar masses, if their abundance is comparable to that of dark matter.

In April 2019, a study was published suggesting this hypothesis may be a dead end. An international team of researchers has put a theory speculated by the late Stephen Hawking to its most rigorous test to date, and their results have ruled out the possibility that primordial black holes smaller than a tenth of a millimeter make up most of dark matter.

In September 2019, a report by James Unwin and Jakub Scholtz proposed the possibility of a primordial black hole (PBH) the size of a tennis ball existing in the extended Kuiper Belt to explain the orbital anomalies that are theorized to be the result of a 9th planet in the solar system.

Formation

Primordial black holes could have formed in the very early Universe (less than one second after the Big Bang), during the so-called *radiation dominated era*. The essential ingredient for the formation of a primordial black hole is a fluctuation in the density of the Universe, inducing its gravitational collapse. One typically requires density contrasts $\delta\rho/\rho \sim 0.1$ (where ρ is the density of the Universe) to form a black hole. There are several mechanisms able to produce such inhomogeneities in the context of cosmic inflation (in hybrid inflation models, for example axion inflation), reheating, or cosmological phase transitions.

Observational Limits and Detection Strategies

A variety of observations have been interpreted to place limits on the abundance and mass of primordial black holes:

- Lifetime, Hawking radiation and gamma-rays: One way to detect primordial black holes, or to constrain their mass and abundance, is by their Hawking radiation. Stephen Hawking theorized in 1974 that large numbers of such smaller primordial black holes might exist in the Milky Way in our galaxy's halo region. All black holes are theorized to emit Hawking radiation at a rate inversely proportional to their mass. Since this emission further decreases their mass, black holes with very small mass would experience runaway evaporation, creating a massive burst of radiation at the final phase, equivalent to a hydrogen bomb yielding millions of megatons of explosive force. A regular black hole (of about 3 solar masses) cannot lose all of its mass within the current age of the universe (they would take about 10^{69} years to do so, even without any matter falling in). However, since primordial black holes are not formed by stellar core collapse, they may be of any size. A black hole with a mass of about 10^{11} kg would have a lifetime about equal to the age of the universe. If such low-mass black holes were created in sufficient number in the Big Bang, we should be able to observe explosions by some of those that are relatively nearby in our own Milky Way galaxy. NASA's Fermi Gamma-ray Space Telescope satellite, launched in June 2008, was designed in part to search for such evaporating primordial black holes. Fermi data set up the limit that less than one percent of dark matter could be made of primordial black holes with masses up to 10^{13} kg. Evaporating primordial black holes would have also had an impact on the Big Bang nucleosynthesis and change the abundances of light elements in the Universe. However, if theoretical Hawking radiation does not actually exist, such primordial black holes would be extremely difficult, if not impossible, to detect in space due to their small size and lack of large gravitational influence.

- Lensing of gamma-ray bursts: Compact objects can induce a change in the luminosity of gamma-ray bursts when passing close to their line-of-sight, through the gravitational lensing effect. The Fermi Gamma-Ray Burst Monitor experiment found that primordial black holes cannot contribute importantly to the dark matter within the mass range $5 \times 10^{14} - 10^{17}$ kg.

- Capture of primordial black holes by neutron stars: If primordial black holes with masses between 10^{15} kg and 10^{22} kg had abundances comparable to that of dark matter, neutron stars in globular clusters should have captured some of them, leading to the rapid destruction of the star. The observation of neutron stars in globular clusters can thus be used to set a limit on primordial black hole abundance.

- Micro-lensing of stars: If a primordial black hole passes between us and a distant star, it induces a magnification of these stars due to the gravitational lensing effect. By monitoring the magnitude of stars in the Magellanic Clouds, the EROS and MACHO surveys have put a limit on the abundance of primordial black holes in the range $10^{23} - 10^{31}$ kg. According to these surveys, primordial black holes within this range cannot constitute an important fraction of the dark matter. However, these limits are model-dependent. It has been also argued that if primordial black holes are regrouped in dense halos, the micro-lensing constraints are then naturally evaded.

- Micro-lensing of Ia supernovae: Primordial black holes with masses larger than 10^{28} kg would magnify distant type Ia supernova (or any other standard candle of known luminosity) due to gravitational lensing. These effects would be apparent if primordial black holes were a significant contribution to the dark matter density, which is constrained by current data sets.

- Temperature anisotropies in the cosmic microwave background: Accretion of matter onto primordial black holes in the early Universe should lead to energy injection in the medium that affects the recombination history of the Universe. This effect induces signatures in the statistical distribution of the cosmic microwave background (CMB) anisotropies. The Planck observations of the CMB exclude that primordial black holes with masses in the range $100 - 10^4$ solar masses contribute importantly to the dark matter, at least in the simplest conservative model. It is still debated whether the constraints are stronger or weaker in more realistic or complex scenarios.

At the time of the detection by LIGO of the gravitational waves emitted during the final coalescence of two 30 solar mass black holes, the mass range between 10 and 100 solar masses was still only poorly constrained. Since then, new observations have been claimed to close this window, at least for models in which the primordial black holes have all the same mass:

- From the absence of X-ray and optical correlations in point sources observed in the direction of the galactic center.
- From the dynamical heating of dwarf galaxies.
- From the observation of a central star cluster in the Eridanus II dwarf galaxy (but these constraints can be relaxed if Eridanus II owns a central intermediate mass black hole, which is suggested by some observations). If primordial black holes exhibit a broad mass distribution, those constraints could nevertheless still be evaded.
- From the gravitational micro-lensing of distant quasars by closer galaxies, allowing only 20% of the galactic matter to be in the form of compact objects with stellar masses, a value consistent with the expected stellar population.
- From micro-lensing of distant stars by galaxy clusters, suggesting that the fraction of dark matter in the form of primordial black holes with masses comparable to those found by LIGO must be less than 10%.

In the future, new limits will be set up by various observations:

- The Square Kilometre Array (SKA) radio telescope will probe the effects of primordial black holes on the reionization history of the Universe, due to energy injection into the intergalactic medium, induced by matter accretion onto primordial black holes.
- LIGO, VIRGO and future gravitational waves detectors will detect new black

hole merging events, from which one could reconstruct the mass distribution of primordial black holes. These detectors could allow distinguishing unambiguously between primordial or stellar origins if merging events involving black holes with a mass lower than 1.4 solar mass are detected. Another way would be to measure the large orbital eccentricity of primordial black hole binaries.

- Gravitational wave detectors, such as the Laser Interferometer Space Antenna (LISA) and pulsar timing arrays will also probe the stochastic background of gravitational waves emitted by primordial black hole binaries, when they are still orbiting relatively far from each other.

- New detections of faint dwarf galaxies, and the observations of their central star cluster, could be used to test the hypothesis that these dark matter-dominated structures contain primordial black holes in abundance.

- Monitoring star positions and velocities within the Milky Way could be used to detect the influence of a nearby primordial black hole.

- It has been suggested that a small black hole passing through the Earth would produce a detectable acoustic signal. Because of its tiny diameter, large mass compared to a nucleon, and relatively high speed, such primordial black holes would simply transit Earth virtually unimpeded with only a few impacts on nucleons, exiting the planet with no ill effects.

- Another way to detect primordial black holes could be by watching for ripples on the surfaces of stars. If the black hole passed through a star, its density would cause observable vibrations.

- Monitoring quasars in the microwave wavelength and detection of the wave optics feature of gravitational microlensing by the primordial black holes.

Implications

The evaporation of primordial black holes has been suggested as one possible explanation for gamma-ray bursts. This explanation is, however, considered unlikely. Other problems for which primordial black holes have been suggested as a solution include the dark matter problem, the cosmological domain wall problem and the cosmological monopole problem. Since a primordial black hole does not necessarily have to be small (they can have any size), primordial black holes may also have contributed to the later formation of galaxies.

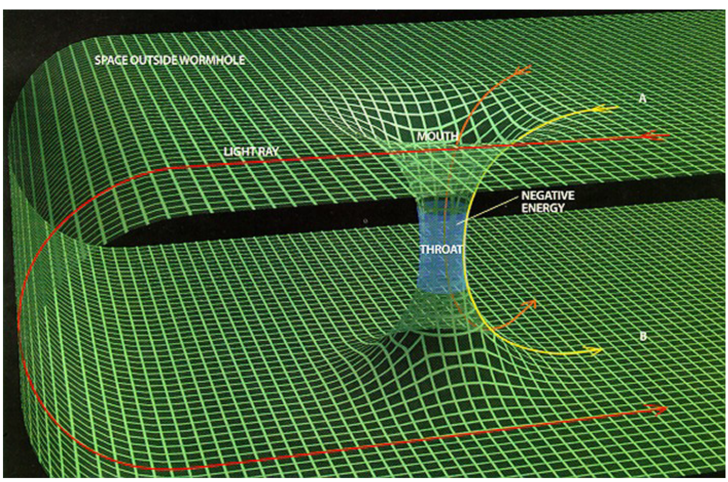
Even if they do not solve these problems, the low number of primordial black holes (as of 2010, only two intermediate mass black holes were confirmed) aids cosmologists by putting constraints on the spectrum of density fluctuations in the early universe.

Wormholes

Like black holes, wormholes arise as valid solutions to the equations of Albert Einstein's General Theory of Relativity, and, like black holes, the phrase was coined by the American physicist John Wheeler. Also like black holes, they have never been observed directly, but they crop up so readily in theory that some physicists are encouraged to think that real counterparts may eventually be found or fabricated.

In 1916, the Austrian physicist Ludwig Flamm, while looking over Karl Schwarzschild's solution to Einstein's field equations, which describes a particular form of black hole known as a Schwarzschild black hole, noticed that another solution was also possible, which described a phenomenon which later came to be known as a "white hole". A white hole is the theoretical time reversal of a black hole and, while a black hole acts as a vacuum, drawing in any matter that crosses the event horizon, a white hole acts as a source that ejects matter from its event horizon. Some have even speculated that there is a white hole on the "other side" of all black holes, where all the matter the black hole sucks up is blown out in some alternative universe, and even that what we think of as the Big Bang might in fact have been the result of just such a phenomenon.

Flamm also noticed that the two solutions, describing two different regions of space-time could be mathematically connected by a kind of space-time conduit, and that, in theory at least, the black hole "entrance" and white hole "exit" could be in totally different parts of the same universe or even in different universes. Einstein himself explored these ideas further in 1935, along with Nathan Rosen, and the two achieved a solution known as an Einstein-Rosen bridge (also known as a Lorentzian wormhole or a Schwarzschild wormhole).



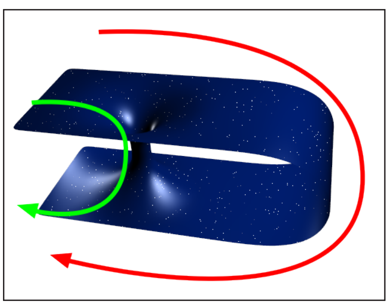
A wormhole is a theoretical "short-cut" between distant regions of space-time.

To better visualize a wormhole, consider the analogy of a piece of paper with two pencil marks drawn on it (to represent two points in space-time), the line between them showing the distance from one point to the other in normal space-time. If the paper is now bent and folded over almost double (the equivalent of drastically warping space-time), then poking the pencil through the paper provides a much shorter way of linking the two points, a short-cut through space-time much like a wormhole.

Some theorists are encouraged to think that real counterparts may eventually be found or fabricated and, perhaps, used as a tunnel or short-cut for high-speed space travel between distant points or even for time travel (with all the potential paradoxes that might entail). However, a generally accepted property of wormholes is that they are inherently highly unstable and would probably collapse in a much shorter time than it would take to get through to the other side. At any rate, it is predicted that they would collapse instantly if even the tiniest amount of matter (even a single photon) attempted to pass through them.

Although some possible theoretical ways around this problem have been suggested (for example, using "cosmic strings" or "negative matter" or some other exotic matter with "negative energy") to prevent the wormhole from pinching closed, the idea remains largely in the realm of science fiction for the time being. It has, however, still not been mathematically proven beyond all doubt that some kind of exotic matter with negative energy density is an absolute requirement for wormholes, nor has it been established that such exotic matter cannot exist, so the possibility of a practical application of the theory still remains.

Because a wormholes is a conduit through 4-dimensional space-time, and not just through space, Stephen Hawking and others have also posited that wormholes might theoretically be utilized for travel through time as well as through space, although it is widely believed that time travel into the past will never be possible due to the potential for paradoxes and self-destructive feedback loops.



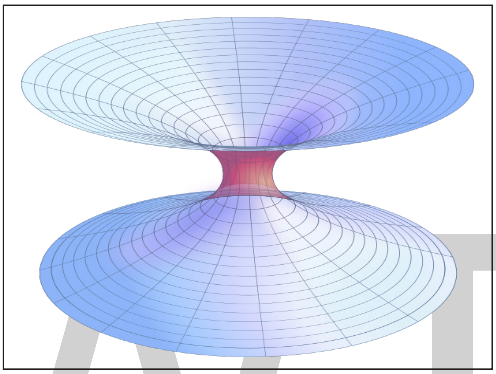
Wormhole visualized in 2D.

For a simplified notion of a wormhole, space can be visualized as a two-dimensional (2D) surface. In this case, a wormhole would appear as a hole in that surface, lead into a 3D tube (the inside surface of a cylinder), then re-emerge at another location on the 2D surface with a hole similar to the entrance. An actual wormhole would be analogous to this, but with the spatial dimensions raised by one. For example, instead of circular holes on a 2D plane, the entry and exit points could be visualized as spheres in 3D space.

Another way to imagine wormholes is to take a sheet of paper and draw two somewhat distant points on one side of the paper. The sheet of paper represents a plane in the spacetime continuum, and the two points represent a distance to be traveled, however theoretically a wormhole could connect these two points by folding that plane so the points are touching. In this way it would be much easier to traverse the distance since the two points are now touching.

Development

Schwarzschild Wormholes



“Embedding diagram” of a Schwarzschild wormhole.

The equations of the theory of general relativity have valid solutions that contain wormholes. The first type of wormhole solution discovered was the *Schwarzschild wormhole*, which would be present in the Schwarzschild metric describing an *eternal black hole*, but it was found that it would collapse too quickly for anything to cross from one end to the other. Wormholes that could be crossed in both directions, known as traversable wormholes, would be possible only if exotic matter with negative energy density could be used to stabilize them.

Einstein-Rosen Bridges

Schwarzschild wormholes, also known as *Einstein–Rosen bridges* (named after Albert Einstein and Nathan Rosen), are connections between areas of space that can be modeled as vacuum solutions to the Einstein field equations, and that are now understood to be intrinsic parts of the maximally extended version of the Schwarzschild metric describing an eternal black hole with no charge and no rotation. Here, “maximally extended” refers to the idea that the spacetime should not have any “edges”: It should be possible to continue this path arbitrarily far into the particle’s future or past for any possible trajectory of a free-falling particle (following a geodesic in the spacetime).

In order to satisfy this requirement, it turns out that in addition to the black hole interi-or region that particles enter when they fall through the event horizon from the outside, there must be a separate white hole interior region that allows us to extrapolate the trajectories of particles that an outside observer sees rising up away from the event horizon. And just as there are two separate interior regions of the maximally extended spacetime, there are also two separate exterior regions, sometimes called two different "universes", with the second universe allowing us to extrapolate some possible particle trajectories in the two interior regions. This means that the interior black hole region can contain a mix of particles that fell in from either universe (and thus an observer who fell in from one universe might be able to see light that fell in from the other one), and likewise particles from the interior white hole region can escape into either universe. All four regions can be seen in a spacetime diagram that uses Kruskal–Szekeres coordinates.

In this spacetime, it is possible to come up with coordinate systems such that if a hypersurface of constant time (a set of points that all have the same time coordinate, such that every point on the surface has a space-like separation, giving what is called a 'space-like surface') is picked and an "embedding diagram" drawn depicting the curvature of space at that time, the embedding diagram will look like a tube connecting the two exterior regions, known as an "Einstein–Rosen bridge". Note that the Schwarzschild metric describes an idealized black hole that exists eternally from the perspective of external observers; a more realistic black hole that forms at some particular time from a collapsing star would require a different metric. When the infalling stellar matter is added to a diagram of a black hole's history, it removes the part of the diagram corresponding to the white hole interior region, along with the part of the diagram corresponding to the other universe.

The Einstein–Rosen bridge was discovered by Ludwig Flamm in 1916, a few months after Schwarzschild published his solution, and was rediscovered by Albert Einstein and his colleague Nathan Rosen, who published their result in 1935. However, in 1962, John Archibald Wheeler and Robert W. Fuller published a paper showing that this type of wormhole is unstable if it connects two parts of the same universe, and that it will pinch off too quickly for light (or any particle moving slower than light) that falls in from one exterior region to make it to the other exterior region.

According to general relativity, the gravitational collapse of a sufficiently compact mass forms a singular Schwarzschild black hole. In the Einstein–Cartan–Sciama–Kibble theory of gravity, however, it forms a regular Einstein–Rosen bridge. This theory extends general relativity by removing a constraint of the symmetry of the affine connection and regarding its antisymmetric part, the torsion tensor, as a dynamical variable. Torsion naturally accounts for the quantum-mechanical, intrinsic angular momentum (spin) of matter. The minimal coupling between torsion and Dirac spinors generates a repulsive spin–spin interaction that is significant in fermionic matter at extremely high densities. Such an interaction prevents the formation of a gravitational singularity. Instead, the collapsing matter reaches an enormous but finite density and rebounds, forming the other side of the bridge.

Although Schwarzschild wormholes are not traversable in both directions, their existence inspired Kip Thorne to imagine traversable wormholes created by holding the "throat" of a Schwarzschild wormhole open with exotic matter (material that has negative mass/energy).

Other non-traversable wormholes include *Lorentzian wormholes*, wormholes creating a spacetime foam in a general relativistic spacetime manifold depicted by a Lorentzian manifold, and *Euclidean wormholes* (named after Euclidean manifold, a structure of Riemannian manifold).

Traversable Wormholes

The Casimir effect shows that quantum field theory allows the energy density in certain regions of space to be negative relative to the ordinary matter vacuum energy, and it has been shown theoretically that quantum field theory allows states where energy can be *arbitrarily* negative at a given point. Many physicists, such as Stephen Hawking, Kip Thorne, and others, argued that such effects might make it possible to stabilize a traversable wormhole. The only known natural process that is theoretically predicted to form a wormhole in the context of general relativity and quantum mechanics was put forth by Leonard Susskind in his ER=EPR conjecture. The quantum foam hypothesis is sometimes used to suggest that tiny wormholes might appear and disappear spontaneously at the Planck scale, and stable versions of such wormholes have been suggested as dark matter candidates. It has also been proposed that, if a tiny wormhole held open by a negative mass cosmic string had appeared around the time of the Big Bang, it could have been inflated to macroscopic size by cosmic inflation.

The image above shows a simulated traversable wormhole that connects the square in front of the physical institutes of University of Tübingen with the sand dunes near Boulogne sur Mer in the north of France. The image is calculated with 4D raytracing in a Morris–Thorne wormhole metric, but the gravitational effects on the wavelength of light have not been simulated.

Lorentzian traversable wormholes would allow travel in both directions from one part of the universe to another part of that same universe very quickly or would allow travel from one universe to another. The possibility of traversable wormholes in general relativity was first demonstrated in a 1973 paper by Homer Ellis and independently in a 1973 paper by K. A. Bronnikov. Ellis analyzed the topology and the geodesics of the

Ellis drainhole, showing it to be geodesically complete, horizonless, singularity-free, and fully traversable in both directions. The drainhole is a solution manifold of Einstein's field equations for a vacuum space-time, modified by inclusion of a scalar field minimally coupled to the Ricci tensor with antiorthodox polarity (negative instead of positive). (Ellis specifically rejected referring to the scalar field as 'exotic' because of the antiorthodox coupling, finding arguments for doing so unpersuasive.) The solution depends on two parameters: m, which fixes the strength of its gravitational field, and n, which determines the curvature of its spatial cross sections. When m is set equal to 0, the drainhole's gravitational field vanishes. What is left is the Ellis wormhole, a non-gravitating, purely geometric, traversable wormhole.

Kip Thorne and his graduate student Mike Morris, unaware of the 1973 papers by Ellis and Bronnikov, manufactured, and in 1988 published, a duplicate of the Ellis wormhole for use as a tool for teaching general relativity. For this reason, the type of traversable wormhole they proposed, held open by a spherical shell of exotic matter, was from 1988 to 2015 referred to in the literature as a *Morris–Thorne wormhole*. Later, other types of traversable wormholes were discovered as allowable solutions to the equations of general relativity, including a variety analyzed in a 1989 paper by Matt Visser, in which a path through the wormhole can be made where the traversing path does not pass through a region of exotic matter. However, in the pure Gauss–Bonnet gravity (a modification to general relativity involving extra spatial dimensions which is sometimes studied in the context of brane cosmology) exotic matter is not needed in order for wormholes to exist—they can exist even with no matter. A type held open by negative mass cosmic strings was put forth by Visser in collaboration with Cramer *et al.*, in which it was proposed that such wormholes could have been naturally created in the early universe.

Wormholes connect two points in spacetime, which means that they would in principle allow travel in time, as well as in space. In 1988, Morris, Thorne and Yurtsever worked out how to convert a wormhole traversing space into one traversing time by accelerating one of its two mouths. However, according to general relativity, it would not be possible to use a wormhole to travel back to a time earlier than when the wormhole was first converted into a time "machine". Until this time it could not have been noticed or have been used.

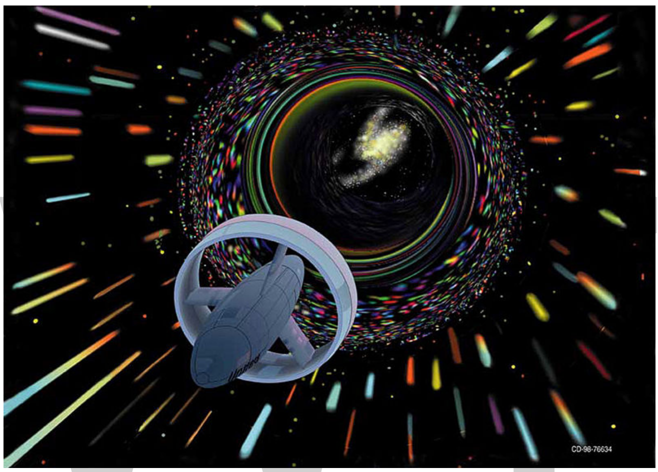
Raychaudhuri's Theorem and Exotic Matter

To see why exotic matter is required, consider an incoming light front traveling along geodesics, which then crosses the wormhole and re-expands on the other side. The expansion goes from negative to positive. As the wormhole neck is of finite size, we would not expect caustics to develop, at least within the vicinity of the neck. According to the optical Raychaudhuri's theorem, this requires a violation of the averaged null energy condition. Quantum effects such as the Casimir effect cannot violate the averaged null energy condition in any neighborhood of space with zero curvature, but calculations in semiclassical gravity suggest that quantum effects may be able to violate this condition in curved spacetime. Although it was hoped recently that quantum effects could not violate an achronal version of the averaged null energy condition, violations have nevertheless been found, so it remains an open possibility that quantum effects might be used to support a wormhole.

Modified General Relativity

In some hypotheses where general relativity is modified, it is possible to have a wormhole that does not collapse without having to resort to exotic matter. For example, this is possible with R^2 gravity, a form of $f(R)$ gravity.

Faster-than-light Travel



Wormhole travel as envisioned by Les Bossinas for NASA.

The impossibility of faster-than-light relative speed only applies locally. Wormholes might allow effective superluminal (faster-than-light) travel by ensuring that the speed of light is not exceeded locally at any time. While traveling through a wormhole, subluminal (slower-than-light) speeds are used. If two points are connected by a wormhole whose length is shorter than the distance between them outside the wormhole, the time taken to traverse it could be less than the time it would take a light beam to make the journey if it took a path through the space outside the wormhole. However, a light beam traveling through the same wormhole would beat the traveler.

Time Travel

If traversable wormholes exist, they could allow time travel. A proposed time-travel machine using a traversable wormhole would hypothetically work in the following way: One end of the wormhole is accelerated to some significant fraction of the speed of light, perhaps with some advanced propulsion system, and then brought back to the point of origin. Alternatively, another way is to take one entrance of the wormhole and move it to within the gravitational field of an object that has higher gravity than the other entrance, and then return it to a position near the other entrance. For both these methods, time dilation causes the end of the wormhole that has been moved to have aged less, or become "younger", than the stationary end as seen by an external observer; however, time connects differently through the wormhole than outside it, so that synchronized clocks at either end of the wormhole will always remain synchronized as seen by an observer passing through the wormhole, no matter how the two ends move around. This means that an observer entering the "younger" end would exit the "older" end at a time when it was the same age as the "younger" end, effectively going back in time as seen by an observer from the outside. One significant limitation of such a time machine is that it is only possible to go as far back in time as the initial creation of the machine; It is more of a path through time rather than it is a device that itself moves through time, and it would not allow the technology itself to be moved backward in time.

According to current theories on the nature of wormholes, construction of a traversable wormhole would require the existence of a substance with negative energy, often referred to as "exotic matter". More technically, the wormhole spacetime requires a distribution of energy that violates various energy conditions, such as the null energy condition along with the weak, strong, and dominant energy conditions. However, it is known that quantum effects can lead to small measurable violations of the null energy condition, and many physicists believe that the required negative energy may actually be possible due to the Casimir effect in quantum physics. Although early calculations suggested a very large amount of negative energy would be required, later calculations showed that the amount of negative energy can be made arbitrarily small.

In 1993, Matt Visser argued that the two mouths of a wormhole with such an induced clock difference could not be brought together without inducing quantum field and gravitational effects that would either make the wormhole collapse or the two mouths repel each other, or otherwise prevent information from passing through the wormhole. Because of this, the two mouths could not be brought close enough for causality violation to take place. However, in a 1997 paper, Visser hypothesized that a complex "Roman ring" (named after Tom Roman) configuration of an N number of wormholes arranged in a symmetric polygon could still act as a time machine, although he concludes that this is more likely a flaw in classical quantum gravity theory rather than proof that causality violation is possible.

Interuniversal Travel

A possible resolution to the paradoxes resulting from wormhole-enabled time travel rests on the many-worlds interpretation of quantum mechanics.

In 1991 David Deutsch showed that quantum theory is fully consistent (in the sense that the so-called density matrix can be made free of discontinuities) in spacetimes with closed timelike curves. However, later it was shown that such model of closed timelike curve can have internal inconsistencies as it will lead to strange phenomena like distinguishing non-orthogonal quantum states and distinguishing proper and improper mixture. Accordingly, the destructive positive feedback loop of virtual particles circulating through a wormhole time machine, a result indicated by semi-classical calculations, is averted. A particle returning from the future does not return to its universe of origination but to a parallel universe. This suggests that a wormhole time machine with an exceedingly short time jump is a theoretical bridge between contemporaneous parallel universes.

Because a wormhole time-machine introduces a type of nonlinearity into quantum theory, this sort of communication between parallel universes is consistent with Joseph Polchinski's proposal of an Everett phone (named after Hugh Everett) in Steven Weinberg's formulation of nonlinear quantum mechanics.

The possibility of communication between parallel universes has been dubbed inter-universal travel.

Metrics

Theories of *wormhole metrics* describe the spacetime geometry of a wormhole and serve as theoretical models for time travel. An example of a (traversable) wormhole metric is the following:

$$ds^2 = -c^2\,dt^2 + d\ell^2 + (k^2 + \ell^2)(d\theta^2 + \sin^2\theta\,d\varphi^2),$$

first presented by Ellis as a special case of the Ellis drainhole.

One type of non-traversable wormhole metric is the Schwarzschild solution:

$$ds^2 = -c^2\left(1 - \frac{2GM}{rc^2}\right)dt^2 + \frac{dr^2}{1 - \frac{2GM}{rc^2}} + r^2(d\theta^2 + \sin^2\theta\,d\varphi^2)$$

For the Schwarzschild spherically symmetric static solution:

$$ds^2 = -\frac{1}{1 - \frac{2m}{r}}dr^2 - r^2(d\theta^2 + \sin^2\theta\,d\varphi^2) + \left(1 - \frac{2m}{r}\right)dt^2,$$

where ds is the proper time and c = 1.

If one replaces r with u according to $u^2 = r - 2m$

$$ds^2 = -4(u^2 + 2m)\,du^2 - (u^2 + 2m)^2(d\theta^2 + \sin^2\theta\,d\varphi^2) + \frac{u^2}{u^2 + 2m}dt^2$$

The four-dimensional space is described mathematically by two congruent parts or "sheets", corresponding to $u > 0$ and $u < 0$, which are joined by a hyperplane $r = 2m$ or $u = 0$ in which vanishes. We call such a connection between the two sheets a "bridge".

For the combined field, gravity and electricity, Einstein and Rosen derived the following Schwarzschild static spherically symmetric solution:

$$\varphi_1 = \varphi_2 = \varphi_3 = 0, \varphi_4 = \frac{\varepsilon}{4},$$

$$ds^2 = -\frac{1}{\left(1 - \frac{2m}{r} - \frac{\varepsilon^2}{2r^2}\right)} dr^2 - r^2 (d\theta^2 + \sin^2\theta \, d\varphi^2) + \left(1 - \frac{2m}{r} - \frac{\varepsilon^2}{2r^2}\right) dt^2,$$

where ε is the electric charge.

The field equations without denominators in the case when $m = 0$ can be written:

$$\varphi_{\mu\nu} = \varphi_{\mu,\nu} - \varphi_{\nu,\mu}$$
$$g^2 \varphi_{\mu\nu;\sigma} g^{\nu\sigma} = 0$$
$$g^2 (R_{ik} + \varphi_{i\alpha} \varphi_k^{\alpha} - \frac{1}{4} g_{ik} \varphi_{\alpha\beta} \varphi^{ab}) = 0$$

In order to eliminate singularities, if one replaces r by u according to the equation:

$$u^2 = r^2 - \frac{\varepsilon^2}{2}$$

and with m = 0 one obtains:

$$\varphi_1 = \varphi_2 = \varphi_3 = 0 \text{ and } \varphi_4 = \frac{\varepsilon}{\left(u^2 + \frac{\varepsilon^2}{2}\right)^{1/2}}$$

$$ds^2 = -du^2 - \left(u^2 + \frac{\varepsilon^2}{2}\right)(d\theta^2 + \sin^2\theta \, d\varphi^2) + \left(\frac{2u^2}{2u^2 + \varepsilon^2}\right) dt^2$$

The solution is free from singularities for all finite points in the space of the two sheets.

The Unified Theory of Formation of Wormholes

Although wormholes have not been observed but mathematical interpretation do suggest existence of wormholes. One reason for which wormholes haven't been detected, is that the conditions required for formation aren't achieved. Though the thought of formation a wormhole from two black holes is a radical thought but the mathematical interpretation does show chances for formation in that way.

Taking in consideration that a wormhole is created from two black hole. The procedure of formation might be merger of the two singularities of the respective black holes.

When the two singularities of the black hole merges to form a region of higher dimension, thereby forming a wormhole. The merger of the two singularities thereby forming a region of higher dimension is called singularity convergence. Singularity convergence occurs either when two black holes bearing opposite charges collide. As known that two black holes collide and merge to form a bigger black hole.

Now taking in consideration that the two black holes are having opposite charges hence as they come closer the two singularities get attracted and merges together even before the event horizon has merged. As the singularity has merged before the event horizon's merger. Hence the singularity merges to form higher dimension and the event horizon's work as a gateway from one point to another.

Equations

$$\delta M = \kappa\delta A / 8\pi G + \Omega\delta J + \phi\delta Q$$

For positively charged black hole:

$$\delta M_\alpha = \kappa\delta A_\alpha / 8\pi G + \Omega_\alpha \delta J_\alpha + \phi_\alpha \delta Q_\alpha$$

For negatively charged black hole:

$$\delta M_\alpha = \kappa\delta A_\beta / 8\pi G + \Omega_\beta \delta J_\beta + \phi_\beta \delta Q_\beta$$

Charge on positively charged black hole:

$$\delta Q_\alpha = \left(\delta M_\alpha - \kappa\delta A_\alpha / 8\pi G - \Omega_\alpha \delta J_\alpha\right) / \phi_\alpha$$

Charge on negatively charged black hole:

$$\delta Q_\beta = \left(\delta M_\beta - \kappa\delta A_\beta / 8\pi G - \Omega_\beta \delta J_\beta\right) / \phi_\beta$$

Force of attraction:

$$\delta F = 1 / 4\pi\varepsilon_0 * \delta Q_\beta * \delta Q_\alpha / r^2$$

Again from black hole mechanics,

$$\delta E / c^2 = \kappa \delta A / 8\pi G + \Omega \delta J + \phi \delta Q$$
$$\delta E = c^2 \kappa \delta A / 8\pi G + c^2 \Omega \delta J + c^2 \phi \delta Q$$
$$-c^2 \kappa \delta A / 8\pi G = c^2 \Omega \delta J + c^2 \phi \delta Q - \delta E$$
$$-\kappa \delta A / 8\pi G = \Omega \delta J + \phi \delta Q - \delta E / c^2$$
$$\kappa \delta A / 8\pi G = \delta M - \Omega \delta J - \phi \delta Q$$
$$\kappa \delta A = 8\pi G \left(\delta M - \Omega \delta J - \phi \delta Q \right)$$
$$\delta A = 8\pi G \left(\delta M - \Omega \delta J - \phi \delta Q \right) / \kappa$$

From the Hawking Radiation:

$$S_{bh} = \kappa_b A / 41p^2$$
$$\delta A = \delta S_{bh} * 4lp^2 / \kappa_b$$
$$\delta S_{bh} * 4lp^2 / \kappa_b = 8\pi G \left(\delta M - \Omega \delta J - \phi \delta Q \right) / \kappa$$
$$\delta S_{bh} * 4lp^2 = \kappa_b = \left\{ 8\pi G \left(\delta M - \Omega \delta J - \phi \delta Q \right) / \kappa \right\}$$
$$\delta S_{bh} = \kappa_b \left\{ 8\pi G \left(\delta M - \Omega \delta J - \phi \delta Q \right) / \kappa \right\} / 4lp^2$$

Entropy of positively charged black hole:

$$\delta S_{bh\alpha} = \kappa_b \left\{ 8\pi G \left(\delta M_\alpha - \Omega_\alpha \delta J_\alpha - \phi_\alpha \delta Q_\alpha \right) / \kappa \right\} / 4lp^2$$

Entropy of negatively charged black hole:

$$\delta S_{bh\beta} = \kappa_b \left\{ 8\pi G \left(\delta M_\beta - \Omega_\beta \delta J_\beta - \phi_\beta \delta Q_\beta \right) / \kappa \right\} / 4lp^2$$

Inserting General Relativity:

$$G_{\mu\nu} = R_{\mu\nu} - 1/2 R g_{\mu\nu} = 8\pi G / c^4 \;.\; T_{\mu\nu}$$
$$R_{\mu\nu} - 1/2 R g_{\mu\nu} = 8\pi G / c^4 \;.\; T_{\mu\nu}$$
$$c^4 \left(R_{\mu\nu} - 1/2 \; R g_{\mu\nu} \right) = 8\pi G \;.\; T_{\mu\nu}$$
$$c^4 \left(R_{\mu\nu} - 1/2 \; R g_{\mu\nu} \right) / T_{\mu\nu} = 8\pi G$$
$$8\pi G = c^4 \left(R_{\mu\nu} - 1/2 \, R g_{\mu\nu} \right) / T_{\mu\nu}$$

Entropy of positively charged black hole:

$$\delta S_{bh\alpha} = \kappa_b \{ c^4 \left(R_{\mu\nu} - 1/2 R g_{\mu\nu} \right) / T_{\mu\nu}$$
$$\left(\delta M_\alpha - \Omega_\alpha \delta J_\alpha - \phi_\alpha \delta Q_\alpha \right) / \kappa \} 4lp^2$$

Entropy of negatively charged black hole:

$$\delta S_{bh\beta} = \kappa_b \{c^4 \left(R_{\mu\nu} - 1/2Rg_{\mu\nu}\right)/T_{\mu\nu}$$
$$\left(\delta M_{\beta} - \Omega_{\beta}\delta J_{\beta} - \phi_{\beta}\delta Q_{\beta}\right)/\kappa\}4lp^2$$

Curvature of Space and time in positively charged black hole:

$$RG_{\mu\nu\alpha} = 2\{-\delta S_{bh\alpha} * 4lp^2 / \kappa_b \left(-\kappa\right)$$
$$\left(-\delta M_{\alpha} + \Omega_{\alpha}\delta J_{\alpha} + \phi_{\alpha}\delta Q_{\alpha}\right)^{-1}\left(-T_{\mu\nu\alpha}\right)\left(-c^4\right) + R_{\mu\nu\alpha}$$

Curvature of Space and time in negatively charged black hole:

$$Rg_{\mu\nu\beta} = 2\{-\delta S_{bh\beta} * 4lp^2 / \kappa_b \left(-\kappa\right)$$
$$\left(-\delta M_{\beta} + \Omega_{\beta}\delta J_{\beta} + \phi_{\beta}\delta Q_{\beta}\right)^{-1}\left(-T_{\mu\nu\beta}\right)\left(-c^4\right) + R_{\mu\nu\beta}$$

References

- Black-hole, science: britannica.com, Retrieved 09 May, 2019
- How-big-is-a-black-hole, astronomy-resources: skyandtelescope.com, Retrieved 18 June, 2019
- Topics-blackholes-blackholes: physicsoftheuniverse.com, Retrieved 27 August, 2019
- Supermassive-Black-Hole, cosmos: astronomy.swin.edu.au, Retrieved 26 April, 2019
- Frampton, Paul H.; Kawasaki, Masahiro; Takahashi, Fuminobu; Yanagida, Tsutomu T. (22 April 2010). “Primordial Black Holes as All Dark Matter”. Journal of Cosmology and Astroparticle Physics. 2010 (04): 023–023. arXiv:1001.2308. doi:10.1088/1475-7516/2010/04/023. ISSN 1475-7516. Retrieved 25 September 2019
- Topics-blackholes-wormholes: physicsoftheuniverse.com, Retrieved 02 March, 2019
- DeBenedictis, Andrew & Das, A. (2001). “On a General Class of Wormhole Geometries”. Classical and Quantum Gravity. 18 (7): 1187–1204. arXiv:gr-qc/0009072. Bibcode:2001CQGra..18.1187D. CiteSeerX 10.1.1.339.8662. doi:10.1088/0264-9381/18/7/304
- Stellar-Black-Hole, cosmos: astronomy.swin.edu.au, Retrieved 14 January, 2019
- Gebhardt, Karl; Rich, R. M.; Ho, Luis C. (December 2005), “An Intermediate-Mass Black Hole in the Globular Cluster G1: Improved Significance from New Keck and Hubble Space Telescope Observations”, The Astrophysical Journal, 634 (2): 1093–1102, arXiv:astro-ph/0508251, Bibcode:2005ApJ...634.1093G, doi:10.1086/497023

Chapter 5

The Galaxy

A gravitationally bound system of stars, stellar remnants, interstellar gas, dust, and dark matter is termed as a galaxy. Some commonly studied galaxies are Milky Way galaxy, Andromeda galaxy, Magellanic cloud, Quasar, etc. This chapter closely examines these types of galaxies to provide an extensive understanding of the subject.

Galaxy is the system of stars and interstellar matter that make up the universe. Many such assemblages are so enormous that they contain hundreds of billions of stars.

Whirlpool Galaxy (M51); NGC 5195. The Whirlpool Galaxy (left), also known as M51, Sc galaxy accompanied by a small irregular companion galaxy, NGC 5195 (right).

Nature has provided an immensely varied array of galaxies, ranging from faint, diffuse dwarf objects to brilliant spiral-shaped giants. Virtually all galaxies appear to have been formed soon after the universe began, and they pervade space, even into the depths of the farthest reaches penetrated by powerful modern telescopes. Galaxies usually exist in clusters, some of which in turn are grouped into larger clusters that measure hundreds of millions of light-years across. (A light-year is the distance traversed by light in one year, traveling at a velocity of 300,000 km per second [km/sec], or 650,000,000 miles per hour.) These so-called superclusters are separated by nearly empty voids, and this causes the gross structure of the universe to look somewhat like a network of sheets and chains of galaxies.

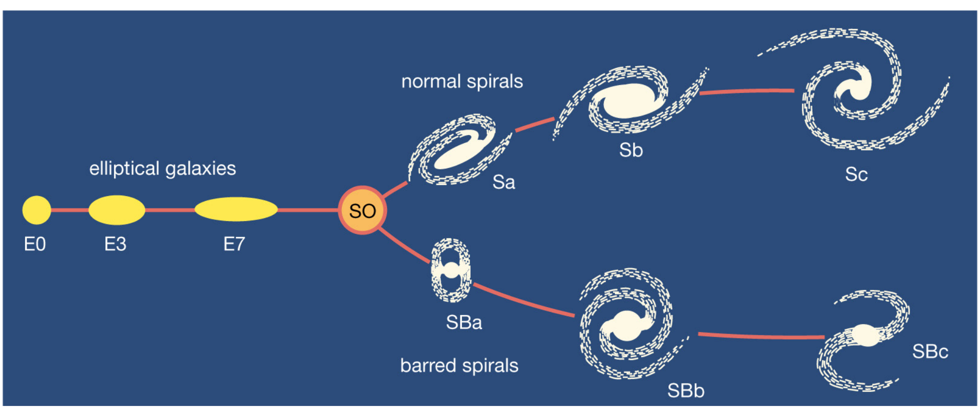
Galaxies differ from one another in shape, with variations resulting from the way in which the systems were formed and subsequently evolved. Galaxies are extremely varied not only in structure but also in the amount of activity observed. Some are the sites of vigorous star formation, with its attendant glowing gas and clouds of dust and molecular complexes. Others, by contrast, are quiescent, having long ago ceased to form new stars. Perhaps the most conspicuous activity in galaxies occurs in their nuclei, where evidence suggests that in many cases supermassive objects—probably black holes—lurk. These central black holes apparently formed several billion years ago; they are now observed forming in galaxies at large distances (and, therefore, because of the time it takes light to travel to Earth, at times in the far distant past) as brilliant objects called quasars.

The existence of galaxies was not recognized until the early 20th century. Since then, however, galaxies have become one of the focal points of astronomical investigation. The notable developments and achievements in the study of galaxies are surveyed here. Included in the discussion are the external galaxies (i.e., those lying outside the Milky Way Galaxy, the local galaxy to which the Sun and Earth belong), their distribution in clusters and superclusters, and the evolution of galaxies and quasars.

Types of Galaxies

Principal Schemes of Classification

Almost all current systems of galaxy classification are outgrowths of the initial scheme proposed by the American astronomer Edwin Hubble in 1926. In Hubble's scheme, which is based on the optical appearance of galaxy images on photographic plates, galaxies are divided into three general classes: ellipticals, spirals, and irregulars. Hubble subdivided these three classes into finer groups.



Hubble's system of classification for galaxies.

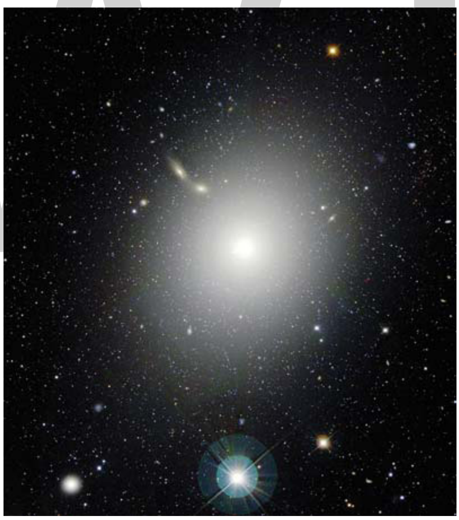
In The Hubble Atlas of Galaxies, the American astronomer Allan R. Sandage drew on Hubble's notes and his own research on galaxy morphology to revise the Hubble classification scheme. Some of the features of this revised scheme are subject to argument because of the findings of very recent research, but its general features, especially the coding of types, remain viable.

Elliptical Galaxies

These systems exhibit certain characteristic properties. They have complete rotational symmetry; i.e., they are figures of revolution with two equal principal axes. They have a third smaller axis that is the presumed axis of rotation. The surface brightness of ellipticals at optical wavelengths decreases monotonically outward from a maximum value at the centre, following a common mathematical law of the form:

$$I = I_0 \left(r / a + 1 \right)^{-2},$$

where I is the intensity of the light, I_o is the central intensity, r is the radius, and a is a scale factor. The isophotal contours exhibited by an elliptical system are similar ellipses with a common orientation, each centred on its nucleus. No galaxy of this type is flatter than b/a = 0.3, with b and a the minor and major axes of the elliptical image, respectively. Ellipticals contain neither interstellar dust nor bright stars of spectral types O and B. Many, however, contain evidence of the presence of low-density gas in their nuclear regions. Ellipticals are red in colour, and their spectra indicate that their light comes mostly from old stars, especially evolved red giants.



The giant elliptical galaxy M87, also known as Virgo A. M87 appears near the centre of the Virgo Cluster of galaxies.

Subclasses of elliptical galaxies are defined by their apparent shape, which is of course not necessarily their three-dimensional shape. The designation is En, where n is an integer defined by,

$$n = 10\left(a - b \right) / a.$$

A perfectly circular image will be an E0 galaxy, while a flatter object might be an E7 galaxy. (Elliptical galaxies are never flatter than this, so there are no E8, E9, or E10 galaxies).

Although the above-cited criteria are generally accepted, current high-quality measurements have shown that some significant deviations exist. Most elliptical galaxies do not, for instance, exactly fit the intensity law formulated by Hubble; deviations are evident in their innermost parts and in their faint outer parts. Furthermore, many elliptical galaxies have slowly varying ellipticity, with the images being more circular in the central regions than in the outer parts. The major axes sometimes do not line up either; their position angles vary in the outer parts. Finally, astronomers have found that a few ellipticals do in fact have small numbers of luminous O and B stars as well as dust lanes.

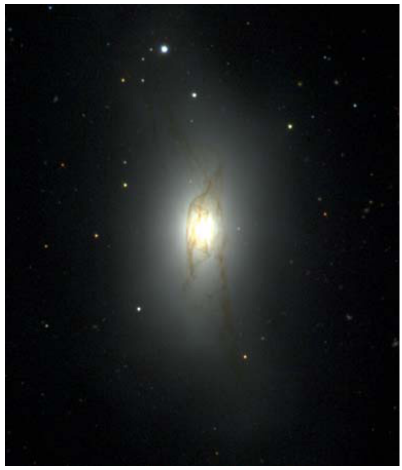
Spiral Galaxies

Spirals are characterized by circular symmetry, a bright nucleus surrounded by a thin outer disk, and a superimposed spiral structure. They are divided into two parallel classes: normal spirals and barred spirals. The normal spirals have arms that emanate from the nucleus, while barred spirals have a bright linear feature called a bar that straddles the nucleus, with the arms unwinding from the ends of the bar. The nucleus of a spiral galaxy is a sharp-peaked area of smooth texture, which can be quite small or, in some cases, can make up the bulk of the galaxy. Both the arms and the disk of a spiral system are blue in colour, whereas its central areas are red like an elliptical galaxy. The normal spirals are designated S and the barred varieties SB. Each of these classes is subclassified into three types according to the size of the nucleus and the degree to which the spiral arms are coiled. The three types are denoted with the lowercase letters a, b, and c. There also exist galaxies that are intermediate between ellipticals and spirals. Such systems have the disk shape characteristic of the latter but no spiral arms. These intermediate forms bear the designation S0.

The Pinwheel Galaxy (M101), as seen in an optical image.

S0 Galaxies

These systems exhibit some of the properties of both the ellipticals and the spirals and seem to be a bridge between these two more common galaxy types. Hubble introduced the S0 class long after his original classification scheme had been universally adopted, largely because he noticed the dearth of highly flattened objects that otherwise had the properties of elliptical galaxies.



S0 galaxy NGC 4753 in an optical image taken by the Sloan Digital Sky Survey Telescope at the Apache Point Observatory in Sunspot, N.M. NGC 4753 is a member of the Virgo Cluster of galaxies.

S0 galaxies have a bright nucleus that is surrounded by a smooth, featureless bulge and a faint outer envelope. They are thin; statistical studies of the ratio of the apparent axes (seen projected onto the sky) indicate that they have intrinsic ratios of minor to major axes in the range 0.1 to 0.3. Their structure does not generally follow the luminosity law of elliptical galaxies but has a form more like that for spiral galaxies. Some S0 systems have a hint of structure in the envelope, either faintly discernible armlike discontinuities or narrow absorption lanes produced by interstellar dust. Several S0 galaxies are otherwise peculiar, and it is difficult to classify them with certainty. They can be thought of as peculiar irregular galaxies or simply as some of the 1 or 2 percent of galaxies that do not fit easily into the Hubble scheme. Among these are such galaxies as NGC 4753, which has irregular dust lanes across its image, and NGC 128, which has a double, almost rectangular bulge around a central nucleus. Another type of peculiar S0 is found in NGC 2685. This nebula in the constellation Ursa Major has an apparently edge-on disk galaxy at its centre, with surrounding hoops of gas, dust, and stars arranged in a plane that is at right angles to the apparent plane of the central object.

Sa Galaxies

These normal spirals have narrow, tightly wound arms, which usually are visible because of the presence of interstellar dust and, in many cases, bright stars. Most of them have a large amorphous bulge in the centre, but there are some that violate this criterion, having a small nucleus around which is arranged an amorphous disk with superimposed faint arms. NGC 1302 is an example of the normal type of Sa galaxy, while NGC 4866 is representative of one with a small nucleus and arms consisting of thin dust lanes on a smooth disk.

The Sombrero Galaxy (M104), which is classified as an Sa/Sb galaxy, in an optical image taken by the Hubble Space Telescope.

Sb Galaxies

This intermediate type of spiral typically has a medium-sized nucleus. Its arms are more widely spread than those of the Sa variety and appear less smooth. They contain stars, star clouds, and interstellar gas and dust. Sb galaxies show wide dispersions in details in terms of their shape. Hubble and Sandage observed, for example, that in certain Sb galaxies the arms emerge at the nucleus, which is often quite small. Other members of this subclass have arms that begin tangent to a bright, nearly circular ring, while still others reveal a small, bright spiral pattern inset into the nuclear bulge. In any of these cases, the spiral arms may be set at different pitch angles. (A pitch angle is defined as the angle between an arm and a circle centred on the nucleus and intersecting the arm).

Sb galaxy NGC 891 as seen edge-on in an optical image taken with the WIYN Telescope. The dark band is made of interstellar dust.

Hubble and Sandage noted further deviations from the standard shape established for Sb galaxies. A few systems exhibit a chaotic dust pattern superimposed upon the tightly wound spiral arms. Some have smooth, thick arms of low surface brightness, frequently bounded on their inner edges with dust lanes. Finally, there are those with a large, smooth nuclear bulge from which the arms emanate, flowing outward tangent to the bulge and forming short arm segments. This is the most familiar type of Sb galaxy and is best exemplified by the giant Andromeda Galaxy.

Many of these variations in shape remain unexplained. Theoretical models of spiral galaxies based on a number of different premises can reproduce the basic Sb galaxy shape, but many of the deviations noted above are somewhat mysterious in origin and must await more detailed and realistic modeling of galactic dynamics.

Sc Galaxies

These galaxies characteristically have a very small nucleus and multiple spiral arms that are open, with relatively large pitch angles. The arms, moreover, are lumpy, containing as they do numerous irregularly distributed star clouds, stellar associations, star clusters, and gas clouds known as emission nebulae.

As in the case of Sb galaxies, there are several recognizable subtypes among the Sc systems. Sandage has cited six subdivisions: (1) galaxies, such as the Whirlpool Galaxy (M51), that have thin branched arms that wind outward from a tiny nucleus, usually extending out about 180° before branching into multiple segments, (2) systems with multiple arms that start tangent to a bright ring centred on the nucleus, (3) those with arms that are poorly defined and that span the entire image of the galaxy, (4) those with a spiral pattern that cannot easily be traced and that are multiple and punctuated with chaotic dust lanes, (5) those with thick, loose arms that are not well defined—e.g., the nearby galaxy M33 (the Triangulum Nebula)—and (6) transition types, which are almost so lacking in order that they could be considered irregular galaxies.

Some classification schemes, such as that of the French-born American astronomer Gerard de Vaucouleurs, give the last of the above-cited subtypes a class of its own, type Sd. It also has been found that some of the variations noted here for Sc galaxies are related to total luminosity. Galaxies of the fifth subtype, in particular, tend to be intrinsically faint, while those of the first subtype are among the most luminous spirals known. This correlation is part of the justification for the luminosity classification.

SB Galaxies

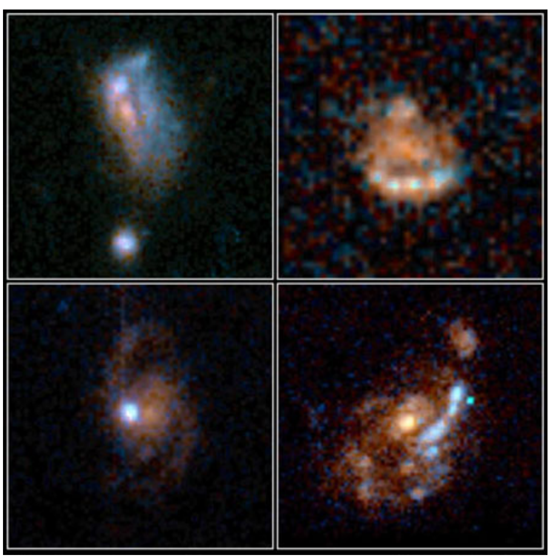
The luminosities, dimensions, spectra, and distributions of the barred spirals tend to be indistinguishable from those of normal spirals. The subclasses of SB systems exist in parallel sequence to those of the latter.

There are SB0 galaxies that feature a large nuclear bulge surrounded by a disklike envelope across which runs a luminous featureless bar. Some SB0 systems have short bars, while others have bars that extend across the entire visible image. Occasionally there is a ringlike feature external to the bar. SBa galaxies have bright, fairly large nuclear bulges and tightly wound, smooth spiral arms that emerge from the ends of the bar or from a circular ring external to the bar. SBb systems have a smooth bar as well as relatively smooth and continuous arms. In some galaxies of this type, the arms start at or near the ends of the bar, with conspicuous dust lanes along the inside of the bar that can be traced right up to the nucleus. Others have arms that start tangent to a ring external to the bar. In SBc galaxies, both the arms and the bar are highly resolved into star clouds and stellar associations. The arms are open in form and can start either at the ends of the bar or tangent to a ring.

Barred spiral galaxy NGC 1300.

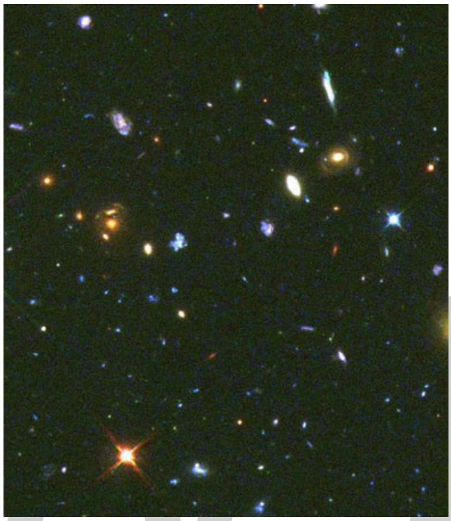
Irregular Galaxies

Most representatives of this class consist of grainy, highly irregular assemblages of luminous areas. They have neither noticeable symmetry nor an obvious central nucleus, and they are generally bluer in colour than are the arms and disks of spiral galaxies. An extremely small number of them, however, are red and have a smooth, though nonsymmetrical, shape.



Four irregular galaxies, as observed by the Hubble Space Telescope.

Hubble recognized these two types of irregular galaxies, Irr I and Irr II. The Irr I type is the most common of the irregular systems, and it seems to fall naturally on an extension of the spiral classes, beyond Sc, into galaxies with no discernible spiral structure. They are blue, are highly resolved, and have little or no nucleus. The Irr II systems are red, rare objects. They include various kinds of chaotic galaxies for which there apparently are many different explanations, including most commonly the results of galaxy-galaxy interactions, both tidal distortions and cannibalism; therefore, this category is no longer seen as a useful way to classify galaxies.



Various galactic types: Most interesting are the numerous blue dwarf galaxies.

Some irregular galaxies, like spirals, are barred. They have a nearly central bar structure dominating an otherwise chaotic arrangement of material. The Large Magellanic Cloud is a well-known example.

Other Classification Schemes and Galaxy Types

Other classification schemes similar to Hubble's follow his pattern but subdivide the galaxies differently. A notable example of one such system is that of de Vaucouleurs. This scheme, which has evolved considerably since its inception in 1959, includes a large number of codes for indicating different kinds of morphological characteristics visible in the images of galaxies. The major Hubble galaxy classes form the framework of de Vaucouleurs's scheme, and its subdivision includes different families, varieties, and stages. The de Vaucouleurs system is so detailed that it is more of a descriptive code for galaxies than a commonly used classification scheme.

Galaxies with unusual properties often have shorthand names that refer to their characteristic properties. Common examples are:

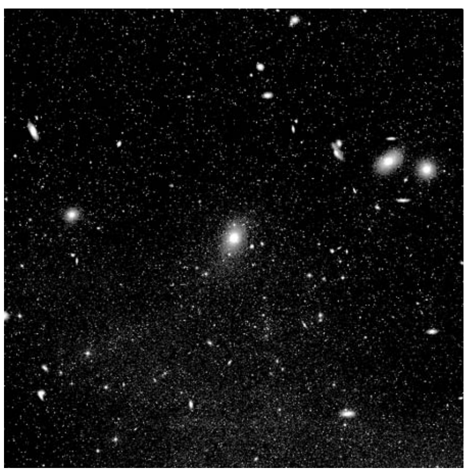
- cD: Galaxies with abnormally large, distended shapes, always found in the central areas of galaxy clusters and hypothesized to consist of merged galaxies.

- S: Seyfert galaxies, originally recognized by the American astronomer Carl K. Seyfert from optical spectra. These objects have very bright nuclei with strong emission lines of hydrogen and other common elements, showing velocities of hundreds or thousands of kilometres per second. Most are radio sources.
- N: Galaxies with small, very bright nuclei and strong radio emission. These are probably similar to Seyfert galaxies but more distant.
- Q: Quasars, or QSOs, small, extremely luminous objects, many of which are strong radio sources. Quasars apparently are related to Seyfert and N galaxies but have such bright nuclei that the underlying galaxy can be detected only with great difficulty.

There are also different schemes used for extremely distant galaxies, which we see in their youth. When a very distant galaxy is examined with a very large telescope, we see its structure as it was when the light was emitted billions of years ago. In such cases, the distinctive Hubble types are not so obvious. Apparently, galaxies are much less well organized in their early years, and these very distant objects tend to be highly irregular and asymmetrical. Although special classification schemes are sometimes used for special purposes, the general scheme of Hubble in its updated form is the one most commonly used.

Clusters of Galaxies

Galaxies tend to cluster together, sometimes in small groups and sometimes in enormous complexes. Most galaxies have companions, either a few nearby objects or a large-scale cluster; isolated galaxies, in other words, are quite rare.



The central portion of the Virgo Cluster in an optical. The galaxy in the centre is M87 (also known as the radio galaxy Virgo A).

Types of Clusters

There are several different classification schemes for galaxy clusters, but the simplest is the most useful. This scheme divides clusters into three classes: groups, irregulars, and sphericals.

Groups

The groups class is composed of small compact groups of 10 to 50 galaxies of mixed types, spanning roughly five million light-years. An example of such an entity is the Local Group, which includes the Milky Way Galaxy, the Magellanic Clouds, the Andromeda Galaxy, and about 50 other systems, mostly of the dwarf variety.

Hickson Compact Group 87, which contains four galaxies.

Irregular Clusters

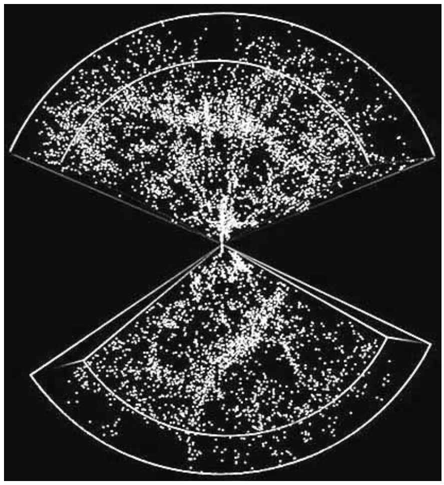
Irregular clusters are large loosely structured assemblages of mixed galaxy types (mostly spirals and ellipticals), totaling perhaps 1,000 or more systems and extending out 10,000,000 to 50,000,000 light-years. The Virgo and Hercules clusters are representative of this class.

Spherical Clusters

Spherical clusters are dense and consist almost exclusively of elliptical and S0 galaxies. They are enormous, having a linear diameter of up to 50,000,000 light-years. Spherical clusters may contain as many as 10,000 galaxies, which are concentrated toward the cluster centre.

Distribution

Clusters of galaxies are found all over the sky. They are difficult to detect along the Milky Way, where high concentrations of the Galaxy's dust and gas obscure virtually everything at optical wavelengths. However, even there clusters can be found in a few galactic "windows," random holes in the dust that permit optical observations.



Distribution of 9,325 galaxies (points) showing large-scale structures like walls where galaxies congregate and voids where galaxies are absent.

The clusters are not evenly spaced in the sky; instead, they are arranged in a way that suggests a certain amount of organization. Clusters are frequently associated with other clusters, forming giant superclusters. These superclusters typically consist of 3 to 10 clusters and span as many as 200 million light-years. There also are immense areas between clusters that are fairly empty, forming voids. Large-scale surveys made in the 1980s of the radial velocities of galaxies revealed an even-larger kind of structure. It was discovered that galaxies and galaxy clusters tend to fall in position along large planes and curves, almost like giant walls, with relatively empty spaces between them. A related large-scale structure was found to exist where there occur departures from the velocity-distance relation in certain directions, indicating that the otherwise uniform expansion is being perturbed by large concentrations of mass. One of these, discovered in 1988, has been dubbed “the Great Attractor.”

Interactions between Cluster Members

Antennae galaxies: The Antennae galaxies colliding.

Galaxies in clusters exist in a part of the universe that is much denser than average, and the result is that they have several unusual features. In the inner parts of dense clusters there are very few, if any, normal spiral galaxies. This condition is probably the result of fairly frequent collisions between the closely packed galaxies, as such violent interactions tend to sweep out the interstellar gas, leaving behind only the spherical component and a gasless disk. What reains is in effect an S0 galaxy.

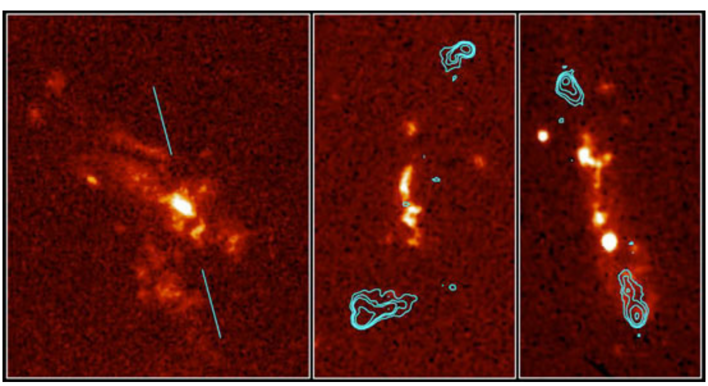
A second and related effect of galaxy interactions is the presence of gas-poor spiral systems at the centres of large irregular clusters. A significant number of the members of such clusters have anomalously small amounts of neutral hydrogen, and their gas components are smaller on average than those for more isolated galaxies. This is thought to be the result of frequent distant encounters between such galaxies involving the disruption of their outer parts.

A third effect of the dense cluster environment is the presence in some clusters—usually rather small dense clusters—of an unusual type of galaxy called a cD galaxy. These objects are somewhat similar in structure to S0 galaxies, but they are considerably larger, having envelopes that extend out to radii as large as one million light-years. Many of them have multiple nuclei, and most are strong sources of radio waves. The most likely explanation for cD galaxies is that they are massive central galactic systems that have captured smaller cluster members because of their dominating gravitational fields and have absorbed the other galaxies into their own structures. Astronomers sometimes refer to this process as galactic cannibalism. In this sense, the outer extended disks of cD systems, as well as their multiple nuclei, represent the remains of past partly digested “meals.”

One more effect that can be traced to the cluster environment is the presence of strong radio and X-ray sources, which tend to occur in or near the centres of clusters of galaxies.

Extragalactic Radio and X-ray Sources

Radio Galaxies



Some of the strongest radio sources in the sky are galaxies. Most of them have a peculiar morphology that is related to the cause of their radio radiation. Some are relatively isolated galaxies, but most galaxies that emit unusually large amounts of radio energy are found in large clusters.

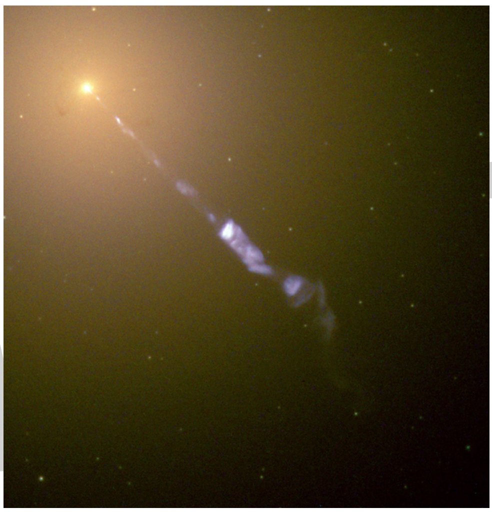
Three radio galaxies: These images of dwarf galaxy 3C 265 (left), 3C 324 (centre), and 3C 368 (right), a galaxy whose main radio emissions are probably caused by a gas jet along one axis, combine observations made by the Hubble Space Telescope with radio source maps (blue contour lines) made by the Very Large Array Radio Interferometer.

The basic characteristics of radio galaxies and the variations that exist among them can be made clear with two examples. The first is Centaurus A, a giant radio structure surrounding a bright, peculiar galaxy of remarkable morphology designated NGC 5128. It exemplifies a type of radio galaxy that consists of an optical galaxy located at the centre of an immensely larger two-lobed radio source. In the particular case of Centaurus A, the extent of the radio structure is so great that it is almost 100 times the size of the central galaxy, which is itself a giant galaxy. This radio structure includes, besides the pair of far-flung radio lobes, two other sets of radio sources: One that is approximately the size of the optical galaxy and that resembles the outer structure in shape, and a second that is an intense small source at the galaxy's nucleus. Optically, NGC 5128 appears as a giant elliptical galaxy with two notable characteristics: An unusual disk of dust and gas surrounding it and thin jets of interstellar gas and young stars radiating outward. The most plausible explanation for this whole array is that a series of energetic events in the nucleus of the galaxy expelled hot ionized gas from the centre at relativistic velocities (i.e., those at nearly the speed of light) in two opposite directions. These clouds of relativistic particles generate synchrotron radiation, which is detected at radio (and X-ray) wavelengths. In this model the very large structure is associated with an old event, while the inner lobes are the result of more-recent ejections. The centre is still active, as evidenced by the presence of the nuclear radio source.

Composite image of radio galaxy Centaurus A, as seen in X-ray data (blue areas) from the Chandra X-ray Observatory, visible-light data (yellow areas) from the UK Schmidt

Telescope at the Siding Spring Observatory in New South Wales, Austl., and radio-wave data (green and pink areas) from the Very Large Array in Socorro, N.M.

The other notable example of a radio galaxy is Virgo A, a powerful radio source that corresponds to a bright elliptical galaxy in the Virgo Cluster, designated as M87. In this type of radio galaxy, most of the radio radiation is emitted from an appreciably smaller area than in the case of Centaurus A. This area coincides in size with the optically visible object. Virgo A is not particularly unusual except for one peculiarity: it has a bright jet of gaseous material that appears to emanate from the nucleus of the galaxy, extending out approximately halfway to its faint outer parts. This gaseous jet can be detected at optical, radio, and other (e.g., X-ray) wavelengths; its spectrum suggests strongly that it shines by means of the synchrotron mechanism.



The jet of radio galaxy Virgo A, as seen by the Hubble Space Telescope.

The only condition that can account for the immense amounts of energy emitted by radio galaxies is the capture of material (interstellar gas and stars) by a supermassive object at their centre. Such an object would resemble the one thought to be in the nucleus of the Milky Way Galaxy but would be far more massive. In short, the most probable type of supermassive object for explaining the details of strong radio sources would be a black hole. For example, M87 has such a black hole, with a mass 6.5 billion times that of the Sun. Large amounts of energy can be released when material is captured by a black hole. An extremely hot high-density accretion disk is first formed around the supermassive object from the material, and then some of the material seems to be ejected explosively from the area, giving rise to the various radio jets and lobes observed.

Black hole at the centre of the massive galaxy M87, about 55 million light-years from Earth, as imaged by the Event Horizon Telescope (EHT). The black hole is 6.5 billion times more massive than the Sun. This image was the first direct visual evidence of a supermassive black hole and its shadow. The ring is brighter on one side because the black hole is rotating, and thus material on the side of the black hole turning toward

Earth has its emission boosted by the Doppler effect. The shadow of the black hole is about five and a half times larger than the event horizon, the boundary marking the black hole's limits, where the escape velocity is equal to the speed of light.

Black hole in M87.

Another kind of event that can result in an explosive eruption around a nuclear black hole involves cases of merging galaxies in which the nuclei of the galaxies "collide." Because many, if not most, galaxy nuclei contain a black hole, such a collision can generate an immense amount of energy as the black holes merge.

X-ray Galaxies

Synchrotron radiation is characteristically emitted at virtually all wavelengths at almost the same intensity. A synchrotron source therefore ought to be detectable at optical and radio wavelengths, as well as at others (e.g., infrared, ultraviolet, X-ray, and gamma-ray wavelengths). For radio galaxies this does seem to be the case, at least in circumstances where the radiation is not screened by absorbing material in the source or in intervening space.

X-rays are absorbed by Earth's atmosphere. Consequently, X-ray galaxies could not be detected until it became possible to place telescopes above the atmosphere, first with balloons and sounding rockets and later with orbiting observatories specially designed for X-ray studies. For example, the Einstein Observatory, which was in operation during the early 1980s, made a fairly complete search for X-ray sources across the sky. Beginning in 1999, the Chandra X-ray Observatory and other orbiting X-ray observatories detected huge numbers of emitters. Many of the sources turned out to be distant galaxies and quasars, while others were relatively nearby objects, including neutron stars (extremely dense stars composed almost exclusively of neutrons) in the Milky Way Galaxy.

A substantial number of the X-ray galaxies so far detected are also well-known radio galaxies. Some X-ray sources, such as certain radio sources, are much too large to be individual galaxies but rather consist of a whole cluster of galaxies.

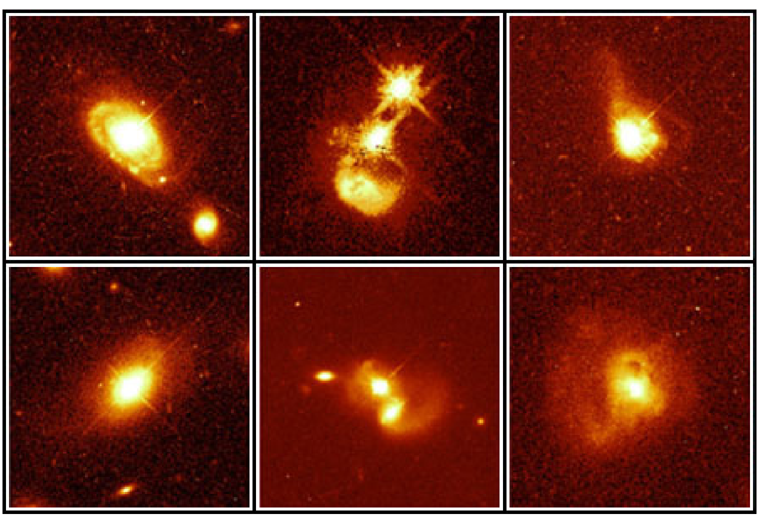
Clusters of Galaxies as Radio and X-ray Sources

Some clusters of galaxies contain a widespread intergalactic cloud of hot gas that can be detected as a diffuse radio source or as a large-scale source of X-rays. The gaseous cloud has a low density but a very high temperature, having been heated by the motion of the cluster's galaxies through it and by the emission of high-energy particles from active galaxies within it.

The form of certain radio galaxies in clusters points rather strongly to the presence of intergalactic gas. These are the "head-tail" galaxies, systems that have a bright source accompanied by a tail or tails that appear swept back by their interaction with the cooler more stationary intergalactic gas. These tails are radio lobes of ejected gas whose shape has been distorted by collisions with the cluster medium.

Quasars

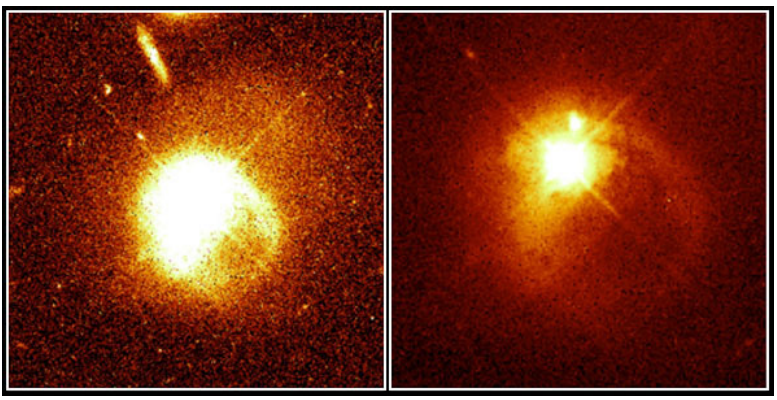
An apparently new kind of radio source was discovered in the early 1960s when radio astronomers identified a very small but powerful radio object designated 3C 48 with a stellar optical image. When they obtained the spectrum of the optical object, they found unexpected and at first unexplainable emission lines superimposed on a flat continuum. This object remained a mystery until another similar but optically brighter object, 3C 273, was examined in 1963. Investigators noticed that 3C 273 had a normal spectrum with the same emission lines as observed in radio galaxies, though greatly redshifted (i.e., the spectral lines are displaced to longer wavelengths), as by the Doppler effect. If the redshift were to be ascribed to velocity, however, it would imply an immense velocity of recession. In the case of 3C 48, the redshift had been so large as to shift familiar lines so far that they were not recognized. Many more such objects were found, and they came to be known as quasi-stellar radio sources, abbreviated as quasars.



Six quasar host galaxies, as observed by the Hubble Space Telescope.Shown are apparently normal, solitary galaxies (left), colliding galaxies (centre), and merging galaxies (right).

Although the first 20 years of quasar studies were noted more for controversy and mystery than for progress in understanding, subsequent years finally saw a solution to the questions raised by these strange objects. It is now clear that quasars are extreme examples of energetic galaxy nuclei. The amount of radiation emitted by such a nucleus overwhelms the light from the rest of the galaxy, so only very special observational techniques can reveal the galaxy's existence.

A quasar has many remarkable properties. Although it is extremely small (only the size of the solar system), it emits up to 100 times as much radiation as an entire galaxy. It is a complex mixture of very hot gas, cooler gas and dust, and particles that emit synchrotron radiation. Its brightness often varies over short periods—days or even hours. The galaxy underlying the brilliant image of a quasar may be fairly normal in some of its properties except for the superficial large-scale effects of the quasar at its centre. Quasars apparently are powered by the same mechanism attributed to radio galaxies. They demonstrate in an extreme way what a supermassive object at the centre of a galaxy can do.



Quasar and its companion galaxy colliding. The image on the left shows a quasar and (pointing lower right) one of its spiral arms. In the image on the right, the companion galaxy is visible as a bright spot just above the quasar.

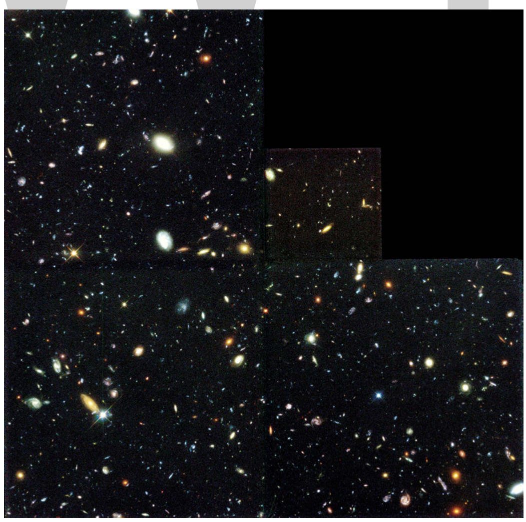
With the gradual recognition of the causes of the quasar phenomenon has come an equally gradual realization that they are simply extreme examples of a process that can be observed in more familiar objects. The black holes that are thought to inhabit the cores of the quasar galaxies are similar to, though more explosive than, those that appear to occur in certain unusual nearer galaxies known as Seyfert galaxies. The radio galaxies fall in between. The reason for the differences in the level of activity is apparently related to the source of the gas and stars that are falling into the centres of such objects, providing the black holes with fuel. In the case of quasars, evidence suggests that an encounter with another galaxy, which causes the latter to be tidally destroyed and its matter to fall into the centre of the more massive quasar galaxy, may be the cause of its activity. As the material approaches the black hole, it is greatly accelerated, and some of it is expelled by the prevailing high temperatures and drastically rapid motions. This process probably also explains the impressive but lower-level activity in the nuclei of radio and Seyfert galaxies. The captured mass may be of lesser amount—i.e., either a smaller galaxy or a portion of the host galaxy itself. Quasars are more common in that part of the universe observed to have redshifts of about 2, meaning that they were more common about 10 or so billion years ago than they are now, which is at least partly a result of the higher density of galaxies at that time.

Gamma-ray Bursts

In the 1970s a new type of object was identified as using orbiting gamma-ray detectors. These "gamma-ray bursts" are identified by extremely energetic flashes of gamma radiation that last only seconds. In some cases the bursts are clearly identified with very distant galaxies, implying immense energies in the bursts. Possibly these are the explosions of "hypernovae," posited to be far more energetic than supernovae and which require some extreme kind of event, such as the merging of two neutron stars.

Evolution of Galaxies and Quasars

The study of the origin and evolution of galaxies and the quasar phenomenon has only just begun. Many models of galaxy formation and evolution have been constructed on the basis of what we know about conditions in the early universe, which is in turn based on models of the expansion of the universe after the big bang (the primordial explosion from which the universe is thought to have originated) and on the characteristics of the cosmic microwave background (the observed photons that show us the light-filled universe as it was when it was a few hundred thousand years old).



The Hubble Deep Field. This image, the result of 10 days', shows 1,500 galaxies in different stages of their evolution.

When the universe had expanded to be cool enough for matter to remain in neutral atoms without being instantly ionized by radiation, structure apparently had already been established in the form of density fluctuations. At a crucial point in time, there condensed from the expanding matter small clouds (protogalaxies) that could collapse under their own gravitational field eventually to form galaxies.

For the latter half of the 20th century, there were two competing models of galaxy formation: “top-down” and “bottom-up.” In the top-down model, galaxies formed out of the collapse of much larger gas clouds. In the bottom-up model, galaxies formed from the merger of smaller entities that were the size of globular clusters. In both models the angular momentum of the original clouds determined the form of the galaxy that eventually evolved. It is thought that a protogalaxy with a large amount of angular momentum tended to form a flat, rapidly rotating system (a spiral galaxy), whereas one with very little angular momentum developed into a more nearly spherical system (an elliptical galaxy).

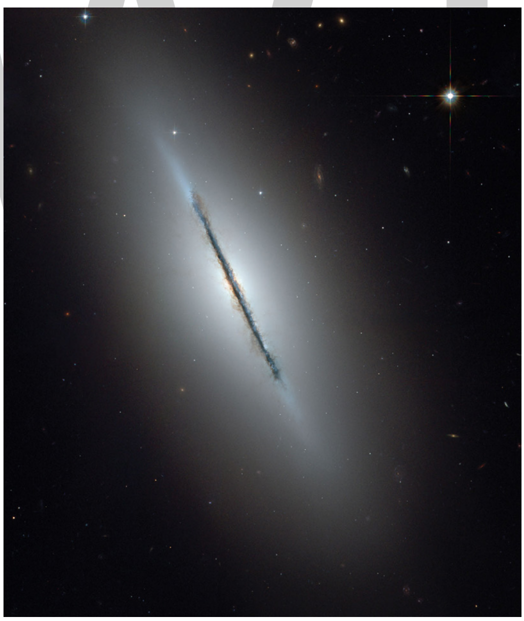
The transition from the 20th to the 21st century coincided with a dramatic transition in our understanding of the evolution of galaxies. It is no longer believed that galaxies have evolved smoothly and alone. Indeed, it has become clear that collisions between galaxies have occurred all during their evolution—and these collisions, far from being rare events, were the mechanism by which galaxies developed in the distant past and are the means by which they are changing their structure and appearance even now. Evidence for this new understanding of galactic evolution comes primarily from two sources: more detailed studies of nearby galaxies with new, more sensitive instruments and deep surveys of extremely distant galaxies, seen when the universe was young.

Recent surveys of nearby galaxies, including the Milky Way Galaxy, have shown evidence of past collisions and capture of galaxies. For the Milky Way the most conspicuous example is the Sagittarius Galaxy, which has been absorbed by our Galaxy. Now its stars lie spread out across the sky, its seven globular clusters intermingling with the globular clusters of the Milky Way Galaxy. Long tails of stars around the Milky Way were formed by the encounter and act as clues to the geometry of the event. A second remnant galaxy, known as the Canis Major Dwarf Galaxy, can also be traced by the detection of star streams in the outer parts of our Galaxy. These galaxies support the idea that the Milky Way Galaxy is a mix of pieces, formed by the amalgamation of many smaller galaxies.

The Andromeda Galaxy (M31) also has a past involving collisions and accretion. Its peculiar close companion, M32, shows a structure that indicates that it was formerly a normal, more massive galaxy that lost much of its outer parts and possibly all of its globular clusters to M31 in a past encounter. Deep surveys of the outer parts of the Andromeda Galaxy have revealed huge coherent structures of star streams and clouds, with properties indicating that these include the outer remnants of smaller galaxies “eaten” by the giant central galaxy, as well as clouds of M31 stars ejected by the strong tidal forces of the collision.

More spectacular are galaxies presently in the process of collision and accretion in the more distant, but still nearby, universe. The symptoms of the collision are the distortion of the galaxies' shape (especially that of the spiral arms), the formation of giant arcs of stars by tidal action, and the enhanced rate of star and star cluster formation. Some of the most massive and luminous young star clusters observed anywhere lie in the regions where two galaxies have come together, with their gas and dust clouds colliding and merging in a spectacular cosmic fireworks display.

A second type of evidence for the fact that galaxies grow by merging comes from very deep surveys of the very distant universe, especially those carried out with the Hubble Space Telescope (HST). These surveys, especially the Hubble Deep Field and the Hubble Ultra Deep Field, found galaxies so far away that the light observed by the HST left them when they were very young, only a few hundred million years old. This enables the direct detection and measurement of young galaxies as they were when the universe was young. The result is a view of a very different universe of galaxies. Instead of giant elliptical galaxies and grand spirals, the universe in its early years was populated with small, irregular objects that looked like mere fragments. These were the building blocks that eventually formed bigger galaxies such as the Milky Way. Many show active formation of stars that are deficient in heavy elements because many of the heavy elements had not yet been created when these stars were formed.

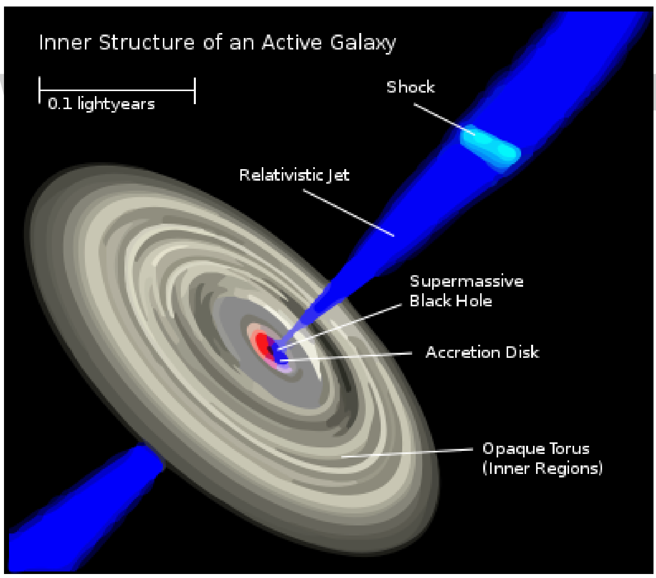


NGC 5866: Image of the disk galaxy NGC 5866 taken with the Advanced Camera for Surveys (ACS) on the Hubble Space Telescope.

The rate of star formation in these early times was significant, but it did not reach a peak until about one billion years later. Galaxies from this time show a maximum in the amount of excited hydrogen, which indicates a high rate of star formation, as young, very hot stars are necessary for exciting interstellar hydrogen so that it can be detected. Since that time, so much matter has been locked up in stars (especially white dwarfs) that not enough interstellar dust and gas are available to achieve such high rates of star formation.

An important development that has helped our understanding of the way galaxies form is the great success of computer simulations. High-speed calculations of the gravitational history of assemblages of stars, interstellar matter, and dark matter suggest that after the big bang the universe developed as a networklike arrangement of material, with gradual condensation of masses where the strands of the network intersected. In simulations of this process, massive galaxies form, but each is surrounded by a hundred or so smaller objects. The small objects may correspond to the dwarf galaxies, such as those that surround the Milky Way Galaxy but of which only a dozen or so remain, the rest having presumably been accreted by the main galaxy. Such computer models, called "n-body simulations," are especially successful in mimicking galaxy collisions and in helping to explain the presence of various tidal arms and jets observed by astronomers.

Active Galactic Nucleus



Inner structure of a galaxy with an active galactic nucleus.

An active galactic nucleus (AGN) is a compact region at the center of a galaxy that has a much higher than normal luminosity over at least some portion of the electromagnetic spectrum with characteristics indicating that the luminosity is not produced by stars. Such excess non-stellar emission has been observed in the radio, microwave, infrared, optical, ultra-violet, X-ray and gamma ray wavebands. A galaxy hosting an AGN is called an "active galaxy". The non-stellar radiation from an AGN is theorized to result from the accretion of matter by a supermassive black hole at the center of its host galaxy.

Active galactic nuclei are the most luminous persistent sources of electromagnetic radiation in the universe, and as such can be used as a means of discovering distant objects; their evolution as a function of cosmic time also puts constraints on models of the cosmos.

The observed characteristics of an AGN depend on several properties such as the mass of the central black hole, the rate of gas accretion onto the black hole, the orientation of the accretion disk, the degree of obscuration of the nucleus by dust, and presence or absence of jets.

Numerous subclasses of AGN have been defined based on their observed characteristics; the most powerful AGN are classified as quasars. A blazar is an AGN with a jet pointed toward the Earth, in which radiation from the jet is enhanced by relativistic beaming.

Models

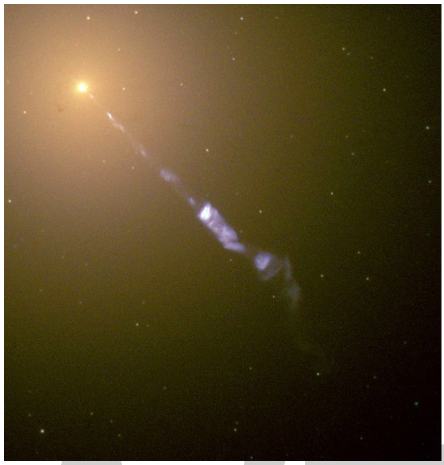
UGC 6093 is classified as an active galaxy, which means that it hosts an active galactic nucleus.

For a long time it has been argued that an AGN must be powered by accretion of mass onto massive black holes (10^6 to 10^{10} times the Solar mass). AGN are both compact and persistently extremely luminous. Accretion can potentially give very efficient conversion of potential and kinetic energy to radiation, and a massive black hole has a high Eddington luminosity, and as a result, it can provide the observed high persistent luminosity. Supermassive black holes are now believed to exist in the centres of most if not all massive galaxies since the mass of the black hole correlates well with the velocity dispersion of the galactic bulge (the M–sigma relation) or with bulge luminosity. Thus AGN-like characteristics are expected whenever a supply of material for accretion comes within the sphere of influence of the central black hole.

Accretion Disc

In the standard model of AGN, cold material close to a black hole forms an accretion disc. Dissipative processes in the accretion disc transport matter inwards and angular momentum outwards, while causing the accretion disc to heat up. The expected spectrum of an accretion disc peaks in the optical-ultraviolet waveband; in addition, a corona of hot material forms above the accretion disc and can inverse-Compton scatter photons up to X-ray energies. The radiation from the accretion disc excites cold atomic material close to the black hole and this in turn radiates at particular emission lines. A large fraction of the AGN's radiation may be obscured by interstellar gas and dust close to the accretion disc, but (in a steady-state situation) this will be re-radiated at some other waveband, most likely the infrared.

Relativistic Jets



Hubble Space Telescope of a 5000-light-year-long jet ejected from the active galaxy M87. The blue synchrotron radiation contrasts with the yellow starlight from the host galaxy.

Some accretion discs produce jets of twin, highly collimated, and fast outflows that emerge in opposite directions from close to the disc. The direction of the jet ejection is determined either by the angular momentum axis of the accretion disc or the spin axis of the black hole. The jet production mechanism and indeed the jet composition on very small scales are not understood at present due to the resolution of astronomical instruments being too low. The jets have their most obvious observational effects in the radio waveband, where very-long-baseline interferometry can be used to study the synchrotron radiation they emit at resolutions of sub-parsec scales. However, they radiate in all wavebands from the radio through to the gamma-ray range via the synchrotron and the inverse-Compton scattering process, and so AGN jets are a second potential source of any observed continuum radiation.

Radiatively Inefficient AGN

There exists a class of 'radiatively inefficient' solutions to the equations that govern accretion. The most widely known of these is the Advection Dominated Accretion Flow (ADAF), but other theories exist. In this type of accretion, which is important for accretion rates well below the Eddington limit, the accreting matter does not form a thin disc and consequently does not efficiently radiate away the energy that it acquired as it moved close to the black hole. Radiatively inefficient accretion has been used to explain the lack of strong AGN-type radiation from massive black holes at the centres of elliptical galaxies in clusters, where otherwise we might expect high accretion rates and correspondingly high luminosities. Radiatively inefficient AGN would be expected to lack many of the characteristic features of standard AGN with an accretion disc.

Observational Characteristics

There is no single observational signature of an AGN. The list below covers some of the features that have allowed systems to be identified as AGN:

- Nuclear optical continuum emission: This is visible whenever there is a direct view of the accretion disc. Jets can also contribute to this component of the AGN emission. The optical emission has a roughly power-law dependence on wavelength.
- Nuclear infra-red emission: This is visible whenever the accretion disc and its environment are obscured by gas and dust close to the nucleus and then re-emitted ('reprocessing'). As it is thermal emission, it can be distinguished from any jet or disc-related emission.
- Broad optical emission lines: These come from cold material close to the central black hole. The lines are broad because the emitting material is revolving around the black hole with high speeds causing a range of Doppler shifts of the emitted photons.
- Narrow optical emission lines: These come from more distant cold material, and so are narrower than the broad lines.
- Radio continuum emission: This is always due to a jet. It shows a spectrum characteristic of synchrotron radiation.
- X-ray continuum emission: This can arise both from a jet and from the hot corona of the accretion disc via a scattering process: in both cases it shows a power-law spectrum. In some radio-quiet AGN there is an excess of soft X-ray emission in addition to the power-law component. The origin of the soft X-rays is not clear at present.
- X-ray line emission: This is a result of illumination of cold heavy elements by the X-ray continuum that causes fluorescence of X-ray emission lines, the best-known of which is the iron feature around 6.4 keV. This line may be narrow or broad: relativistically broadened iron lines can be used to study the dynamics of the accretion disc very close to the nucleus and therefore the nature of the central black hole.

Types of Active Galaxy

It is convenient to divide AGN into two classes, conventionally called radio-quiet and radio-loud. Radio-loud objects have emission contributions from both the jet(s) and the lobes that the jets inflate. These emission contributions dominate the luminosity of the AGN at radio wavelengths and possibly at some or all other wavelengths. Radio-quiet objects are simpler since jet and any jet-related emission can be neglected at all wavelengths.

AGN terminology is often confusing, since the distinctions between different types of AGN sometimes reflect historical differences in how the objects were discovered or initially classified, rather than real physical differences.

Radio-quiet AGN

- Low-ionization nuclear emission-line regions (LINERs): As the name suggests, these systems show only weak nuclear emission-line regions, and no other signatures of AGN emission. It is debatable whether all such systems are true AGN (powered by accretion on to a supermassive black hole). If they are, they constitute the lowest-luminosity class of radio-quiet AGN. Some may be radio-quiet analogues of the low-excitation radio galaxies.

- Seyfert galaxies: Seyferts were the earliest distinct class of AGN to be identified. They show optical range nuclear continuum emission, narrow and occasionally broad emission lines, occasionally strong nuclear X-ray emission and sometimes a weak small-scale radio jet. Originally they were divided into two types known as Seyfert 1 and 2: Seyfert 1s show strong broad emission lines while Seyfert 2s do not, and Seyfert 1s are more likely to show strong low-energy X-ray emission. Various forms of elaboration on this scheme exist: for example, Seyfert 1s with relatively narrow broad lines are sometimes referred to as narrow-line Seyfert 1s. The host galaxies of Seyferts are usually spiral or irregular galaxies.

- Radio-quiet quasars/QSOs: These are essentially more luminous versions of Seyfert 1s: the distinction is arbitrary and is usually expressed in terms of a limiting optical magnitude. Quasars were originally 'quasi-stellar' in optical images as they had optical luminosities that were greater than that of their host galaxy. They always show strong optical continuum emission, X-ray continuum emission, and broad and narrow optical emission lines. Some astronomers use the term QSO (Quasi-Stellar Object) for this class of AGN, reserving 'quasar' for radio-loud objects, while others talk about radio-quiet and radio-loud quasars. The host galaxies of quasars can be spirals, irregulars or ellipticals. There is a correlation between the quasar's luminosity and the mass of its host galaxy, in that the most luminous quasars inhabit the most massive galaxies (ellipticals).

- 'Quasar 2s': By analogy with Seyfert 2s, these are objects with quasar-like luminosities but without strong optical nuclear continuum emission or broad line emission. They are scarce in surveys, though a number of possible candidate quasar 2s have been identified.

Radio-loud AGN

- Radio-loud quasars behave exactly like radio-quiet quasars with the addition of emission from a jet. Thus they show strong optical continuum emission, broad and narrow emission lines, and strong X-ray emission, together with nuclear and often extended radio emission.

- "Blazars" (BL Lac objects and OVV quasars) classes are distinguished by rapidly variable, polarized optical, radio and X-ray emission. BL Lac objects show no optical emission lines, broad or narrow, so that their redshifts can only be determined from features in the spectra of their host galaxies. The emission-line features may be intrinsically absent or simply swamped by the additional variable component. In the latter case, emission lines may become visible when the variable component is at a low level. OVV quasars behave more like standard radio-loud quasars with the addition of a rapidly variable component. In both classes of source, the variable emission is believed to originate in a relativistic jet oriented close to the line of sight. Relativistic effects amplify both the luminosity of the jet and the amplitude of variability.
- Radio galaxies: These objects show nuclear and extended radio emission. Their other AGN properties are heterogeneous. They can broadly be divided into low-excitation and high-excitation classes. Low-excitation objects show no strong narrow or broad emission lines, and the emission lines they do have may be excited by a different mechanism. Their optical and X-ray nuclear emission is consistent with originating purely in a jet. They may be the best current candidates for AGN with radiatively inefficient accretion. By contrast, high-excitation objects (narrow-line radio galaxies) have emission-line spectra similar to those of Seyfert 2s. The small class of broad-line radio galaxies, which show relatively strong nuclear optical continuum emission probably includes some objects that are simply low-luminosity radio-loud quasars. The host galaxies of radio galaxies, whatever their emission-line type, are essentially always ellipticals.

Table: Features of different types of galaxies.

Galaxy type	Active nuclei	Emission lines		X-rays	Excess of		Strong radio	Jets	Variable	Radio loud
		Narrow	Broad		UV	Far-IR				
Normal	no	weak	no	weak	no	no	no	no	no	no
LINER	unknown	weak	weak	weak	no	no	no	no	no	no
Seyfert I	yes	yes	yes	some	some	yes	few	no	yes	no
Seyfert II	yes	yes	no	some	some	yes	few	no	yes	no
Quasar	yes	yes	yes	some	yes	yes	some	some	yes	some
Blazar	yes	no	some	yes	yes	no	yes	yes	yes	yes
BL Lac	yes	no	no/faint	yes	yes	no	yes	yes	yes	yes

OVV	yes	no	stronger than BL Lac	yes	yes	no	yes	yes	yes	yes
Radio galaxy	yes	some	some	some	some	yes	yes	yes	yes	yes

Unification of AGN Species

Unified models propose that different observational classes of AGN are a single type of physical object observed under different conditions. The currently favoured unified models are 'orientation-based unified models' meaning that they propose that the apparent differences between different types of objects arise simply because of their different orientations to the observer. However, they are debated.

Radio-quiet Unification

At low luminosities, the objects to be unified are Seyfert galaxies. The unification models propose that in Seyfert 1s the observer has a direct view of the active nucleus. In Seyfert 2s the nucleus is observed through an obscuring structure which prevents a direct view of the optical continuum, broad-line region or (soft) X-ray emission. The key insight of orientation-dependent accretion models is that the two types of object can be the same if only certain angles to the line of sight are observed. The standard picture is of a torus of obscuring material surrounding the accretion disc. It must be large enough to obscure the broad-line region but not large enough to obscure the narrow-line region, which is seen in both classes of object. Seyfert 2s are seen through the torus. Outside the torus there is material that can scatter some of the nuclear emission into our line of sight, allowing us to see some optical and X-ray continuum and, in some cases, broad emission lines—which are strongly polarized, showing that they have been scattered and proving that some Seyfert 2s really do contain hidden Seyfert 1s. Infrared observations of the nuclei of Seyfert 2s also support this picture.

At higher luminosities, quasars take the place of Seyfert 1s, but, the corresponding 'quasar 2s' are elusive at present. If they do not have the scattering component of Seyfert 2s they would be hard to detect except through their luminous narrow-line and hard X-ray emission.

Radio-loud Unification

Historically, work on radio-loud unification has concentrated on high-luminosity radio-loud quasars. These can be unified with narrow-line radio galaxies in a manner directly analogous to the Seyfert 1/2 unification (but without the complication of much in the way of a reflection component: Narrow-line radio galaxies show no nuclear optical continuum or reflected X-ray component, although they do occasionally show polarized broad-line emission). The large-scale radio structures of these objects provide compelling evidence that the orientation-based unified models really are true. X-ray evidence, where available, supports the unified picture: Radio galaxies show evidence of obscuration from a torus, while quasars do not, although care must be taken since radio-loud objects also have a soft unabsorbed jet-related component, and high resolution is necessary to separate out thermal emission from the sources' large-scale hot-gas environment. At very small angles to the line of sight, relativistic beaming dominates, and we see a blazar of some variety.

However, the population of radio galaxies is completely dominated by low-luminosity, low-excitation objects. These do not show strong nuclear emission lines—broad or narrow—they have optical continua which appear to be entirely jet-related and their X-ray emission is also consistent with coming purely from a jet, with no heavily absorbed nuclear component in general. These objects cannot be unified with quasars, even though they include some high-luminosity objects when looking at radio emission, since the torus can never hide the narrow-line region to the required extent, and since infrared studies show that they have no hidden nuclear component, in fact there is no evidence for a torus in these objects at all. Most likely, they form a separate class in which only jet-related emission is important. At small angles to the line of sight, they will appear as BL Lac objects.

Cosmological uses and Evolution

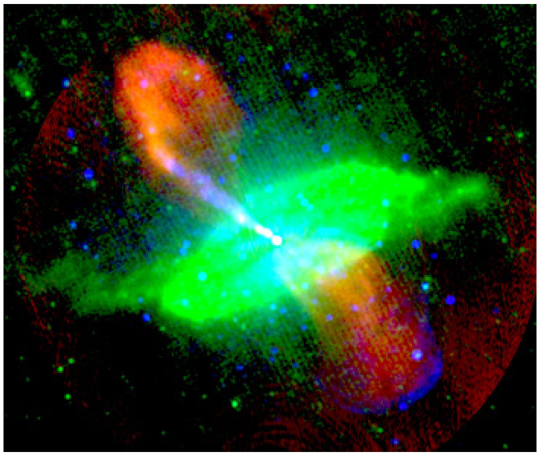
For a long time, active galaxies held all the records for the highest-redshift objects known either in the optical or the radio spectrum, because of their high luminosity. They still have a role to play in studies of the early universe, but it is now recognised that an AGN gives a highly biased picture of the "typical" high-redshift galaxy.

Most luminous classes of AGN (radio-loud and radio-quiet) seem to have been much more numerous in the early universe. This suggests that massive black holes formed early on and that the conditions for the formation of luminous AGN were more common in the early universe, such as a much higher availability of cold gas near the centre of galaxies than at present. It also implies that many objects that were once luminous quasars are now much less luminous, or entirely quiescent. The evolution of the low-luminosity AGN population is much less well understood due to the difficulty of observing these objects at high redshifts.

Radio Galaxy

Radio galaxies and their relatives, radio-loud quasars and blazars, are types of active galaxy nuclei that are very luminous at radio wavelengths, with luminosities up to 10^{39} W between 10 MHz and 100 GHz. The radio emission is due to the synchrotron process. The observed structure in radio emission is determined by the interaction between twin jets and the external medium, modified by the effects of relativistic beaming. The host galaxies are almost exclusively large elliptical galaxies. *Radio-loud* active galaxies can be detected at large distances, making them valuable tools for observational cosmology.

Recently, much work has been done on the effects of these objects on the intergalactic medium, particularly in galaxy groups and clusters. The most distant radio galaxy currently known is TGSS J1530+1049, at a redshift of 5.72.



False-colour image of the nearby radio galaxy Centaurus A, showing radio (red), 24-micrometre infrared (green) and 0.5-5 keV X-ray emission (blue). The jet can be seen to emit synchrotron radiation in all three wavebands. The lobes only emit in the radio frequency range, and so appear red. Gas and dust in the galaxy emits thermal radiation in the infrared. Thermal X-ray radiation from hot gas and non-thermal emission from relativistic electrons can be seen in the blue 'shells' around the lobes, particularly to the south (bottom).

Emission Processes

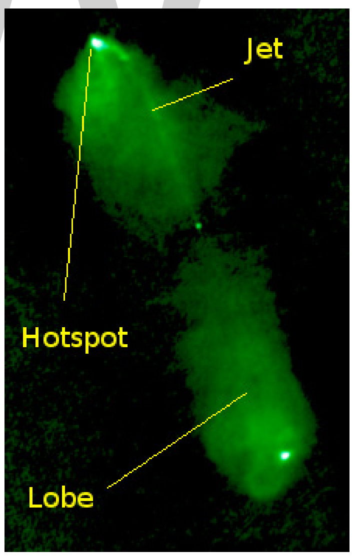
The radio emission from radio-loud active galaxies is synchrotron emission, as inferred from its very smooth, broad-band nature and strong polarization. This implies that the radio-emitting plasma contains, at least, electrons with relativistic speeds (Lorentz factors of ~10^4) and magnetic fields. Since the plasma must be neutral, it must also contain either protons or positrons. There is no way of determining the particle content directly from observations of synchrotron radiation. Moreover, there is no way to determine the energy densities in particles and magnetic fields from observation: the same synchrotron emissivity may be a result of a few electrons and a strong field, or a weak field and many electrons, or something in between. It is possible to determine a minimum energy condition which is the minimum energy density that a region with a given emissivity can have, but for many years there was no particular reason to believe that the true energies were anywhere near the minimum energies.

A sister process to the synchrotron radiation is the inverse-Compton process, in which the relativistic electrons interact with ambient photons and Thomson scatter them to high energies. Inverse-Compton emission from radio-loud sources turns out to be particularly important in X-rays, and, because it depends only on the density of electrons, a detection of inverse-Compton scattering allows a somewhat model-dependent estimate of the energy densities in the particles and magnetic fields. This has been used to argue that many powerful sources are actually quite near the minimum-energy condition.

Synchrotron radiation is not confined to radio wavelengths: if the radio source can accelerate particles to high enough energies, features that are detected in the radio wavelengths may also be seen in the infrared, optical, ultraviolet or even X-ray. In the latter case the responsible electrons must have energies in excess of 1 TeV in typical magnetic field strengths. Again, polarization and continuum spectrum are used to distinguish the synchrotron radiation from other emission processes. Jets and hotspots are the usual sources of high-frequency synchrotron emission. It is hard to distinguish observationally between the synchrotron and inverse-Compton radiation, making them a subject of ongoing research.

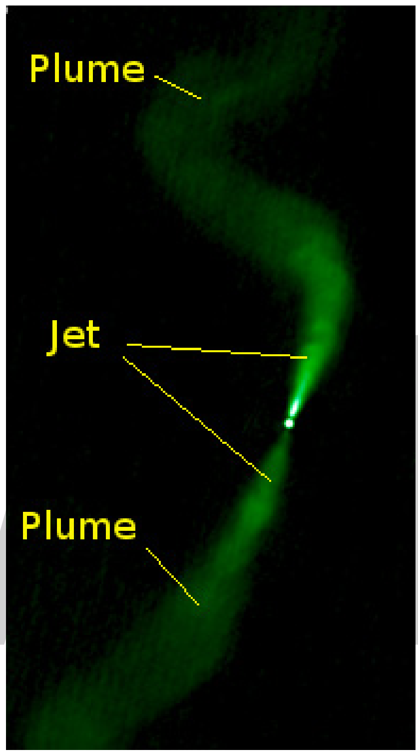
Processes, collectively known as particle acceleration, produce populations of relativistic and non-thermal particles that give rise to synchrotron and inverse-Compton radiation. Fermi acceleration is one plausible particle acceleration process in radio-loud active galaxies.

Radio Structures



Pseudo-colour image of the large-scale radio structure of the FRII radio galaxy 3C98. Lobes, jet and hotspot are labelled.

Radio galaxies, and to a lesser extent, radio-loud quasars display a wide range of structures in radio maps. The most common large-scale structures are called lobes: these are double, often fairly symmetrical, roughly ellipsoidal structures placed on either side of the active nucleus. A significant minority of low-luminosity sources exhibit structures usually known as plumes which are much more elongated. Some radio galaxies show one or two long narrow features known as jets (the most famous example being the giant galaxy M87 in the Virgo cluster) coming directly from the nucleus and going to the lobes. Since the 1970s, the most widely accepted model has been that the lobes or plumes are powered by beams of high-energy particles and magnetic field coming from close to the active nucleus. The jets are believed to be the visible manifestations of the beams, and often the term *jet* is used to refer both to the observable feature and to the underlying flow.



Pseudo-colour image of the large-scale radio structure of the FRI radio galaxy 3C31. Jets and plumes are labelled.

In 1974, radio sources were divided by Fanaroff and Riley into two classes, now known as Fanaroff and Riley Class I (FRI), and Class II (FRII). The distinction was originally made based on the morphology of the large-scale radio emission (the type was determined by the distance between the brightest points in the radio emission): FRI sources were brightest towards the centre, while FRII sources were brightest at the edges. Fanaroff and Riley observed that there was a reasonably sharp divide in luminosity between the two classes: FRIs were low-luminosity, FRIIs were high luminosity. With more detailed radio observations, the morphology turns out to reflect the method of energy transport in the radio source. FRI objects typically have bright jets in the centre, while FRIIs have faint jets but bright *hotspots* at the ends of the lobes. FRIIs appear to be able to transport energy efficiently to the ends of the lobes, while FRI beams are inefficient in the sense that they radiate a significant amount of their energy away as they travel.

The FRI/FRII division depends on host-galaxy environment in the sense that the FRI/FRII transition appears at higher luminosities in more massive galaxies. FRI jets are known to be decelerating in the regions in which their radio emission is brightest, and so it seems that the FRI/FRII transition reflects whether a jet/beam can propagate through the host galaxy without being decelerated to sub-relativistic speeds by interaction with the intergalactic medium. From analysis of relativistic beaming effects, the jets of FRII sources are known to remain relativistic (with speeds of at least 0.5c) out to the ends of the lobes. The hotspots that are usually seen in FRII sources are interpreted as being the visible manifestations of shocks formed when the fast, and therefore supersonic, jet (the speed of sound cannot exceed $c/\sqrt{3}$) abruptly terminates at the end of the source, and their spectral energy distributions are consistent with this picture. Often multiple hotspots are seen, reflecting either continued outflow after the shock or movement of the jet termination point: the overall hotspot region is sometimes called the hotspot complex.

Names are given to several particular types of radio source based on their radio structure:

- Classical double refers to an FRII source with clear hotspots.
- Wide-angle tail normally refers to a source intermediate between standard FRI and FRII structure, with efficient jets and sometimes hotspots, but with plumes rather than lobes, found at or near the centres of clusters.
- Narrow-angle tail or Head-tail source describes an FRI that appears to be bent by ram pressure as it moves through a cluster.
- Fat doubles are sources with diffuse lobes but neither jets nor hotspots. Some such sources may be relics whose energy supply has been permanently or temporarily turned off.

Life Cycles and Dynamics

The largest radio galaxies have lobes or plumes extending to megaparsec scales (more in the case of giant radio galaxies like 3C236), implying a timescale for growth of the order of tens to hundreds of millions of years. This means that, except in the case of very small, very young sources, we cannot observe radio source dynamics directly, and so must resort to theory and inferences from large numbers of objects. Clearly radio sources must start small and grow larger. In the case of sources with lobes, the dynamics are fairly simple: the jets feed the lobes, the pressure of the lobes increases, and the lobes expand. How fast they expand depends on the density and pressure of the external medium. The highest-pressure phase of the external medium, and thus the most important phase from the point of view of the dynamics, is the X-ray emitting diffuse hot gas. For a long time it was assumed that powerful sources would expand supersonically, pushing a shock through the external medium. However, X-ray observations show that the internal lobe pressures of powerful FRII sources are often close to the external thermal pressures and not much higher than the external pressures, as would be required for supersonic expansion. The only unambiguously supersonically expanding system known consists of the inner lobes of the low-power radio galaxy Centaurus A which are probably a result of a comparatively recent outburst of the active nucleus.

Host Galaxies and Environments

These radio sources are almost universally found hosted by elliptical galaxies, though there is one well-documented exception, namely NGC 4151. Some Seyfert galaxies show weak, small radio jets, but they are not radio-luminous enough to be classified as radio-loud. Such information as there is about the host galaxies of radio-loud quasars and blazars suggests that they are also hosted by elliptical galaxies.

There are several possible reasons for this very strong preference for ellipticals. One is that ellipticals generally contain the most massive black holes, and so are capable of powering the most luminous active galaxies. Another is that ellipticals generally inhabit richer environments, providing a large-scale intergalactic medium to confine the radio source. It may also be that the larger amounts of cold gas in spiral galaxies in some way disrupts or stifles a forming jet. To date there is no compelling single explanation for the observations.

Unified Models

The different types of radio-loud active galaxies are linked by unified models. The key observation that led to the adoption of unified models for powerful radio galaxies and radio-loud quasars was that all quasars appear to be beamed towards us, showing superluminal motion in the cores and bright jets on the side of the source nearest to us (the Laing-Garrington effect). If this is the case, there must be a population of objects not beamed towards us, and, since we know the lobes are not affected by beaming, they would appear as radio galaxies, provided that the quasar nucleus is obscured when the source is seen side-on. It is now accepted that at least some powerful radio galaxies have 'hidden' quasars, though it is not clear whether all such radio galaxies would be quasars if viewed from the right angle. In a similar way, low-power radio galaxies are a plausible parent population for BL Lac objects.

Uses of Radio Galaxies

Distant Sources

Radio galaxies and radio-loud quasars have been widely used, particularly in the 80s and 90s, to find distant galaxies: by selecting based on radio spectrum and then observing the host galaxy it was possible to find objects at high redshift at modest cost in telescope time. The problem with this method is that hosts of active galaxies may not be typical of galaxies at their redshift. Similarly, radio galaxies have in the past been used to find distant X-ray emitting clusters, but unbiased selection methods are now preferred.

Standard Rulers

Some work has been done attempting to use radio galaxies as standard rulers to determine cosmological parameters. This method is fraught with difficulty because a radio galaxy's size depends on both its age and its environment. When a model of the radio source is used, though, methods based on radio galaxies can give good agreement with other cosmological observations.

Effects on Environment

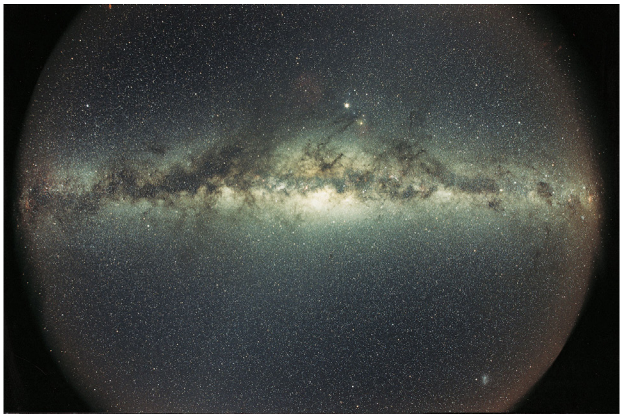
Whether or not a radio source is expanding supersonically, it must do work against the external medium in expanding, and so it puts energy into heating and lifting the external plasma. The minimum energy stored in the lobes of a powerful radio source might be 10^{53} J. The lower limit on the work done on the external medium by such a source is several times this. A good deal of the current interest in radio sources focuses on the effect they must have at the centres of clusters at the present day. Equally interesting is their likely effect on structure formation over cosmological time: it is thought that they may provide a feedback mechanism to slow the formation of the most massive objects.

Terminology

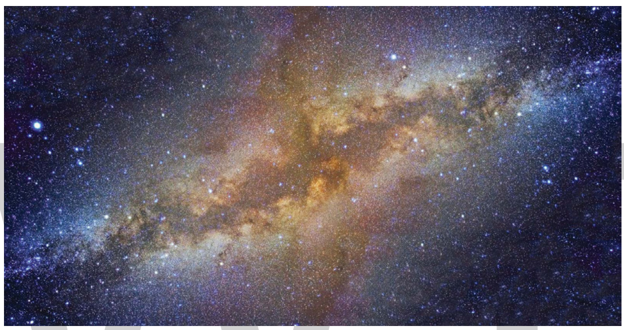
Widely used terminology is awkward now that it is generally accepted that quasars and radio galaxies are the same objects. The acronym *DRAGN* (for 'Double Radiosource Associated with Galactic Nucleus') has been coined but has not yet taken off. Extragalactic radio source is common but can lead to confusion, since many other extragalactic objects are detected in radio surveys, notably starburst galaxies.

The Milky Way Galaxy

The Milky Way is a large spiral system consisting of several hundred billion stars, one of which is the Sun. It takes its name from the Milky Way, the irregular luminous band of stars and gas clouds that stretches across the sky as seen from Earth. Although Earth lies well within the Milky Way Galaxy (sometimes simply called the Galaxy), astronomers do not have as complete an understanding of its nature as they do of some external star systems. A thick layer of interstellar dust obscures much of the Galaxy from scrutiny by optical telescopes, and astronomers can determine its large-scale structure only with the aid of radio and infrared telescopes, which can detect the forms of radiation that penetrate the obscuring matter.



Milky Way Galaxy as seen from Earth.



The Milky Way Galaxy in the night sky.

Major Components of the Galaxy

Emission Nebulae

A conspicuous component of the Galaxy is the collection of large, bright, diffuse gaseous objects generally called nebulae. The brightest of these cloudlike objects are the emission nebulae, large complexes of interstellar gas and stars in which the gas exists in an ionized and excited state (with the electrons of the atoms excited to a higher than normal energy level). This condition is produced by the strong ultraviolet light emitted from the very luminous, hot stars embedded in the gas. Because emission nebulae consist almost entirely of ionized hydrogen, they are usually referred to as H II regions.

Astronomers have identified some 700 young stars in this 2.5-light-year-wide area. They have also detected over 150 protoplanetary disks, or proplyds, which are believed to be embryonic solar systems that will eventually form planets. These stars and proplyds generate most of the nebula's light. This picture is a mosaic combining 45 images taken by the Hubble Space Telescope.

H II regions are found in the plane of the Galaxy intermixed with young stars, stellar associations, and the youngest of the open clusters. They are areas where very massive stars have recently formed, and many contain the uncondensed gas, dust, and molecular complexes commonly associated with ongoing star formation. The H II regions are concentrated in the spiral arms of the Galaxy, though some exist between the arms. Many of them are found at intermediate distances from the centre of the Milky Way Galaxy, with the largest number occurring at a distance of 10,000 light-years. This latter fact can be ascertained even though the H II regions cannot be seen clearly beyond a few thousand light-years from the Sun. They emit radio radiation of a characteristic type, with a thermal spectrum that indicates that their temperatures are about 10,000 kelvins. This thermal radio radiation enables astronomers to map the distribution of H II regions in distant parts of the Galaxy.

Centre of the Orion Nebula (M42).

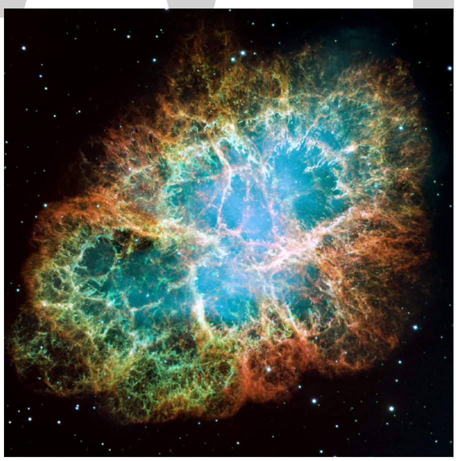
The largest and brightest H II regions in the Galaxy rival the brightest star clusters in total luminosity. Even though most of the visible radiation is concentrated in a few discrete emission lines, the total apparent brightness of the brightest is the equivalent of tens of thousands of solar luminosities. These H II regions are also remarkable in size, having diameters of about 1,000 light-years. More typically, common H II regions such as the Orion Nebula are about 50 light-years across. They contain gas that has a total mass ranging from one or two solar masses up to several thousand. H II regions consist primarily of hydrogen, but they also contain measurable amounts of other gases. Helium is second in abundance, and large amounts of carbon, nitrogen, and oxygen occur as well. Preliminary evidence indicates that the ratio of the abundance of the heavier elements among the detected gases to hydrogen decreases outward from the centre of the Galaxy, a tendency that has been observed in other spiral galaxies.

Planetary Nebulae

The gaseous clouds known as planetary nebulae are only superficially similar to other types of nebulae. So called because the smaller varieties almost resemble planetary disks when viewed through a telescope, planetary nebulae represent a stage at the end of the stellar life cycle rather than one at the beginning. The distribution of such nebulae in the Galaxy is different from that of H II regions. Planetary nebulae belong to an intermediate population and are found throughout the disk and the inner halo. There are more than 1,000 known planetary nebulae in the Galaxy, but more might be overlooked because of obscuration in the Milky Way region.

Composite picture of the Cat's Eye Nebula (NGC 6543), combining three images taken by the Hubble Space Telescope. This planetary nebula has an unusually complicated structure, with concentric shells (seen as bright rings), jets (the projections at upper left and lower right), and a number of details that suggest complex interactions of shock waves.

Supernova Remnants



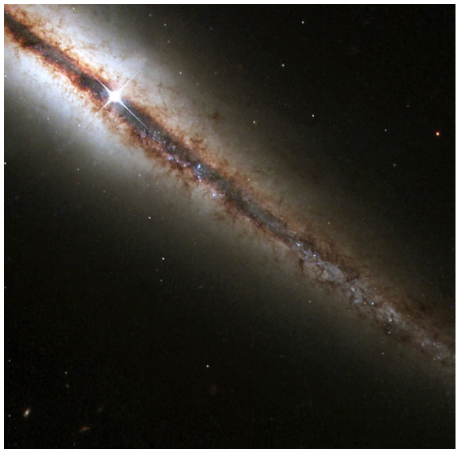
The Crab Nebula, which was formed by a supernova explosion recorded in 1054. This image was made by combining two dozen exposures from the Hubble Space Telescope.

Another type of nebulous object found in the Galaxy is the remnant of the gas blown out from an exploding star that forms a supernova. Occasionally these objects look something like planetary nebulae, as in the case of the Crab Nebula, but they differ from the latter in three ways: (1) the total mass of their gas (they involve a larger mass, essentially all the mass of the exploding star), (2) their kinematics (they are expanding with higher velocities), and (3) their lifetimes (they last for a shorter time as visible nebu-lae). The best-known supernova remnants are those resulting from three historically observed supernovae: That of 1054, which made the Crab Nebula its remnant; that of 1572, called Tycho's Nova; and that of 1604, called Kepler's Nova. These objects and the many others like them in the Galaxy are detected at radio wavelengths. They release radio energy in a nearly flat spectrum because of the emission of radiation by charged particles moving spirally at nearly the speed of light in a magnetic field enmeshed in the gaseous remnant. Radiation generated in this way is called synchrotron radiation and is associated with various types of violent cosmic phenomena besides supernova remnants, as, for example, radio galaxies.

Dust Clouds

The dust clouds of the Galaxy are narrowly limited to the plane of the Milky Way, though very low-density dust can be detected even near the galactic poles. Dust clouds beyond 2,000 to 3,000 light-years from the Sun cannot be detected optically, because intervening clouds of dust and the general dust layer obscure more distant views. Based on the distribution of dust clouds in other galaxies, it can be concluded that they are often most conspicuous within the spiral arms, especially along the inner edge of well-defined ones. The best-observed dust clouds near the Sun have masses of several hundred solar masses and sizes ranging from a maximum of about 200 light-years to a fraction of a light-year. The smallest tend to be the densest, possibly partly because of evolution: as a dust complex contracts, it also becomes denser and more opaque. The very smallest dust clouds are the so-called Bok globules, named after the Dutch American astronomer Bart J. Bok; these objects are about one light-year across and have masses of 1–20 solar masses.

The Eagle Nebula. Stars are forming in this column of cold dust and gas, which is 9.5 light-years in length.



NGC 4013, a spiral galaxy, which has a prominent dust lane like the Milky Way Galaxy, in an image taken by the Hubble Space Telescope.

More complete information on the dust in the Galaxy comes from infrared observations. While optical instruments can detect the dust when it obscures more distant objects or when it is illuminated by very nearby stars, infrared telescopes are able to register the long-wavelength radiation that the cool dust clouds themselves emit. A complete survey of the sky at infrared wavelengths made during the early 1980s by an unmanned orbiting observatory, the Infrared Astronomical Satellite (IRAS), revealed a large number of dense dust clouds in the Milky Way.

Thick clouds of dust in the Milky Way can be studied by still another means. Many such objects contain detectable amounts of molecules that emit radio radiation at wavelengths that allow them to be identified and analyzed. More than 50 different molecules, including carbon monoxide and formaldehyde, and radicals have been detected in dust clouds.

The General Interstellar Medium

Horsehead Nebula.

The stars in the Galaxy, especially along the Milky Way, reveal the presence of a general, all-pervasive interstellar medium by the way in which they gradually fade with distance. This occurs primarily because of interstellar dust, which obscures and reddens starlight. On the average, stars near the Sun are dimmed by a factor of two for every 3,000 light-years. Thus, a star that is 6,000 light-years away in the plane of the Galaxy will appear four times fainter than it would otherwise were it not for the interstellar dust.

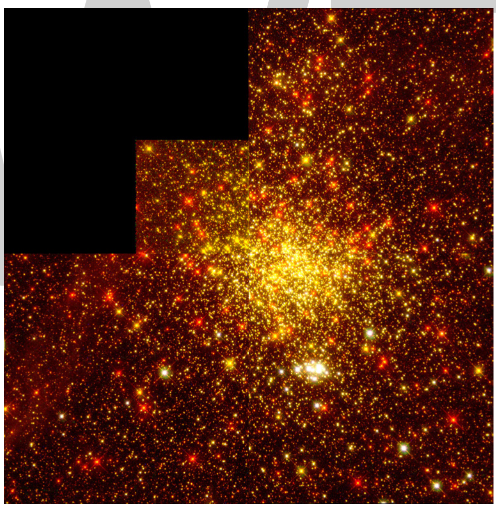
Central regions of the Milky Way Galaxy. The image on the left is in visible light, and the image on the right is in infrared; the marked difference between the two images shows how infrared radiation can penetrate galactic dust. The infrared image is part of the Two Micron All Sky Survey (2MASS), a survey of the entire sky in infrared light.

Another way in which the effects of interstellar dust become apparent is through the polarization of background starlight. Dust is aligned in space to some extent, and this results in selective absorption such that there is a preferred plane of vibration for the light waves. The electric vectors tend to lie preferentially along the galactic plane, though there are areas where the distribution is more complicated. It is likely that the polarization arises because the dust grains are partially aligned by the galactic magnetic field. If the dust grains are paramagnetic so that they act somewhat like a magnet, then the general magnetic field, though very weak, can in time line up the grains with their short axes in the direction of the field. As a consequence, the directions of polarization for stars in different parts of the sky make it possible to plot the direction of the magnetic field in the Milky Way.

The dust is accompanied by gas, which is thinly dispersed among the stars, filling the space between them. This interstellar gas consists mostly of hydrogen in its neutral form. Radio telescopes can detect neutral hydrogen because it emits radiation at a wavelength of 21 cm. Such radio wavelength is long enough to penetrate interstellar dust and so can be detected from all parts of the Galaxy. Most of what astronomers have learned about the large-scale structure and motions of the Galaxy has been derived from the radio waves of interstellar neutral hydrogen. The distance to the gas detected is not easily determined. Statistical arguments must be used in many cases, but the velocities of the gas, when compared with the velocities found for stars and those anticipated on the basis of the dynamics of the Galaxy, provide useful clues as to the location of the different sources of hydrogen radio emission. Near the Sun the average density of interstellar gas is 10^{-21} gm/cm^3, which is the equivalent of about one hydrogen atom per cubic centimetre.

Even before they first detected the emission from neutral hydrogen in 1951, astronomers were aware of interstellar gas. Minor components of the gas, such as sodium and calcium, absorb light at specific wavelengths, and they thus cause the appearance of absorption lines in the spectra of the stars that lie beyond the gas. Since the lines originating from stars are usually different, it is possible to distinguish the lines of the interstellar gas and to measure both the density and velocity of the gas. Frequently it is even possible to observe the effects of several concentrations of interstellar gas between Earth and the background stars and thereby determine the kinematics of the gas in different parts of the Galaxy.

Companion Galaxies



Globular cluster NGC 1850 in the Large Magellanic Cloud. Most of the cluster consists of yellow stars; the bright white stars are members of a second, open cluster about 200 light-years beyond NGC.

The Magellanic Clouds were recognized early in the 20th century as companion objects to the Galaxy. When American astronomer Edwin Hubble established the extragalactic nature of what we now call galaxies, it became plain that the Clouds had to be separate systems, both of the irregular class and more than 100,000 light-years distant. (The current best values for their distances are 163,000 and 202,000 light-years for the Large and Small Clouds, respectively). Additional close companions have been found, all of them small and inconspicuous objects of the dwarf elliptical class. The nearest of these is the Sagittarius dwarf, a galaxy that is falling into the Milky Way Galaxy, having been captured tidally by the Galaxy's much stronger gravity. The core of this galaxy is about 90,000 light-years distant. Other close companions are the well-studied Carina, Draco, Fornax, Leo I, Leo II, Sextans, Sculptor, and Ursa Minor galaxies, as well as several very faint, less well-known objects. Distances for them range from approximately 200,000 to 800,000 light-years. The grouping of these galaxies around the Milky Way Galaxy is mimicked in the case of the Andromeda Galaxy, which is also accompanied by several dwarf companions.

Star Populations and Movement

Stars and Stellar Populations

The concept of different populations of stars has undergone considerable change over the last several decades. Before the 1940s, astronomers were aware of differences between stars and had largely accounted for most of them in terms of different masses, luminosities, and orbital characteristics around the Galaxy. Understanding of evolutionary differences, however, had not yet been achieved, and, although differences in the chemical abundances in the stars were known, their significance was not comprehended. At this juncture, chemical differences seemed exceptional and erratic and remained uncorrelated with other stellar properties. There was still no systematic division of stars even into different kinematic families, in spite of the advances in theoretical work on the dynamics of the Galaxy.

Principal Population Types

In 1944 the German-born astronomer Walter Baade announced the successful resolution into stars of the centre of the Andromeda Galaxy, M31, and its two elliptical companions, M32 and NGC 205. He found that the central parts of Andromeda and the accompanying galaxies were resolved at very much fainter magnitudes than were the outer spiral arm areas of M31. Furthermore, by using plates of different spectral sensitivity and coloured filters, he discovered that the two ellipticals and the centre of the spiral had red giants as their brightest stars rather than blue main-sequence stars, as in the case of the spiral arms. This finding led Baade to suggest that these galaxies, and also the Milky Way Galaxy, are made of two populations of stars that are distinct in their physical properties as well as their locations. He applied the term Population I to the stars that constitute the spiral arms of Andromeda and to most of the stars that are visible in the Milky Way system in the neighbourhood of the Sun. He found that these Population I objects were limited to the flat disk of the spirals and suggested that they were absent from the centres of such galaxies and from the ellipticals entirely. Baade designated as Population II the bright red giant stars that he discovered in the ellipticals and in the nucleus of Andromeda. Other objects that seemed to contain the brightest stars of this class were the globular clusters of the Galaxy. Baade further suggested that the high-velocity stars near the Sun were Population II objects that happened to be passing through the disk.

As a result of Baade's pioneering work on other galaxies in the Local Group (the cluster of star systems to which the Milky Way Galaxy belongs), astronomers immediately applied the notion of two stellar populations to the Galaxy. It is possible to segregate various components of the Galaxy into the two population types by applying both the idea of kinematics of different populations suggested by their position in the Andromeda system and the dynamical theories that relate galactic orbital properties with z distances (the distances above the plane of the Galaxy) for different stars. For many of these objects, the kinematic data on velocities are the prime source of population classification. The Population I component of the Galaxy, highly limited to the flat plane of the system, contains such objects as open star clusters, O and B stars, Cepheid variables, emission nebulae, and neutral hydrogen. Its Population II component, spread over a more nearly spherical volume of space, includes globular clusters, RR Lyrae variables, high-velocity stars, and certain other rarer objects.

As time progressed, it was possible for astronomers to subdivide the different populations in the Galaxy further. These subdivisions ranged from the nearly spherical “halo Population II” system to the very thin “extreme Population I” system. Each subdivision was found to contain (though not exclusively) characteristic types of stars, and it was even possible to divide some of the variable-star types into subgroups according to their population subdivision. The RR Lyrae variables of type ab, for example, could be separated into different groups by their spectral classifications and their mean periods. Those with mean periods longer than 0.4 days were classified as halo Population II, while those with periods less than 0.4 days were placed in the “disk population.” Similarly, long-period variables were divided into different subgroups, such that those with periods of less than 250 days and of relatively early spectral type (earlier than M5e) were considered “intermediate Population II,” whereas the longer period variables fell into the “older Population I” category. As dynamical properties were more thoroughly investigated, many astronomers divided the Galaxy’s stellar populations into a “thin disk,” a “thick disk,” and a “halo.”

An understanding of the physical differences in the stellar populations became increasingly clearer during the 1950s with improved calculations of stellar evolution. Evolving-star models showed that giants and supergiants are evolved objects recently derived from the main sequence after the exhaustion of hydrogen in the stellar core. As this became better understood, it was found that the luminosity of such giants was not only a function of the masses of the initial main-sequence stars from which they evolved but was also dependent on the chemical composition of the stellar atmosphere. Therefore, not only was the existence of giants in the different stellar populations understood, but differences between the giants with relation to the main sequence of star groups came to be understood in terms of the chemistry of the stars.

At the same time, progress was made in determining the abundances of stars of the different population types by means of high-dispersion spectra obtained with large reflecting telescopes having a coudé focus arrangement. A curve of growth analysis demonstrated beyond a doubt that the two population types exhibited very different chemistries. In 1959 H. Lawrence Helfer, George Wallerstein, and Jesse L. Greenstein of the United States showed that the giant stars in globular clusters have chemical abundances quite different from those of Population I stars such as typified by the Sun. Population II stars have considerably lower abundances of the heavy elements—by amounts ranging from a factor of 5 or 10 up to a factor of several hundred. The total abundance of heavy elements, Z, for typical Population I stars is 0.04 (given in terms of the mass percent for all elements with atomic weights heavier than helium, a common practice in calculating stellar models). The values of Z for halo population globular clusters, on the other hand, were typically as small as 0.003.

A further difference between the two populations became clear as the study of stellar evolution advanced. It was found that Population II was exclusively made up of stars that are very old. Estimates of the age of Population II stars have varied over the years, depending on the degree of sophistication of the calculated models and the manner in which observations for globular clusters are fitted to these models. They have ranged from 10^9 years up to 2×10^{10} years. Recent comparisons of these data suggest that the halo globular clusters have ages of approximately $1.1–1.3 \times 10^{10}$ years. The work of American astronomer Allan Sandage and his collaborators proved without a doubt that the range in age for globular clusters was relatively small and that the detailed characteristics of the giant branches of their colour-magnitude diagrams were correlated with age and small differences in chemical abundances. On the other hand, stars of Population I were found to have a wide range of ages. Stellar associations and galactic clusters with bright blue main-sequence stars have ages of a few million years (stars are still in the process of forming in some of them) to a few hundred million years. Studies of the stars nearest the Sun indicate a mixture of ages with a considerable number of stars of great age—on the order of 10^9 years. Careful searches, however, have shown that there are no stars in the solar neighbourhood and no galactic clusters whatsoever that are older than the globular clusters. This is an indication that globular clusters, and thus Population II objects, formed first in the Galaxy and that Population I stars have been forming since.

In short, as the understanding of stellar populations grew, the division into Population I and Population II became understood in terms of three parameters: Age, chemical composition, and kinematics. A fourth parameter, spatial distribution, appeared to be clearly another manifestation of kinematics. The correlations between these three parameters were not perfect but seemed to be reasonably good for the Galaxy, even though it was not yet known whether these correlations were applicable to other galaxies. As various types of galaxies were explored more completely, it became clear that the mix of populations in galaxies was correlated with Hubble type. Spiral galaxies such as the Milky Way Galaxy have Population I concentrated in the spiral disk and Population II spread out in a thick disk and/or a spherical halo. Elliptical galaxies are nearly pure Population II, while irregular galaxies are dominated by a thick disk of Population I, with only a small number of Population II stars. Furthermore, the populations vary with galaxy mass; while the Milky Way Galaxy, a massive example of a spiral galaxy, contains no stars of young age and a low heavy-metal abundance, low-mass galaxies, such as the dwarf irregulars, contain young, low heavy-element stars, as the buildup of heavy elements in stars has not proceeded far in such small galaxies.

The Stellar Luminosity Function

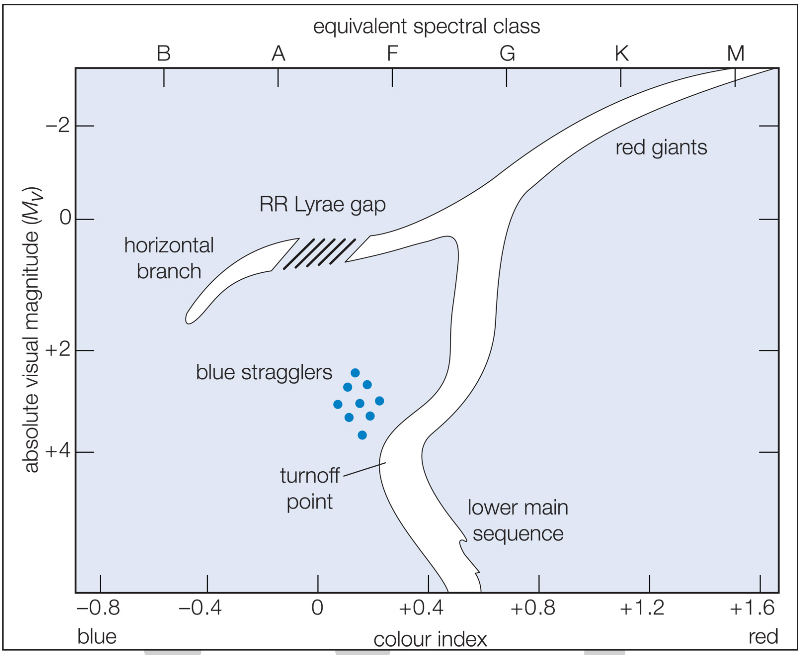
The stellar luminosity function is a description of the relative number of stars of different absolute luminosities. It is often used to describe the stellar content of various parts of the Galaxy or other groups of stars, but it most commonly refers to the absolute number of stars of different absolute magnitudes in the solar neighbourhood. In this form it is usually called the van Rhijn function, named after the Dutch astronomer Pieter J. van Rhijn. The van Rhijn function is a basic datum for the local portion of the Galaxy, but it is not necessarily representative for an area larger than the immediate solar neighbourhood. Investigators have found that elsewhere in the Galaxy, and in the external galaxies (as well as in star clusters), the form of the luminosity function differs in various respects from the van Rhijn function.

The detailed determination of the luminosity function of the solar neighbourhood is an extremely complicated process. Difficulties arise because of (1) the incompleteness of existing surveys of stars of all luminosities in any sample of space and (2) the uncertainties in the basic data (distances and magnitudes). In determining the van Rhijn function, it is normally preferable to specify exactly what volume of space is being sampled and to state explicitly the way in which problems of incompleteness and data uncertainties are handled.

In general there are four different methods for determining the local luminosity function. Most commonly, trigonometric parallaxes are employed as the basic sample. Alternative but somewhat less certain methods include the use of spectroscopic parallaxes, which can involve much larger volumes of space. A third method entails the use of mean parallaxes of a star of a given proper motion and apparent magnitude; this yields a statistical sample of stars of approximately known and uniform distance. The fourth method involves examining the distribution of proper motions and tangential velocities (the speeds at which stellar objects move at right angles to the line of sight) of stars near the Sun.

Because the solar neighbourhood is a mixture of stars of various ages and different types, it is difficult to interpret the van Rhijn function in physical terms without recourse to other sources of information, such as the study of star clusters of various types, ages, and dynamical families. Globular clusters are the best samples to use for determining the luminosity function of old stars having a low abundance of heavy elements (Population II stars).

Globular-cluster luminosity functions show a conspicuous peak at absolute magnitude $M_V = 0.5$, and this is clearly due to the enrichment of stars at that magnitude from the horizontal branch of the cluster. The height of this peak in the data is related to the richness of the horizontal branch, which is in turn related to the age and chemical composition of the stars in the cluster. A comparison of the observed M3 luminosity function with the van Rhijn function shows a depletion of stars, relative to fainter stars, for absolute magnitudes brighter than roughly $M_V = 3.5$. This discrepancy is important in the discussion of the physical significance of the van Rhijn function and luminosity functions for clusters of different ages.



Colour-magnitude (Hertzsprung-Russell) diagram for an old globular cluster made up of Population II stars.

Many studies of the component stars of open clusters have shown that the luminosity functions of these objects vary widely. The two most conspicuous differences are the overabundance of stars of brighter absolute luminosities and the underabundance or absence of stars of faint absolute luminosities. The overabundance at the bright end is clearly related to the age of the cluster (as determined from the main-sequence turnoff point) in the sense that younger star clusters have more of the highly luminous stars. This is completely understandable in terms of the evolution of the clusters and can be accounted for in detail by calculations of the rate of evolution of stars of different absolute magnitudes and mass. For example, the luminosity function for the young clusters h and χ Persei, when compared with the van Rhijn function, clearly shows a large overabundance of bright stars due to the extremely young age of the cluster, which is on the order of 10^6 years. Calculations of stellar evolution indicate that in an additional 10^9 or 10^{10} years all of these stars will have evolved away and disappeared from the bright end of the luminosity function.

In 1955 the first detailed attempt to interpret the shape of the general van Rhijn luminosity function was made by the Austrian-born American astronomer Edwin E.

Salpeter, who pointed out that the change in slope of this function near $M_V = +3.5$ is most likely the result of the depletion of the stars brighter than this limit. Salpeter noted that this particular absolute luminosity is very close to the turnoff point of the main sequence for stars of an age equal to the oldest in the solar neighbourhood—approximately 10^{10} years. Thus, all stars of the luminosity function with fainter absolute magnitudes have not suffered depletion of their numbers because of stellar evolution, as there has not been enough time for them to have evolved from the main sequence.

On the other hand, the ranks of stars of brighter absolute luminosity have been variously depleted by evolution, and so the form of the luminosity function in this range is a composite curve contributed by stars of ages ranging from 0 to 10^{10} years. Salpeter hypothesized that there might exist a time-independent function, the so-called formation function, which would describe the general initial distribution of luminosities, taking into account all stars at the time of formation. Then, by assuming that the rate of star formation in the solar neighbourhood has been uniform since the beginning of this process and by using available calculations of the rate of evolution of stars of different masses and luminosities, he showed that it is possible to apply a correction to the van Rhijn function in order to obtain the form of the initial luminosity function. Comparisons of open clusters of various ages have shown that these clusters agree much more closely with the initial formation function than with the van Rhijn function; this is especially true for the very young clusters. Consequently, investigators believe that the formation function, as derived by Salpeter, is a reasonable representation of the distribution of star luminosities at the time of formation, even though they are not certain that the assumption of a uniform rate of formation of stars can be precisely true or that the rate is uniform throughout a galaxy.

It was stated above that open-cluster luminosity functions show two discrepancies when compared with the van Rhijn function. The first is due to the evolution of stars from the bright end of the luminosity function such that young clusters have too many stars of high luminosity, as compared with the solar neighbourhood. The second discrepancy is that very old clusters such as the globular clusters have too few high-luminosity stars, as compared with the van Rhijn function, and this is clearly the result of stellar evolution away from the main sequence. Stars do not, however, disappear completely from the luminosity function; most become white dwarfs and reappear at the faint end. In his early comparisons of formation functions with luminosity functions of galactic clusters, Sandage calculated the number of white dwarfs expected in various clusters; present searches for these objects in a few of the clusters (e.g., the Hyades) have supported his conclusions.

Open clusters also disagree with the van Rhijn function at the faint end—i.e., for absolute magnitudes fainter than approximately $M_V = +6$. In all likelihood this is mainly due to a depletion of another sort, the result of dynamical effects on the clusters that arise because of internal and external forces. Stars of low mass in such clusters escape from the system under certain common conditions. The formation functions for these clusters may be different from the Salpeter function and may exclude faint stars. A further effect is the result of the finite amount of time it takes for stars to condense; very young clusters have few faint stars partly because there has not been sufficient time for them to have reached their main-sequence luminosity.

Density Distribution

The Stellar Density near the Sun

The density distribution of stars near the Sun can be used to calculate the mass density of material (in the form of stars) at the Sun's distance within the Galaxy. It is therefore of interest not only from the point of view of stellar statistics but also in relation to galactic dynamics. In principle, the density distribution can be calculated by integrating the stellar luminosity function. In practice, because of uncertainties in the luminosity function at the faint end and because of variations at the bright end, the local density distribution is not simply derived nor is there agreement between different studies in the final result.

In the vicinity of the Sun, stellar density can be determined from the various surveys of nearby stars and from estimates of their completeness. For example, the RECONS (Research Consortium on Nearby Stars) has sought all-stars within 10 parsecs of the Sun and found a density in the solar neighbourhood of about 0.003 star per cubic light-year.

The density distribution of stars can be combined with the luminosity-mass relationship to obtain the mass density in the solar neighbourhood, which includes only stars and not interstellar material. This mass density is about 0.001 solar mass per cubic light-year.

Density Distribution of Various Types of Stars

To examine what kinds of stars contribute to the overall density distribution in the immediate solar neighbourhood, various statistical sampling arguments can be applied to catalogs and lists of stars. The result of such a procedure is summarized in the table, which lists some of the kinds of objects and gives the calculated mean density over an appropriate volume centred on the Sun. Note that the table are given in terms of number density.

Table: Space densities of stars.

Object	Density (solar mass per cubic light-year)
O, B stars	0.00003
A, F stars	0.0001
dG, dK stars	0.0004
dM stars	0.0008

gG, gK stars	0.00003
gM stars	0.0000003
dark companions	0.00014
white dwarfs	0.0002
long-period variables	0.00000003
RR Lyrae stars	0.0000000003
Cepheids	0.00000003
planetary nebulae	0.00000000015
open clusters	0.0000011
globular clusters	0.00000003

The most common stars and those that contribute the most to the local stellar mass density are the red dwarf M (dM) stars, which provide a total of 0.0026 star per cubic light-year. White dwarf stars, which are difficult to observe and of which very few are known, are among the more significant contributors.

Variations in the Stellar Density

The star density in the wider solar neighbourhood beyond 10 parsecs is not perfectly uniform. The most conspicuous variations occur in the z direction, above and below the plane of the Galaxy, where the number density falls off rapidly. The more difficult problem of variations within the plane is dealt with here.

Density variations are conspicuous for early-type stars (i.e., stars of higher temperatures) even after allowance has been made for interstellar absorption. For the stars earlier than type B3, for example, large stellar groupings in which the density is abnormally high are conspicuous in several galactic longitudes. The Sun in fact appears to be in a somewhat lower density region than the immediate surroundings, where early B stars are relatively scarce. There is a conspicuous grouping of stars, sometimes called the Cassiopeia-Taurus association that has a centroid at approximately 600 light-years distance. A deficiency of early-type stars is readily noticeable, for instance, in the direction of the constellation Perseus at distances beyond 600 light-years. Of course, the nearby stellar associations are striking density anomalies for early-type stars in the solar neighbourhood. The early-type stars within 2,000 light-years are significantly concentrated at negative galactic latitudes. This is a manifestation of a phenomenon referred to as "the Gould Belt," a tilt of the nearby bright stars in this direction with respect to the galactic plane, which was first noted by the English astronomer John Herschel in 1847. Such anomalous behaviour is true only for the immediate neighbourhood of the Sun; faint B stars are strictly concentrated along the galactic equator.

Generally speaking, the large variations in stellar density near the Sun are less conspicuous for the late-type dwarf stars (those of lower temperatures) than for the earlier types. This fact is explained as the result of the mixing of stellar orbits over long time intervals available for the older stars, which are primarily those stars of later spectral types. The young stars (O, B, and A types) are still close to the areas of star formation and show a common motion and common concentration due to initial formation distributions. In this connection it is interesting to note that the concentration of A-type stars at galactic longitudes 160° to 210° is coincident with a similar concentration of hydrogen detected by means of 21-cm line radiation. Correlations between densities of early-type stars on the one hand and interstellar hydrogen on the other are conspicuous but not fixed; there are areas where neutral-hydrogen concentrations exist but for which no anomalous star density is found.

Variation of Star Density with z Distances

For all stars, variation of star density above and below the galactic plane rapidly decreases with height. Stars of different types, however, exhibit widely differing behaviour in this respect, and this tendency is one of the important clues as to the kinds of stars that occur in different stellar populations.

The luminosity function of stars is different at different galactic latitudes, and this is still another phenomenon connected with the z distribution of stars of different types. At a height of z = 3,000 light-years, stars of absolute magnitude 13 and fainter are nearly as abundant as at the galactic plane, while stars with absolute magnitude 0 are depleted by a factor of 100.

The values of the scale height for various kinds of objects form the basis for the segregation of these objects into different population types. Such objects as open clusters and Cepheid variables that have very small values of the scale height are the objects most restricted to the plane of the Galaxy, while globular clusters and other extreme Population II objects have scale heights of thousands of parsecs, indicating little or no concentration at the plane. Such data and the variation of star density with z distance bear on the mixture of stellar orbit types. They show the range from those stars having nearly circular orbits that are strictly limited to a very flat volume centred at the galactic plane to stars with highly elliptical orbits that are not restricted to the plane.

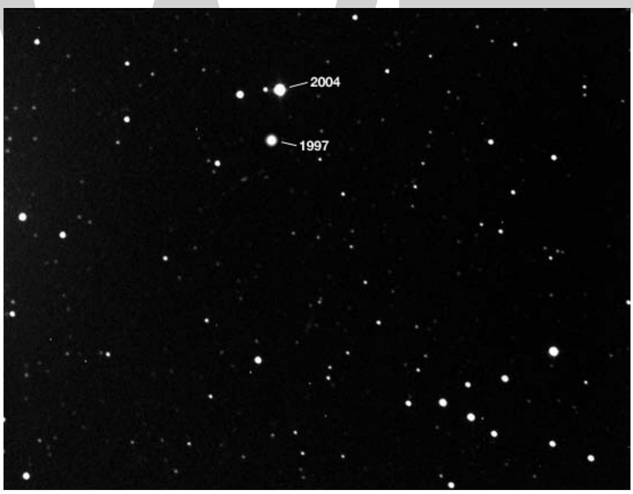
Stellar Motions

A complete knowledge of a star's motion in space is possible only when both its proper motion and radial velocity can be measured. Proper motion is the motion of a star across an observer's line of sight and constitutes the rate at which the direction of the star changes in the celestial sphere. It is usually measured in seconds of arc per year. (One degree equals 3,600 seconds of arc). Radial velocity is the motion of a star along the line of sight and as such is the speed with which the star approaches or recedes from the observer. It is expressed in kilometres per second and is given as either a positive or negative figure, depending on whether the star is moving away from or toward the observer.

Astronomers are able to measure both the proper motions and radial velocities of stars lying near the Sun. They can, however, determine only the radial velocities of stellar objects in more distant parts of the Galaxy and so must use these data, along with the information gleaned from the local sample of nearby stars, to ascertain the large-scale motions of stars in the Milky Way system.

Proper Motions

The proper motions of the stars in the immediate neighbourhood of the Sun are usually very large, as compared with those of most other stars. Those of stars within 17 light-years of the Sun, for instance, range from 0.44 to 10.36 arc seconds per year. The latter value is that of Barnard's star, which is the star with the largest known proper motion. The tangential velocity of Barnard's star is 90 km/sec, and, from its radial velocity (−110.5 km/sec) and distance (6 light-years), astronomers have found that its space velocity (total velocity with respect to the Sun) is 143 km/sec. The distance to this star is rapidly decreasing; it will reach a minimum value of 3.5 light-years about the year 11,800.



Movement of Barnard's star.

Radial Velocities

Radial velocities, measured along the line of sight spectroscopically using the Doppler effect, are known for nearly all of the recognized stars near the Sun. Of the 54 systems within 17 light-years, most have well-determined radial velocities. The radial velocities of the rest are not known, either because of faintness or because of problems resulting from the nature of their spectrum. For example, radial velocities of white dwarfs are often very difficult to obtain because of the extremely broad and faint spectral lines in some of these objects. Moreover, the radial velocities that are determined for such stars are subject to further complication because a gravitational redshift generally affects the positions of their spectral lines. The average gravitational redshift for white dwarfs has been shown to be the equivalent of a velocity of −51 km/sec. To study the true motions of these objects, it is necessary to make such a correction to the observed shifts of their spectral lines.

For nearby stars, radial velocities are with very few exceptions rather small. For stars closer than 17 light-years, radial velocities range from −85 km/sec to +263 km/sec. Most values are on the order of ±20 km/sec, with a mean value of 2 km/sec.

Space Motions

Space motions comprise a three-dimensional determination of stellar motion. They may be divided into a set of components related to directions in the Galaxy: U, directed away from the galactic centre; V, in the direction of galactic rotation; and W, toward the north galactic pole. For the nearby stars the average values for these galactic components are as follows: U = −8 km/sec, V = −28 km/sec, and W = −12 km/sec. These values are fairly similar to those for the galactic circular velocity components, which give U = −9 km/sec, V = −12 km/sec, and W = −7 km/sec. Note that the largest difference between these two sets of values is for the average V, which shows an excess of 16 km/sec for the nearby stars as compared with the circular velocity. Since V is the velocity in the direction of galactic rotation, this can be understood as resulting from the presence of stars in the local neighbourhood that have significantly elliptical orbits for which the apparent velocity in this direction is much less than the circular velocity. This fact was noted long before the kinematics of the Galaxy was understood and is referred to as the asymmetry of stellar motion.

The average components of the velocities of the local stellar neighbourhood also can be used to demonstrate the so-called stream motion. Calculations based on the Dutch-born American astronomer Peter van de Kamp's table of stars within 17 light-years, excluding the star of greatest anomalous velocity, reveal that dispersions in the V direction and the W direction are approximately half the size of the dispersion in the U direction. This is an indication of a commonality of motion for the nearby stars; i.e., these stars are not moving entirely at random but show a preferential direction of motion—the stream motion—confined somewhat to the galactic plane and to the direction of galactic rotation.

High-velocity Stars

One of the nearest 45 stars, called Kapteyn's star, is an example of the high-velocity stars that lie near the Sun. Its observed radial velocity is −245 km/sec, and the components of its space velocity are U = 19 km/sec, V = −288 km/sec, and W = −52 km/sec. The very large value for V indicates that, with respect to circular velocity, this star has practically no motion in the direction of galactic rotation at all. As the Sun's motion in its orbit around the Galaxy is estimated to be approximately 250 km/sec in this direction, the value V of −288 km/sec is primarily just a reflection of the solar orbital motion.

Solar Motion

Solar motion is defined as the calculated motion of the Sun with respect to a specified reference frame. In practice, calculations of solar motion provide information not only on the Sun's motion with respect to its neighbours in the Galaxy but also on the kinematic properties of various kinds of stars within the system. These properties in turn can be used to deduce information on the dynamical history of the Galaxy and of its stellar components. Solutions for solar motion involving many stars of a given class are the prime source of information on the patterns of motion for that class. Furthermore, astronomers obtain information on the large-scale motions of galaxies in the neighbourhood of the Galaxy from solar motion solutions because it is necessary to know the space motion of the Sun with respect to the centre of the Galaxy (its orbital motion) before such velocities can be calculated.

The Sun's motion can be calculated by reference to any of three stellar motion elements: (1) the radial velocities of stars, (2) the proper motions of stars, or (3) the space motions of stars.

Solar Motion Calculations from Radial Velocities

For objects beyond the immediate neighbourhood of the Sun, initially it is necessary to choose a standard of rest (the reference frame) from which the solar motion is to be calculated. This is usually done by selecting a particular kind of star or a portion of space. To solve for solar motion, two assumptions are made. The first is that the stars that form the standard of rest are symmetrically distributed over the sky, and the second is that the peculiar motions—the motions of individual stars with respect to that standard of rest—are randomly distributed. Considering the geometry then provides a mathematical solution for the motion of the Sun through the average rest frame of the stars being considered.

In astronomical literature where solar motion solutions are published, there is often employed a "K-term," a term that is added to the equations to account for systematic errors, the stream motions of stars, or the expansion or contraction of the member stars of the reference frame. Recent determinations of solar motion from high-dispersion radial velocities have suggested that most previous K-terms (which averaged a few kilometres per second) were the result of systematic errors in stellar spectra caused by blends of spectral lines. Of course, the K-term that arises when a solution for solar motions is calculated for galaxies results from the expansion of the system of galaxies and is very large if galaxies at great distances from the Milky Way Galaxy are included.

Solar Motion Calculations from Proper Motions

Solutions for solar motion based on the proper motions of the stars in proper mo-tion catalogs can be carried out even when the distances are not known and the radial velocities are not given. It is necessary to consider groups of stars of limited dispersion in distance so as to have a well-defined and reasonably spatially-uniform reference frame. This can be accomplished by limiting the selection of stars according to their apparent magnitudes. The procedure is the same as the above except that the proper motion components are used instead of the radial velocities. The average distance of the stars of the reference frame enters into the solution of these equations and is related to the term often referred to as the secular parallax. The secular parallax is defined as 0.24h/r, where h is the solar motion in astronomical units per year and r is the mean distance for the solar motion solution.

Solar Motion Calculations from Space Motions

For nearby well-observed stars, it is possible to determine complete space motions and to use these for calculating the solar motion. One must have six quantities: α (the right ascension of the star); δ (the declination of the star); μ_α (the proper motion in right ascension); μ_δ (the proper motion in declination); ρ (the radial velocity as reduced to the Sun); and r (the distance of the star). To find the solar motion, one calculates the velocity components of each star of the sample and the averages of all of these.

Solar motion solutions give values for the Sun's motion in terms of velocity components, which are normally reduced to a single velocity and a direction. The direction in which the Sun is apparently moving with respect to the reference frame is called the apex of solar motion. In addition, the calculation of the solar motion provides dispersion in velocity. Such dispersions are as intrinsically interesting as the solar motions themselves because dispersion is an indication of the integrity of the selection of stars used as a reference frame and of its uniformity of kinematic properties. It is found, for example, that dispersions are very small for certain kinds of stars (e.g., A-type stars, all of which apparently have nearly similar, almost circular orbits in the Galaxy) and are very large for some other kinds of objects (e.g., the RR Lyrae variables, which show a dispersion of almost 100 km/sec due to the wide variation in the shapes and orientations of orbits for these stars).

Solar Motion Solutions

The motion of the Sun with respect to the nearest common stars is of primary interest. If stars within about 80 light-years of the Sun are used exclusively, the result is often called the standard solar motion. This average, taken for all kinds of stars, leads to a velocity $V_{\odot}$ = 19.5 km/sec. The apex of this solar motion is in the direction of α = 270°, δ = +30°. The exact values depend on the selection of data and method of solution. These values suggest that the Sun's motion with respect to its neighbours is moderate but certainly not zero. The velocity difference is larger than the velocity dispersions for common stars of the earlier spectral types, but it is very similar in value to the dispersion for stars of a spectral type similar to the Sun. The solar velocity for, say, G5 stars is 10 km/sec, and the dispersion is 21 km/sec. Thus, the Sun's motion can be considered fairly typical for its class in its neighbourhood. The peculiar motion of the Sun is a result of its relatively large age and a somewhat noncircular orbit. It is generally true that stars of later spectral types show both greater dispersions and greater values for solar motion, and this characteristic is interpreted to be the result of a mixture of orbital properties for the later spectral types, with increasingly large numbers of stars having more highly elliptical orbits.

The term basic solar motion has been used by some astronomers to define the motion of the Sun relative to stars moving in its neighbourhood in perfectly circular orbits around the galactic centre. The basic solar motion differs from the standard solar motion because of the noncircular motion of the Sun and because of the contamination of the local population of stars by the presence of older stars in noncircular orbits within the limits of the reference frame. The most commonly quoted value for the basic solar motion is a velocity of 16.5 km/sec toward an apex with a position $\alpha = 265°$, $\delta = 25°$.

When the solutions for solar motion are determined according to the spectral class of the stars, there is a correlation between the result and the spectral class. The table summarizes values obtained from various sources and illustrates this fact. The apex of the solar motion, the solar motion velocity, and its dispersion are all correlated with spectral type. Generally speaking (with the exception of the very early type stars), the solar motion velocity increases with decreasing temperature of the stars, ranging from 16 km/sec for late B-type and early A-type stars to 24 km/sec for late K-type and early M-type stars. The dispersion similarly increases from a value near 10 km/sec to a value of 22 km/sec. The reason for this is related to the dynamical history of the Galaxy and the mean age and mixture of ages for stars of the different spectral types. It is quite clear, for example, that stars of early spectral type are all young, whereas stars of late spectral types are a mixture of young and old. Connected with this is the fact that the solar motion apex shows a trend for the latitude to decrease and the longitude to increase with later spectral types.

The solar motion can be based on reference frames defined by various kinds of stars and clusters of astrophysical interest. Data of this sort are interesting because of the way in which they make it possible to distinguish between objects with different kinematic properties in the Galaxy. For example, it is clear that interstellar calcium lines have relatively small solar motion and extremely small dispersion because they are primarily connected with the dust that is limited to the galactic plane and with objects that are decidedly of the Population I class. On the other hand, RR Lyrae variables and globular clusters have very large values of solar motion and very large dispersions, indicating that they are extreme Population II object that do not all equally share in the rotational motion of the Galaxy. The solar motion of these various objects is an important consideration in determining to what population the objects belong and what their kinematic history has been.

When some of these classes of objects are examined in greater detail, it is possible to separate them into subgroups and find correlations with other astrophysical properties. Take, for example, globular clusters, for which the solar motion is correlated with the spectral type of the clusters. The clusters of spectral types G0–G5 (the more metal-rich clusters) have a mean solar motion of 80 ± 82 km/sec (corrected for the standard solar motion). The earlier globular clusters of types F2–F9, on the other hand, have a mean velocity of 162 ± 36 km/sec, suggesting that they partake much less extensively in the general rotation of the Galaxy. Similarly, the most distant globular clusters have a larger solar motion than the ones closer to the galactic centre. Studies of RR Lyrae variables also show correlations of this sort. The period of an RR Lyrae variable, for example, is correlated with its motion with respect to the Sun. For type ab RR Lyrae variables, periods frequently vary from 0.3 to 0.7 days, and the range of solar motion for this range of period extends from 30 to 205 km/sec, respectively. This condition is believed to be primarily the result of the effects of the spread in age and composition for the RR Lyrae variables in the field, which is similar to, but larger than, the spread in the properties of the globular clusters.

Since the direction of the centre of the Galaxy is well established by radio measurements and since the galactic plane is clearly established by both radio and optical studies, it is possible to determine the motion of the Sun with respect to a fixed frame of reference centred at the Galaxy and not rotating (i.e., tied to the external galaxies). The value for this motion is generally accepted to be 225 km/sec in the direction $\ell^{II} = 90°$. It is not a firmly established number, but it is used by convention in most studies.

In order to arrive at a clear idea of the Sun's motion in the Galaxy as well as of the motion of the Galaxy with respect to neighbouring systems, solar motion has been studied with respect to the Local Group galaxies and those in nearby space. Hubble determined the Sun's motion with respect to the galaxies beyond the Local Group and found the value of 300 km/sec in the direction toward galactic longitude 120°, latitude +35°. This velocity includes the Sun's motion in relation to its proper circular velocity, its circular velocity around the galactic centre, the motion of the Galaxy with respect to the Local Group, and the latter's motion with respect to its neighbours.

The Structure and Dynamics of the Milky Way Galaxy

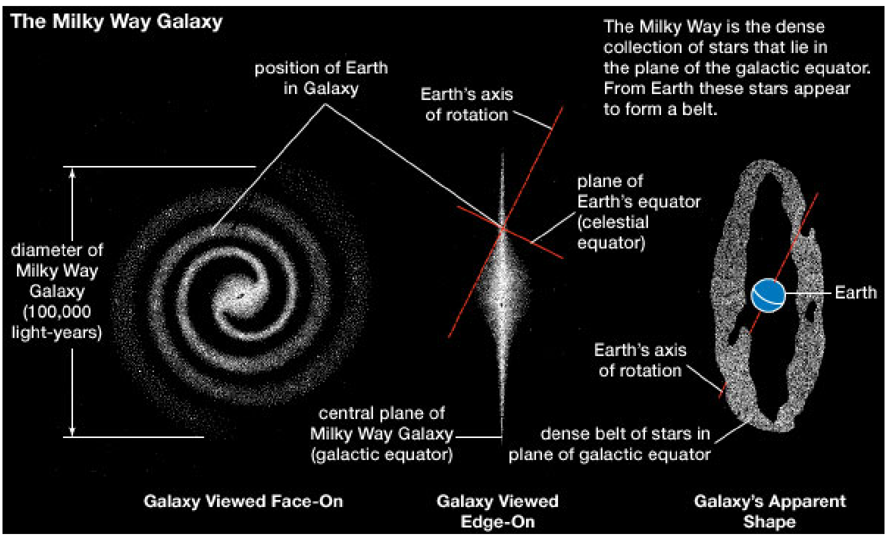
Size

The first reliable measurement of the size of the Galaxy was made in 1917 by American astronomer Harlow Shapley. He arrived at his size determination by establishing the spatial distribution of globular clusters. Shapley found that, instead of a relatively small system with the Sun near its centre, as had previously been thought, the Galaxy is immense, with the Sun nearer the edge than the centre. Assuming that the globular clusters outlined the Galaxy, he determined that it has a diameter of about 100,000 light-years and that the Sun lies about 30,000 light-years from the centre. (A light-year is the distance traveled by light in one year and is roughly 9,460,000,000,000 km [5,880,000,000,000 miles]). His values have held up remarkably well over the years. Depending in part on the particular component being discussed, the stellar disk of the

Milky Way system is just about as large as Shapley's model predicted, with neutral hydrogen somewhat more widely dispersed and dark (i.e., unobservable) matter perhaps filling an even larger volume than expected. The most-distant stars and gas clouds of the system that have had their distance reliably determined lie roughly 100,000 light-years from the galactic centre, while the distance of the Sun from the centre has been found to be approximately 25,000 light-years.

Structure of the Spiral System

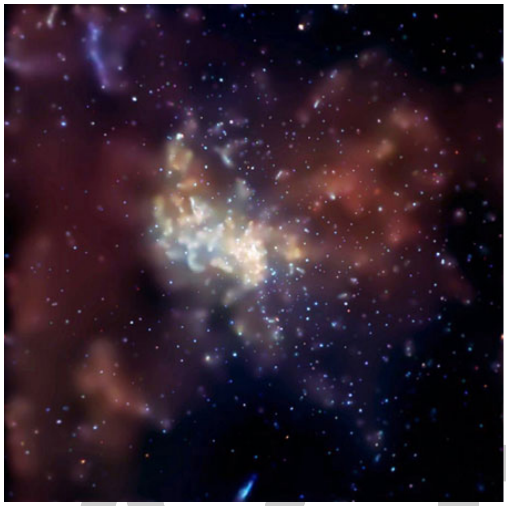
The Milky Way Galaxy's structure is fairly typical of a large spiral system. This structure can be viewed as consisting of six separate parts: (1) a nucleus, (2) a central bulge, (3) a disk (both a thin and a thick disk), (4) spiral arms, (5) a spherical component, and (6) a massive halo. Some of these components blend into each other.



Three views of the Milky Way Galaxy.

The Nucleus

At the very centre of the Galaxy lies a remarkable object—in all likelihood a massive black hole surrounded by an accretion disk of high-temperature gas. Neither the central object nor any of the material immediately around it can be observed at optical wavelengths because of the thick screen of intervening dust in the Milky Way. The object, however, is readily detectable at radio wavelengths and has been dubbed Sagittarius A* by radio astronomers. Somewhat similar to the centres of active galaxies, though on a lesser scale, the galactic nucleus is the site of a wide range of activity apparently powered by the black hole. Infrared radiation and X-rays are emitted from the area, and rapidly moving gas clouds can be observed there. Data strongly indicate that material is being pulled into the black hole from outside the nuclear region, including some gas from the z direction (i.e., perpendicular to the galactic plane). As the gas nears the black hole, the central object's strong gravitational force squeezes the gas into a rapidly rotating disk, which extends outward about 5–30 light-years from the black hole. Rotation measurements of the disk and the orbital motions of stars (seen at infrared wavelengths) indicate that the black hole has a mass 4,000,000 times that of the Sun.



Cosmic radio-wave source Sagittarius A* in an image from the Chandra X-ray Observatory. Sagittarius A* is an extremely bright source within the larger Sagittarius A complex and contains the black hole at the Milky Way Galaxy's centre.

The Central Bulge

Surrounding the nucleus is an extended bulge of stars that is nearly spherical in shape and that consists primarily of Population II stars, though they are comparatively rich in heavy elements. Mixed with the stars are several globular clusters of similar stars, and both the stars and the clusters have nearly radial orbits around the nucleus. The bulge stars can be seen optically where they stick up above the obscuring dust of the galactic plane.

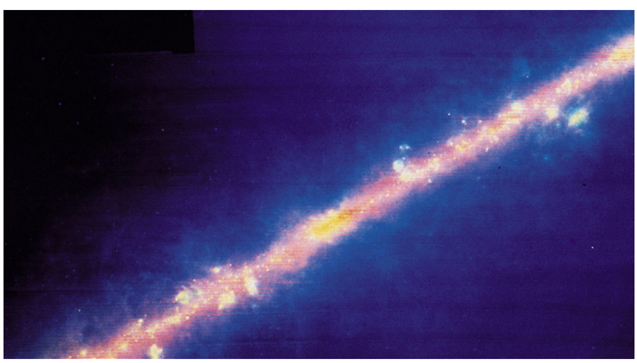


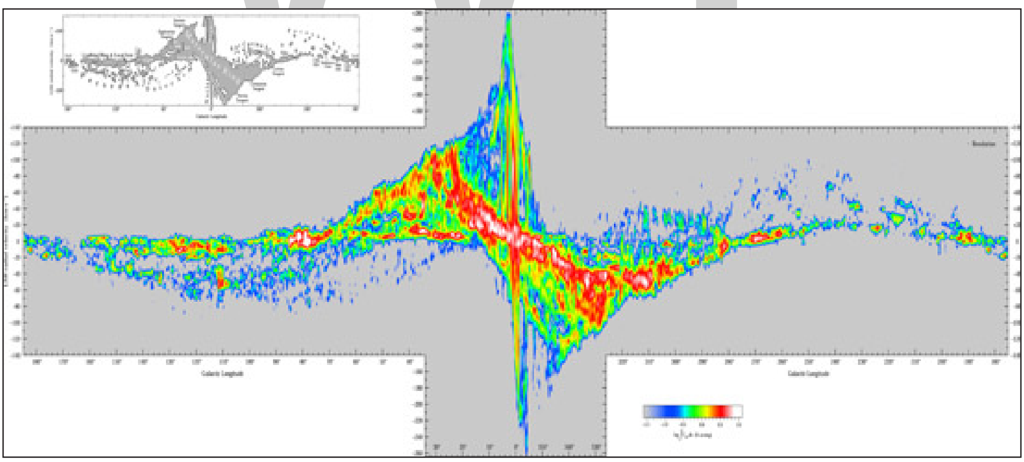
Image of the centre of the Milky Way Galaxy, produced from the observations made by the Infrared Astronomy Satellite (IRAS). The bulge in the band is the centre of the Galaxy. The yellow and green spots and blobs are giant clouds of interstellar gas and dust. The warmest material appears blue and colder material red.

The Disk

From a distance the most conspicuous part of the Galaxy would be the disk, which extends from the nucleus out to approximately 75,000 light-years. The Galaxy resembles other spiral systems, featuring as it does a bright, flat arrangement of stars and gas clouds that is spread out over its entirety and marked by a spiral structure. The disk can be thought of as being the underlying body of stars upon which the arms are superimposed. This body has a thickness that is roughly one-fifth its diameter, but different components have different characteristic thicknesses. The thinnest component, often called the "thin disk," includes the dust and gas and the youngest stars, while a thicker component, the "thick disk," includes somewhat older stars.

The Spiral Arms

Astronomers did not know that the Galaxy had a spiral structure until 1953, when the distances to stellar associations were first obtained reliably. Because of the obscuring interstellar dust and the interior location of the solar system, the spiral structure is very difficult to detect optically. This structure is easier to discern from radio maps of either neutral hydrogen or molecular clouds, since both can be detected through the dust. Distances to the observed neutral hydrogen atoms must be estimated on the basis of measured velocities used in conjunction with a rotation curve for the Galaxy, which can be built up from measurements made at different galactic longitudes.



Longitude-velocity map of the Milky Way Galaxy as shown by spectral line emission of carbon monoxide in molecular clouds. The vertical axis represents velocity and the horizontal axis longitude. The gentle curves in the left and right portions of the map trace the spiral arms of the Milky Way Galaxy. The vertical structure in the middle of the map is the centre of the Galaxy. The emission stretching from the upper left to the lower right in the middle portion of the map is the "molecular ring," a ring of gas and dust in orbit between 4 and 8 kiloparsecs from the centre of the Galaxy.

From studies of other galaxies it can be shown that spiral arms generally follow a logarithmic spiral form such that,

$$\log r = a - b\varphi,$$

where φ is a position angle measured from the centre to the outermost part of the arm, r is the distance from the centre of the galaxy, and a and b are constants. The range in pitch angles for galaxies is from about 50° to approximately 85°. The pitch angle is constant for any given galaxy if it follows a true logarithmic spiral. The pitch angle for the spiral arms of the Galaxy is difficult to determine from the limited optical data, but most measurements indicate a value of about 75°. There are three spiral arms in the part of the Milky Way Galaxy wherein the solar system is located.

Theoretical understanding of the Galaxy's spiral arms has progressed greatly since the 1950s, but there is still no complete understanding of the relative importance of the various effects thought to determine their structure. The overall pattern is almost certainly the result of a general dynamical effect known as a density-wave pattern. The American astronomers Chia-Chiao Lin and Frank H. Shu showed that a spiral shape is a natural result of any large-scale disturbance of the density distribution of stars in a galactic disk. When the interaction of the stars with one another is calculated, it is found that the resulting density distribution takes on a spiral pattern that does not rotate with the stars but rather moves around the nucleus more slowly as a fixed pattern. Individual stars in their orbits pass in and out of the spiral arms, slowing down in the arms temporarily and thereby causing the density enhancement. For the Galaxy, comparison of neutral hydrogen data with the calculations of Lin and Shu have shown that the pattern speed is 4 km/sec per 1,000 light-years.

Other effects that can influence a galaxy's spiral shape have been explored. It has been demonstrated, for example, that a general spiral pattern will result simply from the fact that the galaxy has differential rotation; i.e., the rotation speed is different at different distances from the galactic centre. Any disturbance, such as a sequence of stellar formation events that are sometimes found drawn out in a near-linear pattern, will eventually take on a spiral shape simply because of the differential rotation. For example, the outer spiral structure in some galaxies may be the result of tidal encounters with other galaxies or galactic cannibalism. Distortions that also can be included are the results of massive explosions such as supernova events. These, however, tend to have only fairly local effects.

The Spherical Component

The space above and below the disk of the Galaxy is occupied by a thinly populated extension of the central bulge. Nearly spherical in shape, this region is populated by the outer globular clusters, but it also contains many individual field stars of extreme Population II, such as RR Lyrae variables and dwarf stars deficient in the heavy elements. Structurally, the spherical component resembles an elliptical galaxy, following the same simple mathematical law of how density varies with distance from the centre.

The Massive Halo

The least-understood component of the Galaxy is the giant massive halo that is exterior to the entire visible part. The existence of the massive halo is demonstrated by its effect on the outer rotation curve of the Galaxy. All that can be said with any certainty is that the halo extends considerably beyond a distance of 100,000 light-years from the centre and that its mass is several times greater than the mass of the rest of the Galaxy taken together. It is not known what its shape is, what its constituents are, or how far into intergalactic space it extends.

Magnetic Field

It was once thought that the spiral structure of galaxies might be controlled by a strong magnetic field. However, when the general magnetic field was detected by radio techniques, it was found to be too weak to have large-scale effects on galactic structure. The strength of the galactic field is only about 0.000001 times the strength of Earth's field at its surface, a value that is much too low to have dynamical effects on the interstellar gas that could account for the order represented by the spiral-arm structure. This is, however, sufficient strength to cause a general alignment of the dust grains in interstellar space, a feature that is detected by measurements of the polarization of starlight. In the prevailing model of interstellar dust grains, the particles are shown to be rapidly spinning and to contain small amounts of metal (probably iron), though the primary constituents are ice and carbon. The magnetic field of the Galaxy can gradually act on the dust particles and cause their rotational axes to line up in such a way that their short axes are parallel to the direction of the field. The field itself is aligned along the Milky Way band, so that the short axes of the particles also become aligned along the galactic plane. Polarization measurements of stars at low galactic latitudes confirm this pattern.

Rotation

The motions of stars in the local stellar neighbourhood can be understood in terms of a general population of stars that have circular orbits of rotation around the distant galactic nucleus, with an admixture of stars that have more highly elliptical orbits and that appear to be high-velocity stars to a terrestrial observer as Earth moves with the Sun in its circular orbit. The general rotation of the disk stars was first detected through studies made in the 1920s, notably those of the Swedish astronomer Bertil Lindblad, who correctly interpreted the apparent asymmetries in stellar motions as the result of this multiple nature of stellar orbital characteristics.

The disk component of the Galaxy rotates around the nucleus in a manner similar to the pattern for the planets of the solar system, which has nearly circular orbits around the Sun. Because the rotation rate is different at different distances from the centre of the Galaxy, the measured velocities of disk stars in different directions along the Milky Way exhibit different patterns. The Dutch astronomer Jan H. Oort first interpreted this effect in terms of galactic rotation motions, employing the radial velocities and proper motions of stars. He demonstrated that differential rotation leads to a systematic variation of the radial velocities of stars with galactic longitude following the mathematical expression.

radial velocity = Ar sin 2l,

where A is called Oort's constant and is approximately 15 km/sec/kiloparsec (1 kiloparsec is 3,260 light-years), r is the distance to the star, and l is the galactic longitude.

A similar expression can be derived for measured proper motions of stars. The agreement of observed data with Oort's formulas was a landmark demonstration of the correctness of Lindblad's ideas about stellar motions. It led to the modern understanding of the Galaxy as consisting of a giant rotating disk with other, more spherical, and more slowly rotating components superimposed.

Mass

The total mass of the Galaxy, which had seemed reasonably well established during the 1960s, has become a matter of considerable uncertainty. Measuring the mass out to the distance of the farthest large hydrogen clouds is a relatively straightforward procedure. The measurements required are the velocities and positions of neutral hydrogen gas, combined with the approximation that the gas is rotating in nearly circular orbits around the centre of the Galaxy. A rotation curve, which relates the circular velocity of the gas to its distance from the galactic centre, is constructed. The shape of this curve and its values are determined by the amount of gravitational pull that the Galaxy exerts on the gas. Velocities are low in the central parts of the system because not much mass is interior to the orbit of the gas; most of the Galaxy is exterior to it and does not exert an inward gravitational pull. Velocities are high at intermediate distances because most of the mass in that case is inside the orbit of the gas clouds and the gravitational pull inward is at a maximum. At the farthest distances, the velocities decrease because nearly all the mass is interior to the clouds. This portion of the Galaxy is said to have Keplerian orbits, since the material should move in the same manner that the German astronomer Johannes Kepler discovered the planets to move within the solar system, where virtually all the mass is concentrated inside the orbits of the orbiting bodies. The total mass of the Galaxy is then found by constructing mathematical models of the system with different amounts of material distributed in various ways and by comparing the resulting velocity curves with the observed one. As applied in the 1960s, this procedure indicated that the total mass of the Galaxy was approximately 200 billion times the mass of the Sun.

During the 1980s, however, refinements in the determination of the velocity curve began to cast doubts on the earlier results. The downward trend to lower velocities in the outer parts of the Galaxy was found to have been in error. Instead, the curve remained almost constant, indicating that there continue to be substantial amounts of matter exterior to the measured hydrogen gas. This in turn indicates that there must be some undetected material out there that is completely unexpected. It must extend considerably beyond the previously accepted positions of the edge of the Galaxy, and it must be dark at virtually all wavelengths, as it remains undetected even when searched for with radio, X-ray, ultraviolet, infrared, and optical telescopes. Until the dark matter is identified and its distribution determined, it will be impossible to measure the total mass of the Galaxy, so all that can be said is that the mass is at least several hundred billion, and possibly one trillion, times the mass of the Sun.

The nature of the dark matter in the Galaxy remains one of the major questions of galactic astronomy. Many other galaxies also appear to have such undetected matter. In the 1990s astronomers carried out exhaustive lensing experiments involving the study of millions of stars in the galactic central areas and in the Magellanic Clouds to search for dark objects whose masses would cause lensed brightenings of background stars. Some lensing events were detected, but the number of dark objects inferred is not enough to explain completely the dark matter in galaxies and galaxy clusters. It is likely that the dark matter consists of some undiscovered particle, such as a WIMP (weakly interacting massive particle).

Galactic Center

The Galactic Center, as seen by one of the 2MASS infrared telescopes, is located in the bright upper left portion of the image.

The Galactic Center, or Galactic Centre, is the rotational center of the Milky Way. It is 8,122 ± 31 parsecs (26,490 ± 100 ly) away from Earth in the direction of the constellations Sagittarius, Ophiuchus, and Scorpius where the Milky Way appears brightest. It coincides with the compact radio source Sagittarius A*.

There are around 10 million stars within one parsec of the Galactic Center, dominated by red giants, with a significant population of massive supergiants and Wolf-Rayet stars from a star formation event around one million years ago, and one supermassive black hole of 4.100 ± 0.034 million solar masses at the Galactic Center, which powers the Sagittarius A* radio source.

Discovery

Because of interstellar dust along the line of sight, the Galactic Center cannot be studied at visible, ultraviolet, or soft (low-energy) X-ray wavelengths. The available information about the Galactic Center comes from observations at gamma ray, hard (high-energy) X-ray, infrared, submillimetre, and radio wavelengths.

Immanuel Kant stated in *General Natural History and Theory of the Heavens* that a large star was at the center of the Milky Way Galaxy, and that Sirius might be the star. Harlow Shapley stated in 1918 that the halo of globular clusters surrounding the Milky Way seemed to be centered on the star swarms in the constellation of Sagittarius, but the dark molecular clouds in the area blocked the view for optical astronomy. In the early 1940s Walter Baade at Mount Wilson Observatory took advantage of wartime blackout conditions in nearby Los Angeles to conduct a search for the center with the 100-inch (250 cm) Hooker Telescope. He found that near the star Alnasl (Gamma Sagittarii) there is a one-degree-wide void in the interstellar dust lanes, which provides a relatively clear view of the swarms of stars around the nucleus of our Milky Way Galaxy. This gap has been known as Baade's Window ever since.

At Dover Heights in Sydney, Australia, a team of radio astronomers from the Division of Radiophysics at the CSIRO, led by Joseph Lade Pawsey, used 'sea interferometry' to discover some of the first interstellar and intergalactic radio sources, including Taurus A, Virgo A and Centaurus A. By 1954 they had built an 80-foot (24 m) fixed dish antenna and used it to make a detailed study of an extended, extremely powerful belt of radio emission that was detected in Sagittarius. They named an intense point-source near the center of this belt Sagittarius A, and realised that it was located at the very center of our Galaxy, despite being some 32 degrees south-west of the conjectured galactic center of the time.

In 1958 the International Astronomical Union (IAU) decided to adopt the position of Sagittarius A as the true zero co-ordinate point for the system of galactic latitude and longitude. In the equatorial coordinate system the location is: RA $17^h\ 45^m\ 40.04^s$, Dec $-29°\ 00'\ 28.1''$ (J2000 epoch).

Distance to the Galactic Center

The exact distance between the Solar System and the Galactic Center is not certain, although estimates since 2000 have remained within the range 24–28.4 kilolight-years (7.4–8.7 kiloparsecs). The latest estimates from geometric-based methods and standard candles yield the following distances to the Galactic Center:

- 7.4 ± 0.2(stat) ± 0.2(syst) or 7.4±0.3 kpc (≈24 ± 1 kly).
- 7.62 ± 0.32 kpc (≈24.8 ± 1 kly).
- 7.7 ± 0.7 kpc (≈25.1 ± 2.3 kly).
- 7.94 or 8.0±0.5 kpc (≈26 ± 1.6 kly).
- 7.98 ± 0.15(stat) ± 0.20(syst) or 8.0±0.25 kpc (≈26 ± 0.8 kly).
- 8.33 ± 0.35 kpc (≈27 ± 1.1 kly).
- 8.7 ± 0.5 kpc (≈28.4 ± 1.6 kly).

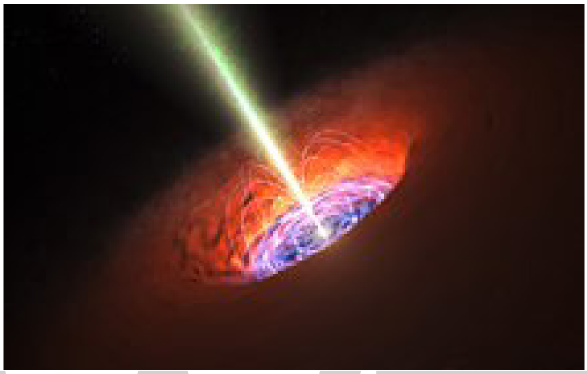
An accurate determination of the distance to the Galactic Center as established from variable stars (e.g. RR Lyrae variables) or standard candles (e.g. red-clump stars) is hindered by countless effects, which include: an ambiguous reddening law; a bias for smaller values of the distance to the Galactic Center because of a preferential sampling of stars toward the near side of the Galactic bulge owing to interstellar extinction; and an uncertainty in characterizing how a mean distance to a group of variable stars found in the direction of the Galactic bulge relates to the distance to the Galactic Center.

The nature of the Milky Way's bar, which extends across the Galactic Center, is also actively debated, with estimates for its half-length and orientation spanning between 1–5 kpc (short or a long bar) and 10–50°. Certain authors advocate that the Milky Way features two distinct bars, one nestled within the other. The bar is delineated by red-clump stars however, RR Lyrae variables do not trace a prominent Galactic bar. The bar may be surrounded by a ring called the *5-kpc ring* that contains a large fraction of the molecular hydrogen present in the Milky Way, and most of the Milky Way's star formation activity. Viewed from the Andromeda Galaxy, it would be the brightest feature of the Milky Way.

Supermassive Black Hole

There is a supermassive black hole in the bright white area to the right of the center of the image. This composite photograph covers about half of a degree.

The complex astronomical radio source Sagittarius A appears to be located almost exactly at the Galactic Center (approx. 18 hrs, −29 deg), and contains an intense compact radio source, Sagittarius A*, which coincides with a supermassive black hole at the center of the Milky Way. Accretion of gas onto the black hole, probably involving an accretion disk around it, would release energy to power the radio source, itself much larger than the black hole. The latter is too small to see with present instruments.



Artist impression of a supermassive black hole at the center of a galaxy.

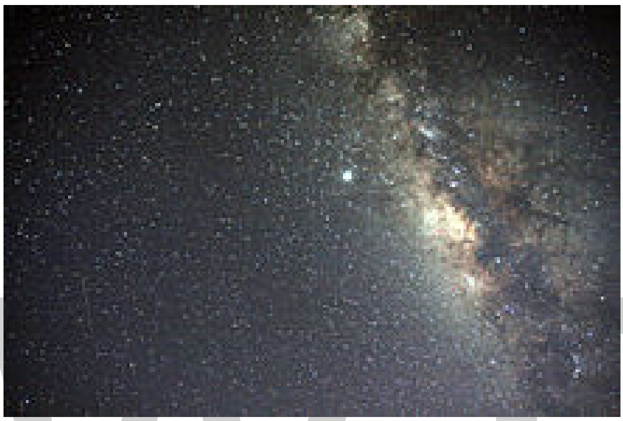
A study in 2008 which linked radio telescopes in Hawaii, Arizona and California (Very Long Baseline Interferometry) measured the diameter of Sagittarius A* to be 44 million kilometers (0.3 AU). For comparison, the radius of Earth's orbit around the Sun is about 150 million kilometers (1.0 AU), whereas the distance of Mercury from the Sun at closest approach (perihelion) is 46 million kilometers (0.3 AU). Thus, the diameter of the radio source is slightly less than the distance from Mercury to the Sun.

Scientists at the Max Planck Institute for Extraterrestrial Physics in Germany using Chilean telescopes have confirmed the existence of a supermassive black hole at the Galactic Center, on the order of 4.3 million solar masses.

On 5 January 2015, NASA reported observing an X-ray flare 400 times brighter than usual, a record-breaker, from Sagittarius A*. The unusual event may have been caused by the breaking apart of an asteroid falling into the black hole or by the entanglement of magnetic field lines within gas flowing into Sagittarius A*, according to astronomers.

Stellar Population

The central cubic parsec around Sagittarius A* contains around 10 million stars. Although most of them are old red giant stars, the Galactic Center is also rich in massive stars. More than 100 OB and Wolf–Rayet stars have been identified there so far. They seem to have all been formed in a single star formation event a few million years ago. The existence of these relatively young stars was a surprise to experts, who expected the tidal forces from the central black hole to prevent their formation. This *paradox of youth* is even stronger for stars that are on very tight orbits around Sagittarius A*, such as S2 and S0-102. The scenarios invoked to explain this formation involve either star formation in a massive star cluster offset from the Galactic Center that would have migrated to its current location once formed, or star formation within a massive, compact gas accretion disk around the central black-hole. Current evidence favors the latter theory, as formation through a large accretion disk is more likely to lead to the observed discrete edge of the young stellar cluster at roughly 0.5 parsec. Most of these 100 young, massive stars seem to be concentrated within one or two disks, rather than randomly distributed within the central parsec. This observation however does not allow definite conclusions to be drawn at this point.



Galactic Center of the Milky Way and a meteor.

Star formation does not seem to be occurring currently at the Galactic Center, although the Circumnuclear Disk of molecular gas that orbits the Galactic Center at two parsecs seems a fairly favorable site for star formation. Work presented in 2002 by Antony Stark and Chris Martin mapping the gas density in a 400-light-year region around the Galactic Center has revealed an accumulating ring with a mass several million times that of the Sun and near the critical density for star formation. They predict that in approximately 200 million years there will be an episode of starburst in the Galactic Center, with many stars forming rapidly and undergoing supernovae at a hundred times the current rate. This starburst may also be accompanied by the formation of galactic relativistic jets as matter falls into the central black hole. It is thought that the Milky Way undergoes a starburst of this sort every 500 million years.

In addition to the paradox of youth, there is also a “conundrum of old age” associated with the distribution of the old stars at the Galactic Center. Theoretical models had predicted that the old stars—which far outnumber young stars—should have a steeply-rising density near the black hole, a so-called Bahcall–Wolf cusp. Instead, it was discovered in 2009 that the density of the old stars peaks at a distance of roughly 0.5 parsec from Sgr A*, then falls inward: instead of a dense cluster, there is a “hole”, or core, around the black hole. Several suggestions have been put forward to explain this puzzling observation, but none is completely satisfactory. For instance, although the black hole would eat stars near it, creating a region of low density, this region would be much smaller than a parsec. Because the observed stars are a fraction of the total number, it is theoretically possible that the overall stellar distribution is different from what is observed, although no plausible models of this sort have been proposed yet.

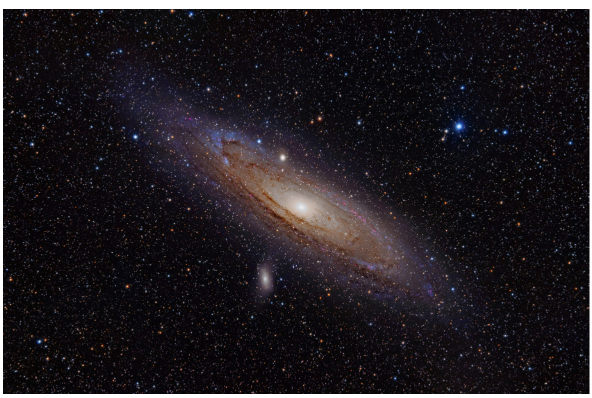
Gamma- and X-ray Emitting Fermi Bubbles

In November 2010, it was announced that two large elliptical lobe structures of energetic plasma, termed “bubbles”, which emit gamma- and X-rays, were detected astride the Milky Way galaxy’s core. These so-called “Fermi bubbles” extend up to about 25,000 light years above and below the galactic center. The galaxy’s diffuse gamma-ray fog hampered prior observations, but the discovery team led by D. Finkbeiner, building on research by G. Dobler, worked around this problem. The 2014 Bruno Rossi Prize went to Tracy Slatyer, Douglas Finkeiner, and Meng Su “for their discovery, in gamma rays, of the large unanticipated Galactic structure called the *Fermi bubbles*”.

The origin of the bubbles is being researched. The bubbles are connected and seemingly coupled, via energy transport, to the galactic core by columnar structures of energetic plasma dubbed “chimneys”.

The Andromeda Galaxy

The Andromeda galaxy, our Milky Way’s closest neighbor, is the most distant object in the sky that you can see with your unaided eye — but only on a clear night from a location with a very dark sky. The galaxy is a beautiful spiral, but one fact you may not be aware of: We’re safe for a few billion years, but Andromeda is headed our way and on a collision course with the Milky Way.



Andromeda’s close proximity to Earth — at only 2.5 million light-years away — makes it a convenient target to observe for extrapolations about other spiral galaxies. In recent years, scientists have done detailed studies of black holes, stars and other objects within the galaxy. This included a stunning mosaic of Andromeda galaxy images taken by the Hubble Space Telescope in 2015.

Location of the Andromeda Galaxy in the Universe

This mosaic of M31 merges 330 individual images taken by the Ultraviolet/Optical Telescope aboard NASA's Swift spacecraft. It is the highest-resolution image of the galaxy ever recorded in the ultraviolet. The image shows a region 200,000 light-years wide and 100,000 light-years high (100 arcminutes by 50 arcminutes).

The visible fuzzy patch of stars stretches about as long as the width of the full moon, and half as wide; only with significant magnification can you tell it stretches six times that length in fullness.

A spiral galaxy like the Milky Way, Andromeda contains a concentrated bulge of matter in the middle, surrounded by a disk of gas, dust, and stars and an immense halo. Though Andromeda contains approximately a trillion stars to the 250 billion in the Milky Way, our galaxy is actually more massive, because it is thought to contain more dark matter.

Collision Course

Artist's conception of the Milkomeda galaxy a trillion years from now.

Andromeda and the Milky Way are heading on a collision course that will alter the structure of the two galaxies forever. The galaxies are rushing closer to one another at about 70 miles per second (112 kilometers per second). Astronomers estimate that Andromeda will collide with the Milky Way in 4 billion years, with the merger concluding 6 billion years from now. By that time, the sun will have swollen into a red giant and swallowed up the terrestrial planets, so Earth will have other things to worry about.

Still, the fresh influx of dust should boost star formation in the new "Milkomeda" galaxy, and the Earthless sun may well leave the Milky Way for good. After a messy phase, where arms project crazily from the combined pair, the two should settle into a smooth elliptical galaxy.

Galaxy collisions are a normal part of the universe's evolution. In fact, both Andromeda and the Milky Way bear signs of having already crashed into other galaxies. Andromeda boasts a large ring of dust in its center, giving it an interesting shape. Astronomers believe this dust may have formed when it swallowed an existing galaxy.

Andromeda–Milky Way Collision

The Andromeda–Milky Way collision is a galactic collision predicted to occur in about 4.5 billion years between two galaxies in the Local Group—the Milky Way (which contains the Solar System and Earth) and the Andromeda Galaxy. The stars involved are sufficiently far apart that it is improbable that any of them will individually collide. Some stars will be ejected from the resulting galaxy, nicknamed *Milkomeda* or *Milkdromeda*.

Certainty

Based on data from the Hubble Space Telescope, the Milky Way Galaxy (pictured right-centre) and Andromeda Galaxy (left-centre) are predicted to distort each other with tidal pull in 3.75 billion years.

The Andromeda Galaxy is approaching the Milky Way at about 110 kilometres per second (68 mi/s) as indicated by blueshift. However, the lateral speed (measured as proper motion) is very difficult to measure with a precision to draw reasonable conclusions: a lateral speed of only 7.7 km/s would mean that the Andromeda Galaxy is moving toward a point 177,800 light-years to the side of the Milky Way ((7.7 km/s)/(110 km/s)

× (2,540,000 ly)), and such a speed over an eight-year timeframe amounts to only 1/3,000th of a Hubble Space Telescope pixel (Hubble's resolution≈0.05 arcsec: (7.7 km/s)/(300,000 km/s)×(8 y)/(2,540,000 ly)×180°/π×3600 = 0.000017 arcsec). Until 2012, it was not known whether the possible collision was definitely going to happen or not. In 2012, researchers concluded that the collision is sure using Hubble to track the motion of stars in Andromeda between 2002 and 2010 with sub-pixel accuracy. Andromeda's tangential or sideways velocity with respect to the Milky Way was found to be much smaller than the speed of approach and therefore it is expected that it will directly collide with the Milky Way in around four and a half billion years.

Such collisions are relatively common, considering galaxies' long lifespans. Andromeda, for example, is believed to have collided with at least one other galaxy in the past, and several dwarf galaxies such as Sgr dSph are currently colliding with the Milky Way and being merged into it.

The studies also suggest that M33, the Triangulum Galaxy—the third-largest and third-brightest galaxy of the Local Group—will participate in the collision event, too. Its most likely fate is to end up orbiting the merger remnant of the Milky Way and Andromeda galaxies and finally to merge with it in an even more distant future. However, a collision with the Milky Way, before it collides with the Andromeda Galaxy, or an ejection from the Local Group cannot be ruled out.

Stellar Collisions

While the Andromeda Galaxy contains about 1 trillion (10^{12}) stars and the Milky Way contains about 300 billion ($3×10^{11}$), the chance of even two stars colliding is negligible because of the huge distances between the stars. For example, the nearest star to the Sun is Proxima Centauri, about 4.2 light-years ($4.0×10^{13}$ km; $2.5×10^{13}$ mi) or 30 million ($3×10^{7}$) solar diameters away.

To visualize that scale, if the Sun were a ping-pong ball, Proxima Centauri would be a pea about 1,100 km (680 mi) away, and the Milky Way would be about 30 million km (19 million mi) wide. Although stars are more common near the centers of each galaxy, the average distance between stars is still 160 billion ($1.6×10^{11}$) km (100 billion mi). That is analogous to one ping-pong ball every 3.2 km (2.0 mi). Thus, it is extremely unlikely that any two stars from the merging galaxies would collide.

Black Hole Collisions

The Milky Way and Andromeda galaxies each contain a central supermassive black hole (SMBH), these being Sagittarius A* (c. $3.6×10^{6}$ $M_{\odot}$) and an object within the P2 concentration of Andromeda's nucleus ($1–2×10^{8}$ $M_{\odot}$). These black holes will converge near the centre of the newly formed galaxy over a period that may take millions of years, due to a process known as dynamical friction: as the SMBHs move relative to the surrounding cloud of much less massive stars, gravitational interactions lead to a net transfer of orbital energy from the SMBHs to the stars, causing the stars to be "slingshotted" into higher-radius orbits, and the SMBHs to "sink" toward the galactic core. When the SMBHs come within one light-year of one another, they will begin to strongly emit gravitational waves that will radiate further orbital energy until they merge completely. Gas taken up by the combined black hole could create a luminous quasar or an active galactic nucleus, releasing as much energy as 100 million supernova explosions. As of 2006, simulations indicated that the Sun might be brought near the centre of the combined galaxy, potentially coming near one of the black holes before being ejected entirely out of the galaxy. Alternatively, the Sun might approach one of the black holes a bit closer and be torn apart by its gravity. Parts of the former Sun would be pulled into the black hole.

Fate of the Solar System

Two scientists with the Harvard–Smithsonian Center for Astrophysics stated that when, and even whether, the two galaxies collide will depend on Andromeda's transverse velocity. Based on current calculations they predict a 50% chance that in a merged galaxy, the Solar System will be swept out three times farther from the galactic core than its current distance. They also predict a 12% chance that the Solar System will be ejected from the new galaxy sometime during the collision. Such an event would have no adverse effect on the system and the chances of any sort of disturbance to the Sun or planets themselves may be remote.

Excluding planetary engineering, by the time the two galaxies collide the surface of the Earth will have already become far too hot for liquid water to exist, ending all terrestrial life; that is currently estimated to occur in about 3.75 billion years due to gradually increasing luminosity of the Sun (it will have risen by 35–40% above the current luminosity).

Possible Triggered Stellar Events

When two spiral galaxies collide, the hydrogen present on their disks is compressed, producing strong star formation as can be seen on interacting systems like the Antennae Galaxies. In the case of the Andromeda–Milky Way collision, it is believed that there will be little gas remaining in the disks of both galaxies, so the mentioned starburst will be relatively weak, though it still may be enough to form a quasar.

Merger Remnant

The galaxy product of the collision has been nicknamed *Milkomeda* or *Milkdromeda*. According to simulations, this object will look like a giant elliptical galaxy, but with a centre showing less stellar density than current elliptical galaxies. It is, however, possible the resulting object will be a large lenticular galaxy, depending on the amount of remaining gas in the Milky Way and Andromeda.

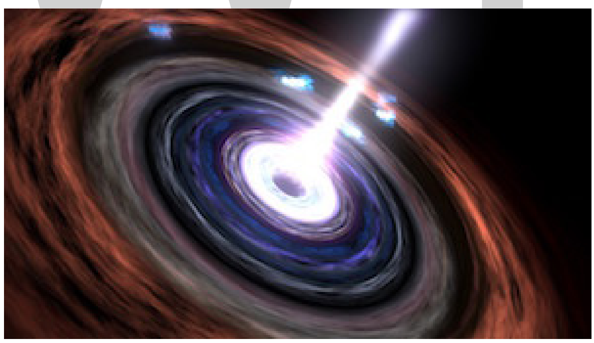
In the far future, roughly 150 billion years from now, the remaining galaxies of the Local Group will coalesce into this object, that being the next evolutionary stage of the local group of galaxies.

Active Galaxies

Active galaxies are galaxies that have a small core of emission embedded at the center of an otherwise typical galaxy. This core is typically highly variable and very bright compared to the rest of the galaxy.

For normal galaxies, we think of the total energy they emit as the sum of the emission from each of the stars found in the galaxy, but in active galaxies, this is not true. There is a great deal more emitted energy in active galaxies than there should be and this excess energy is found in the infrared, radio, UV, and X-ray regions of the electromagnetic spectrum. The energy emitted by an active galaxy, AGN for short, is anything but normal.

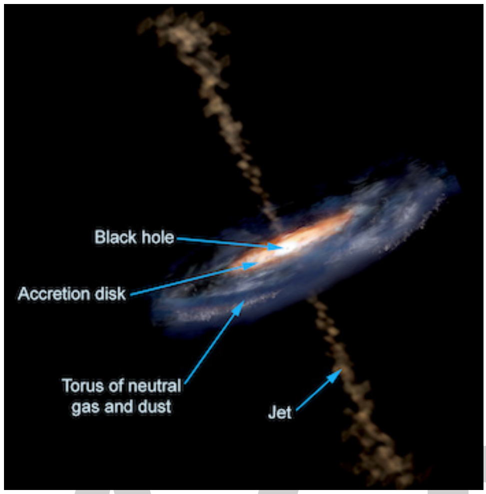
Most, if not all, normal galaxies have a supermassive black hole at their center. In an active galaxy, its supermassive black hole is accreting material from the galaxy's dense central region. As the material falls in toward the black hole, angular momentum will cause it to spiral in and form into a disk. This disk, called an accretion disk, heats up due to the gravitational and frictional forces at work.



An artist's concept of the central region of an active galaxy.

Models of active galaxies also include a region of cold gas and dust, thought to be in the shape of a giant donut with the black hole and accretion disk nestled in the donut's hole. In about one out of ten AGN, the black hole and accretion disk produce narrow beams of energetic particles and ejects them outward in opposite directions away from the disk. These jets, which emerge at nearly the speed of light, become a powerful source of radio wave emission.

The properties of an active galaxy are determined by the black hole's mass, the rate of accretion onto the black hole, whether or not it has a powerful jet, and the angle at which we view the galaxy. Radio galaxies, quasars, and blazars are AGN with strong jets that can travel outward into large regions of intergalactic space. Some of the apparent differences between types of AGN are due to our having different orientations with respect to the disk. With blazars and quasars, we are looking down the jet.



This illustration shows the different features of an active galactic nucleus (AGN). The extreme luminosity of an AGN is powered by accretion onto a supermassive black hole. Some AGN have jets, while others do not.

Active galaxies are intensely studied at all wavelengths. Since they can change their behavior on short timescales, it is useful to study them simultaneously at all energies. X-ray and gamma-ray observations have proven to be important parts of this multi wavelength approach since many high-energy quasars emit a large fraction of their power at such energies. The X-rays in AGN originate from very near the black hole, so X-ray studies can provide scientists with unique insights into the physical processes occurring in the central engine. In addition, gamma-ray observations alone can provide valuable information on the nature of particle acceleration in the quasar jet and clues as to how the particles interact with their surroundings.

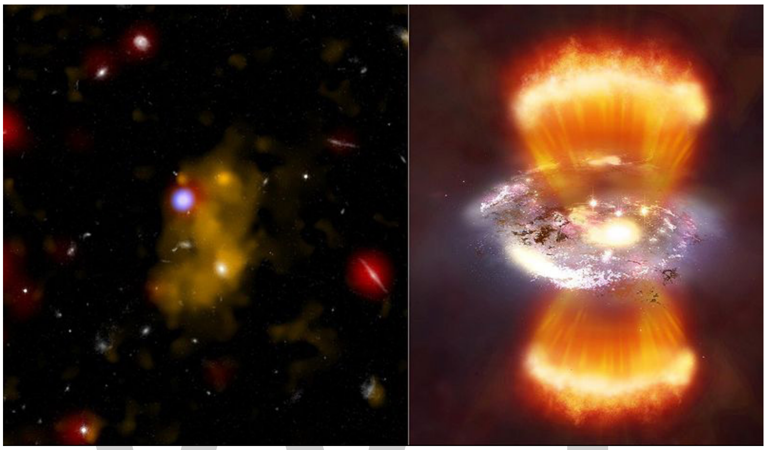
Lyman-alpha Emitter

A Lyman-alpha emitter (LAE) is a type of distant galaxy that emits Lyman-alpha radiation from neutral hydrogen.

Most known LAEs are extremely distant, and because of the finite travel time of light they provide glimpses into the history of the universe. They are thought to be the progenitors of most modern Milky Way type galaxies. These galaxies can be found nowadays rather easily in narrow-band searches by an excess of their narrow-band flux at a wavelength which may be interpreted as their redshift:

$$1 + z = \frac{\lambda}{1215.67\mathring{A}}$$

where z is the redshift, λ is the observed wavelength, and 1215.67 Å is the wavelength of Lyman-alpha emission. The Lyman-alpha line in most LAEs is thought to be caused by recombination of interstellar hydrogen that is ionized by an ongoing burst of star-formation. Such Lyman alpha emission was first suggested as a signature of young galaxies by Bruce Partridge and P. J. E. Peebles in 1967. Experimental observations of the redshift of LAEs are important in cosmology because they trace dark matter halos and subsequently the evolution of matter distribution in the universe.



A Lyman alpha emitter (left) and an artist's impression of what one might look like if viewed at a relatively close distance (right).

Properties

Lyman-alpha emitters are typically low mass galaxies of 10^8 or 10^10 solar masses. They are typically young galaxies that are 200 to 600 million years old, and they have the highest specific star formation rate of any galaxies known. All of these properties indicate that Lyman-alpha emitters are important clues as to the progenitors of modern Milky Way type galaxies.

Lyman-alpha emitters have many unknown properties. The Lyman-alpha photon escape fraction varies greatly in these galaxies. This is what portion of the light emitted at the Lyman-alpha line wavelength inside the galaxy actually escapes and will be visible to distant observers. There is much evidence that the dust content of these galaxies could be significant and therefore is obscuring the brightness of these galaxies. It is also possible that anisotropic distribution of hydrogen density and velocity play a significant role in the varying escape fraction due to the photons' continued interaction with the hydrogen gas (radiative transfer). Evidence now shows strong evolution in the Lyman Alpha escape fraction with redshift, most likely associated with the buildup of dust in the ISM. Dust is shown to be the main parameter setting the escape of Lyman Alpha photons. Additionally the metallicity, outflows, and detailed evolution with redshift is unknown.

Importance in Cosmology

LAEs are important probes of reionization, cosmology (BAO), and they allow probing of the faint end of the luminosity function at high redshift.

The baryonic acoustic oscillation signal should be evident in the power spectrum of Lyman-alpha emitters at high redshift. Baryonic acoustic oscillations are imprints of sound waves on scales where radiation pressure stabilized the density perturbations against gravitational collapse in the early universe. The three-dimensional distribution of the characteristically homogeneous Lyman-alpha galaxy population will allow a robust probe of cosmology. They are a good tool because the Lyman-alpha bias, the propensity for galaxies to form in the highest overdensity of the underlying dark matter distribution, can be modeled and accounted for. Lyman-alpha emitters are over dense in clusters.

Luminous Infrared Galaxy

Luminous infrared galaxies or LIRGs are galaxies with luminosities, the measurement of brightness, above $10^{11}\ L_\odot$. They are also referred to as submillimeter galaxies (SMGs) through their normal method of detection. LIRGs are more abundant than starburst galaxies, Seyfert galaxies and quasi-stellar objects at comparable luminosity. Infrared galaxies emit more energy in the infrared than at all other wavelengths combined. A LIRG's luminosity is 100 billion times that of our sun.

Galaxies with luminosities above $10^{12}\ L_\odot$ are ultraluminous infrared galaxies (ULIRGs). Galaxies exceeding $10^{13}\ L_\odot$ are characterised as hyper-luminous infrared galaxies (HyLIRGs). Those exceeding $10^{14}\ L_\odot$ are extremely luminous infrared galaxies (ELIRGs). Many of the LIRGs and ULIRGs are showing interactions and disruptions. Many of these types of galaxies spawn about 100 new stars a year as compared to ours which spawn one a year; this creates the high level of luminosity.

Discovery and Characteristics

Infrared galaxies appear to be single, gas-rich spirals whose infrared luminosity is created largely by the formation of stars within them. These types of galaxies were discovered in 1983 with IRAS. A LIRG's excess infrared luminosity may also come from the presence of an active galactic nucleus (AGN) residing at the center.

These galaxies emit more energy in the infrared portion of the spectrum, not visible to the naked eye. The energy given off by LIRGs is comparable to that of a quasar (a type of AGN), which formerly was known as the most energetic object in the universe.

LIRGs are brighter in the infrared than in the optical spectrum because the visible light is absorbed by the high amounts of gas and dust, and the dust re-emits thermal energy in the infrared spectrum. LIRGs are known to live in denser parts of the universe than non-LIRGs.

ULIRG

IRAS 14348-1447 is an ultraluminous infrared galaxy, located over a billion light-years away.

LIRGs are also capable of becoming Ultra Luminous Infrared Galaxys (ULIRGs) but there is no perfect timetable because not all LIRGs turn into ULIRGs, Newtonian mechanics is used in the calculations and because the constraints are not quite approximate. Studies have shown that ULIRGs are more likely to contain an AGN than LIRGs.

According to one study a ULIRG is just part of an evolutionary galaxy merger scenario. In essence, two or more spiral galaxies, galaxies that consist of a flat, rotating disk containing stars, gas and dust and a central concentration of stars known as the bulge, merge to form an early stage merger. An early stage merger in this case can also be identified as a LIRG. After that, it becomes a late stage merger, which is a ULIRG. It then becomes a quasar and in the final stage of the evolution it becomes an elliptical galaxy. This can be evidenced by the fact that stars are much older in elliptical galaxies than those found in the earlier stages of the evolution.

HyLIRG

Hyper luminous Infrared Galaxies (HyLIRG), also referred to as HiLIRGs and HLIRGs, are considered to be some of the most luminous persistent objects in the Universe, exhibiting extremely high star formation rates, and most of which are known to harbour Active Galactic Nuclei (AGN). They are defined as galaxies with luminosities above 10^{13} $L_\odot$, as distinct from the less luminous population of ULIRGs ($L = 10^{12} - 10^{13}$ $L_\odot$). HLIRGs were first identified through follow-up observations of the IRAS mission.

IRAS F10214+4724, a HyLIRG being gravitationally lensed by a foreground elliptical galaxy, was considered to be one of the most luminous objects in the Universe having an intrinsic luminosity of $\sim 2 \times 10^{13}$ $L_\odot$. It is believed that the bolometric luminosity of this HLIRG is likely amplified by a factor of ~30 as a result of the gravitational lensing.

The majority (~80%) of the mid-infrared spectrum of these objects is found to be dominated by AGN emission. However, the starburst (SB) activity is known to be significant in all known sources with a mean SB contribution of ~30%. Star formation rates in HLIRGs have been shown to reach ~ $3\times10^2 - 3\times10^3\ M_\odot\ yr^{-1}$.

ELIRG

The Extremely Luminous Infrared Galaxy WISE J224607.57-052635.0, with a luminosity of 300 trillion suns was discovered by NASA's Wide-field Infrared Survey Explorer (WISE), and as of May 2015 is the most luminous galaxy found. The galaxy belongs to a new class of objects discovered by WISE, extremely luminous infrared galaxies, or ELIRGs.

Light from the WISE J224607.57-052635.0 galaxy has traveled 12.5 billion years. The black hole at its center was billions of times the mass of our sun when the universe was a tenth (1.3 billion years) of its present age of 13.8 billion years.

There are three reasons the black holes in the ELIRGs could be massive. First, the embryonic black holes might be bigger than thought possible. Second, the Eddington limit was exceeded. When a black hole feeds, gas falls in and heats, emitting light. The pressure of the emitted light forces the gas outward, creating a limit to how fast the black hole can continuously absorb matter. If a black hole broke this limit, it could theoretically increase in size at a fast rate. Black holes have previously been observed breaking this limit; the black hole in the study would have had to repeatedly break the limit to grow this large. Third, the black holes might just be bending this limit, absorbing gas faster than thought possible, if the black hole is not spinning fast. If a black hole spins slowly, it will not repel its gas absorption as much. A slow-spinning black hole can absorb more matter than a fast-spinning black hole. The massive black holes in ELIRGs could be absorbing matter for a longer time.

Twenty new ELIRGs, including the most luminous galaxy found to date, have been discovered. These galaxies were not found earlier because of their distance, and because dust converts their visible light into infrared light. One has been observed to have three star-forming areas.

IRAS

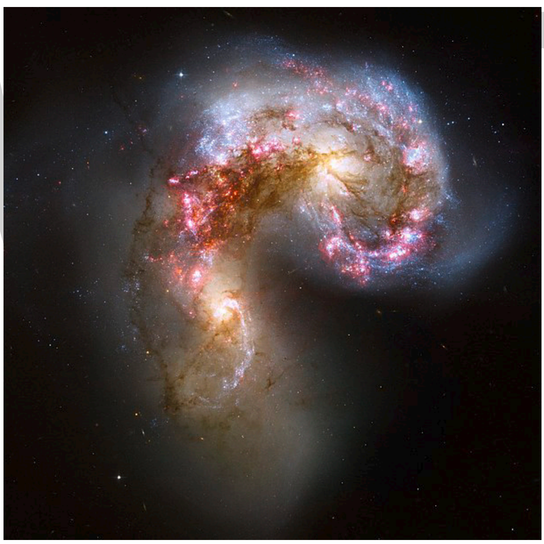
The Infrared Astronomical Satellite (IRAS) was the first all-sky survey which used far-infrared wavelengths, in 1983. In that survey, tens of thousands of galaxies were detected, many of which would not have been recorded in previous surveys. It is now clear that the reason the number of detections has risen is that the majority of LIRGs in the universe emitted the bulk of their energy in the far infrared. Using the IRAS, scientists were able to determine the luminosity of the galactic objects discovered. The telescope was a joint project of the United States (NASA), Netherlands (NIVR), and the

United Kingdom (SERC). Over 250,000 infrared sources were observed during this 10-month mission.

GOALS

The Great Observatories All-sky LIRG Survey (GOALS) is a multi-wavelength study of luminous infrared galaxies, incorporating observations with NASA's Great Observatories and other ground and space-based telescopes. Using information from NASA's Spitzer, Hubble, Chandra and Galex observations in a study over 200 of the most luminous infrared selected galaxies in the local universe. Approximately 180 LIRGs were identified along with over 20 ULIRGs. The LIRGs and ULIRGs targeted in GOALS span the full range of nuclear spectral types (type-1 and type 2 Active Galactic Nuclei, LINERS's, and starbursts) and interaction stages (major mergers, minor mergers, and isolated galaxies).

Starburst Galaxy



The Antennae Galaxies are an example of a starburst galaxy occurring from the collision of NGC 4038/NGC 4039.

A starburst galaxy is a galaxy undergoing an exceptionally high rate of star formation, as compared to the long-term average rate of star formation in the galaxy or the star formation rate observed in most other galaxies. For example, the star formation rate of the Milky Way galaxy is approximately 3 $M_{\odot}$/yr, however, starburst galaxies can experience star formation rates that are more than a factor of 100 times greater. In a starburst galaxy, the rate of star formation is so large that the galaxy will consume all of its gas reservoir, from which the stars are forming, on a timescale much shorter than the age of the galaxy. As such, the starburst nature of a galaxy is a phase, and one that typically occupies a brief period of a galaxy's evolution. The majority of starburst galaxies are in the midst of a merger or close encounter with another galaxy. Starburst galaxies include M82, NGC 4038/NGC 4039 (the Antennae Galaxies), and IC 10.

Starburst galaxies are defined by these three interrelated factors:

- The rate at which the galaxy is currently converting gas into stars (the star-formation rate, or SFR).
- The available quantity of gas from which stars can be formed.
- A comparison of the timescale on which star formation will consume the available gas with the age or rotation period of the galaxy.

Commonly used definitions include:

- Continued star-formation where the current SFR would exhaust the available gas reservoir in much less than the age of the Universe (the Hubble Time).
- Continued star-formation where the current SFR would exhaust the available gas reservoir in much less than the dynamical timescale of the galaxy (perhaps one rotation period in a disk type galaxy).
- The current SFR, normalised by the past-averaged SFR, is much greater than unity. This ratio is referred to as the "birthrate parameter".

Light and dust in a nearby starburst galaxy.

Triggering Mechanisms

Mergers and tidal interactions between gas-rich galaxies play a large role in driving starbursts. Galaxies in the midst of a starburst frequently show tidal tails, an indication of a close encounter with another galaxy, or are in the midst of a merger. Interactions between galaxies that do not merge can trigger unstable rotation modes, such as the bar instability, which causes gas to be funneled towards the nucleus and ignites bursts of star formation near the galactic nucleus. It has been shown that there is a strong correlation between the lopsidedness of a galaxy and the youth of its stellar population, with more lopsided galaxies having younger central stellar populations. As lopsidedness can be caused by tidal interactions and mergers between galaxies, this result gives further evidence that mergers and tidal interactions can induce central star formation in a galaxy and drive a starburst.

Types

Artist's impression of a galaxy undergoing a starburst.

Classifying types of starburst galaxies is difficult since starburst galaxies do not represent a specific type in and of themselves. Starbursts can occur in disk galaxies, and irregular galaxies often exhibit knots of starburst spread throughout the irregular galaxy. Nevertheless, astronomers typically classify starburst galaxies based on their most distinct observational characteristics. Some of the categorizations include:

- Blue compact galaxies (BCGs): These galaxies are often low mass, low metallicity, dust-free objects. Because they are dust-free and contain a large number of hot, young stars, they are often blue in optical and ultraviolet colours. It was initially thought that BCGs were genuinely young galaxies in the process of forming their first generation of stars, thus explaining their low metal content. However, old stellar populations have been found in most BCGs, and it is thought that efficient mixing may explain the apparent lack of dust and metals. Most BCGs show signs of recent mergers and/or close interactions. Well-studied BCGs include IZw18 (the most metal poor galaxy known), ESO338-IG04 and Haro11.
    - Blue compact dwarf galaxies (BCD galaxies) are small compact galaxies.
    - Green Pea galaxies (GPs) are small compact galaxies resembling primordial starbursts. They were found by citizen scientists taking part in the Galaxy Zoo project.
- Luminous infrared galaxies (LIRGs):
    - Ultra-luminous Infrared Galaxies ULIRGs): These galaxies are generally extremely dusty objects. The ultraviolet radiation produced by the obscured star-formation is absorbed by the dust and reradiated in the infrared spectrum at wavelengths of around 100 micrometres. This explains the extreme red colours associated with ULIRGs. It is not known for sure that the UV radiation is produced purely by star-formation, and some astronomers

believe ULIRGs to be powered (at least in part) by active galactic nuclei (AGN). X-ray observations of many ULIRGs that penetrate the dust suggest that many starburst galaxies are double-cored systems, lending support to the hypothesis that ULIRGs are powered by star-formation triggered by major mergers. Well-studied ULIRGs include Arp 220.

  - Hyperluminous Infrared galaxies (HLIRGs), sometimes called Submillimeter galaxies.

- Wolf-Rayet galaxies (WR galaxies), galaxies where a large portion of the bright stars are Wolf-Rayet stars. The Wolf-Rayet phase is a relatively short-lived phase in the life of massive stars, typically 10% of the total life-time of these stars and as such any galaxy is likely to contain few of these. However, because the stars are both very luminous and have very distinctive spectral features, it is possible to identify these stars in the spectra of entire galaxies and doing so allows good constraints to be placed on the properties of the starbursts in these galaxies.

Ingredients

Messier 82 is the prototype nearby starburst galaxy about 12 million light-years away in the constellation Ursa Major.

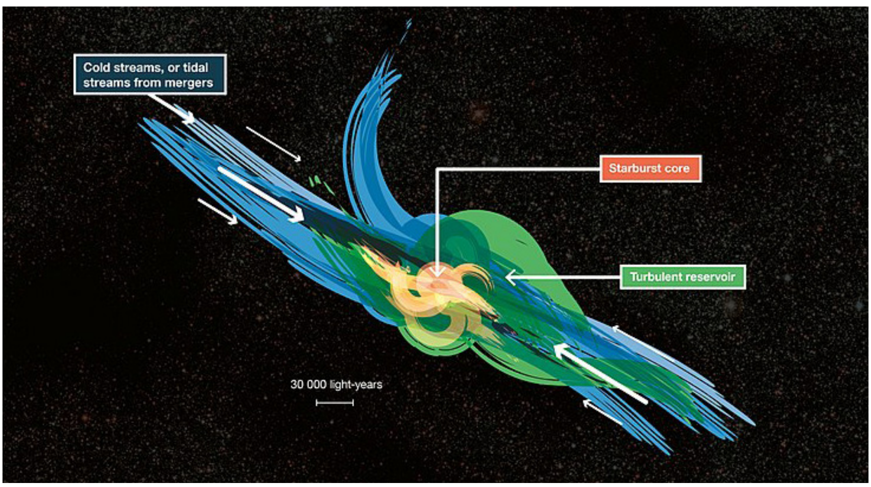
Firstly, a starburst galaxy must have a large supply of gas available to form stars. The burst itself may be triggered by a close encounter with another galaxy (such as M81/M82), a collision with another galaxy (such as the Antennae), or by another process which forces material into the centre of the galaxy (such as a stellar bar).

The inside of the starburst is quite an extreme environment. The large amounts of gas mean that very massive stars are formed. Young, hot stars ionize the gas (mainly hydrogen) around them, creating H II regions. Groups of very hot stars are known as OB associations. These stars burn very bright and very fast, and are quite likely to explode at the end of their lives as supernovae.

After the supernova explosion, the ejected material expands and becomes a supernova remnant. These remnants interact with the surrounding environment within the starburst (the interstellar medium) and can be the site of naturally occurring masers.

Studying nearby starburst galaxies can help us determine the history of galaxy formation and evolution. Large numbers of the very distant galaxies seen, for example, in the Hubble Deep Field are known to be starbursts, but they are too far away to be studied in any detail. Observing nearby examples and exploring their characteristics can give us an idea of what was happening in the early universe as the light we see from these distant galaxies left them when the universe was much younger. However, starburst galaxies seem to be quite rare in our local universe, and are more common further away - indicating that there were more of them billions of years ago. All galaxies were closer together then, and therefore more likely to be influenced by each other's gravity. More frequent encounters produced more starbursts as galactic forms evolved with the expanding universe.

Examples:



Artist's impression of gas fueling distant starburst galaxies.

M82 is the archetypal starburst galaxy. Its high level of star formation is due to a close encounter with the nearby spiral M81. Maps of the regions made with radio telescopes show large streams of neutral hydrogen connecting the two galaxies, also as a result of the encounter. Radio images of the central regions of M82 also show a large number of young supernova remnants, left behind when the more massive stars created in the starburst came to the end of their lives.

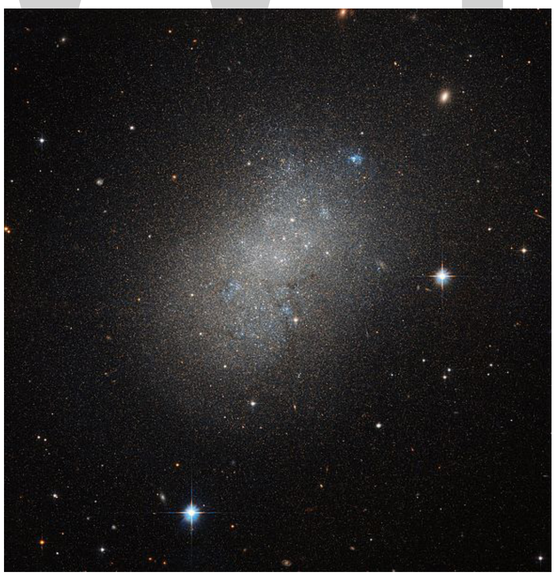
Dwarf Galaxy

A dwarf galaxy is a small galaxy composed of about 100 million up to several billion stars, a small number compared to the Milky Way's 200–400 billion stars. The Large Magellanic Cloud, which closely orbits the Milky Way and contains over 30 billion stars, is sometimes classified as a dwarf galaxy; others consider it a full-fledged galaxy. Dwarf galaxies' formation and activity are thought to be heavily influenced by interactions with larger galaxies. Astronomers identify numerous types of dwarf galaxies, based on their shape and composition.

The Large Magellanic Cloud, a satellite galaxy of the Milky Way.

Formation

Current theory states that most galaxies, including dwarf galaxies, form in association with dark matter, or from gas that contains metals. However, NASA's Galaxy Evolution Explorer space probe identified new dwarf galaxies forming out of gases with low metallicity. These galaxies were located in the Leo Ring, a cloud of hydrogen and helium around two massive galaxies in the constellation Leo.



Dwarf galaxies like NGC 5264 typically possess around a billion stars.

Because of their small size, dwarf galaxies have been observed being pulled toward and ripped by neighbouring spiral galaxies, resulting in galaxy merger.

Local Dwarf Galaxies

The Phoenix Dwarf Galaxy is a dwarf irregular galaxy, featuring younger stars in its inner regions and older ones at its outskirts.

There are many dwarf galaxies in the Local Group; these small galaxies frequently orbit larger galaxies, such as the Milky Way, the Andromeda Galaxy and the Triangulum Galaxy. A 2007 paper has suggested that many dwarf galaxies were created by galactic tides during the early evolutions of the Milky Way and Andromeda. Tidal dwarf galaxies are produced when galaxies collide and their gravitational masses interact. Streams of galactic material are pulled away from the parent galaxies and the halos of dark matter that surround them. A 2018 study suggests that some local dwarf galaxies formed extremely early, during the Dark Ages within the first billion years after the big bang.

More than 20 known dwarf galaxies orbit the Milky Way, and recent observations have also led astronomers to believe the largest globular cluster in the Milky Way, Omega Centauri, is in fact the core of a dwarf galaxy with a black hole at its centre, which was at some time absorbed by the Milky Way.

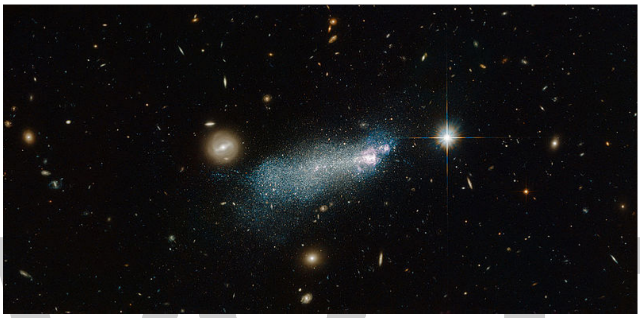
Common Types

UGC 11411 is a galaxy known as an irregular blue compact dwarf (BCD) galaxy.

- Elliptical galaxy: Dwarf elliptical galaxy (dE).
- Dwarf spheroidal galaxy (dSph): Once a subtype of dwarf ellipticals, now regarded as a distinct type.

- Irregular galaxy: Dwarf irregular galaxy (dIrr).
- Spiral galaxy: Dwarf spiral galaxy (dS).
- Magellanic type dwarfs.
- Blue compact dwarf galaxies.
- Ultra-compact dwarf galaxies.

Blue Compact Dwarf Galaxies



Blue compact dwarf PGC 51017.

In astronomy, a blue compact dwarf galaxy (BCD galaxy) is a small galaxy which contains large clusters of young, hot, massive stars. These stars, the brightest of which are blue, cause the galaxy itself to appear blue in colour. Most BCD galaxies are also classified as dwarf irregular galaxies or as dwarf lenticular galaxies. Because they are composed of star clusters, BCD galaxies lack a uniform shape. They consume gas intensely, which causes their stars to become very violent when forming.

BCD galaxies cool in the process of forming new stars. The galaxies' stars are all formed at different time periods, so the galaxies have time to cool and to build up matter to form new stars. As time passes, this star formation changes the shape of the galaxies.

Nearby examples include NGC 1705, NGC 2915, NGC 3353 and UGCA 281.

Ultra-compact Dwarfs

Ultra-compact dwarf galaxies (UCD) are a class of very compact galaxies with very high stellar densities, discovered in the 2000s. They are thought to be on the order of 200 light years across, containing about 100 million stars. It is theorised that these are the cores of nucleated dwarf elliptical galaxies that have been stripped of gas and outlying stars by tidal interactions, travelling through the hearts of rich clusters. UCDs have been found in the Virgo Cluster, Fornax Cluster, Abell 1689, and the Coma Cluster, amongst others. In particular, an unprecedentedly large sample of ~ 100 UCDs has been found in the core region of the Virgo cluster by the Next Generation Virgo Cluster Survey team. The first ever relatively robust studies of the global properties of

Virgo UCDs suggest that UCDs have distinct dynamical and structural properties from normal globular clusters. An extreme example of UCD is M60-UCD1, about 54 million light years away, which contains approximately 200 million solar masses within a 160 light year radius; its central region packs in stars about 25 times closer together than stars in Earth's region in the Milky Way. M59-UCD3 is approximately the same size as M60-UCD1 with a half-light radius, r_h, of approximately 20 parsecs but is 40% more luminous with an absolute visual magnitude of approximately −14.6. This makes M59-UCD3 the densest known galaxy. Based on stellar orbital velocities, two UCD in the Virgo Cluster are claimed to have supermassive black holes weighing 13% and 18% of the galaxies masses.

Spiral Galaxy

Spiral galaxies form a class of galaxy originally described by Edwin Hubble in his 1936 work *The Realm of the Nebulae* and, as such, form part of the Hubble sequence. Most spiral galaxies consist of a flat, rotating disk containing stars, gas and dust, and a central concentration of stars known as the bulge. These are often surrounded by a much fainter halo of stars, many of which reside in globular clusters.

Spiral galaxies are named by their spiral structures that extend from the center into the galactic disc. The spiral arms are sites of ongoing star formation and are brighter than the surrounding disc because of the young, hot OB stars that inhabit them.

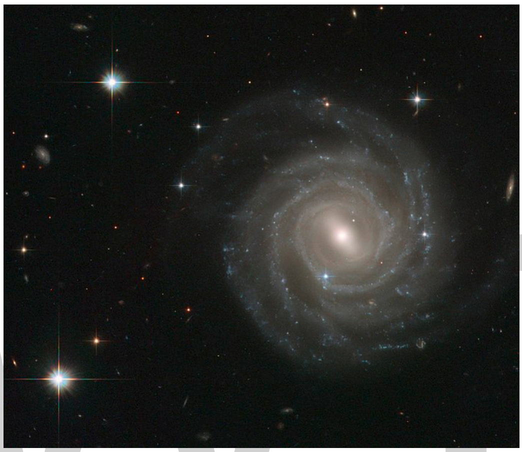
An example of a spiral galaxy, the Pinwheel Galaxy (also known as Messier 101 or NGC 5457).

Roughly two-thirds of all spirals are observed to have an additional component in the form of a bar-like structure, extending from the central bulge, at the ends of which the spiral arms begin. The proportion of barred spirals relative to their barless cousins has likely changed over the history of the Universe, with only about 10% containing bars about 8 billion years ago, to roughly a quarter 2.5 billion years ago, until present, where over two-thirds of the galaxies in the visible universe (Hubble volume) have bars.

The Milky Way is a barred spiral, although the bar itself is difficult to observe from Earth's current position within the galactic disc. The most convincing evidence for the stars forming a bar in the galactic center comes from several recent surveys, including the Spitzer Space Telescope.

Together with irregular galaxies, spiral galaxies make up approximately 60% of galaxies in today's universe. They are mostly found in low-density regions and are rare in the centers of galaxy clusters.

Structure



Barred spiral galaxy UGC 12158.

Spiral galaxies may consist of several distinct components:

- A flat, rotating disc of stars and interstellar matter of which spiral arms are prominent components.
- A central stellar bulge of mainly older stars, which resembles an elliptical galaxy.
- A bar-shaped distribution of stars.
- A near-spherical halo of stars, including many in globular clusters.
- A supermassive black hole at the very center of the central bulge.
- A near-spherical dark matter halo.

The relative importance, in terms of mass, brightness and size, of the different components varies from galaxy to galaxy.

Spiral Arms

Spiral arms are regions of stars that extend from the center of spiral and barred spiral galaxies. These long, thin regions resemble a spiral and thus give spiral galaxies their name. Naturally, different classifications of spiral galaxies have distinct arm-structures. Sc and SBc galaxies, for instance, have very "loose" arms, whereas Sa and SBa galaxies have tightly wrapped arms (with reference to the Hubble sequence). Either way, spiral arms contain many young, blue stars (due to the high mass density and the high rate of star formation), which make the arms so bright.

NGC 1300 in infrared light.

Bulge

A bulge is a large, tightly packed group of stars. The term refers to the central group of stars found in most spiral galaxies, often defined as the excess of stellar light above the inward extrapolation of the outer (exponential) disk light.

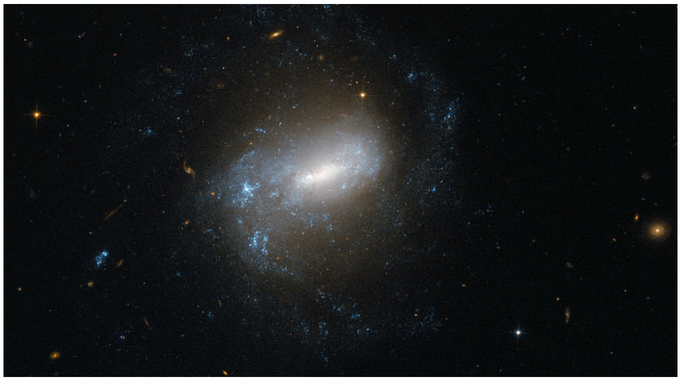
Using the Hubble classification, the bulge of Sa galaxies is usually composed of Population II stars, which are old, red stars with low metal content. Further, the bulge of Sa and SBa galaxies tends to be large. In contrast, the bulges of Sc and SBc galaxies are much smaller and are composed of young, blue Population I stars. Some bulges have similar properties to those of elliptical galaxies (scaled down to lower mass and luminosity); others simply appear as higher density centers of disks, with properties similar to disk galaxies.

Many bulges are thought to host a supermassive black hole at their centers. In our own galaxy, for instance, the object called Sagittarius A* is believed to be a supermassive black hole. There are many lines of evidence for the existence of black holes in spiral galaxy centers, including the presence of active nuclei in some spiral galaxies, and dynamical measurements that find large compact central masses in galaxies such as NGC 4258.

Bar

Bar-shaped elongations of stars are observed in roughly two-thirds of all spiral galaxies. Their presence may be either strong or weak. In edge-on spiral (and lenticular) galaxies, the presence of the bar can sometimes be discerned by the out-of-plane X-shaped or (peanut shell)-shaped structures which typically have a maximum visibility at half the length of the in-plane bar.

Spheroid



Spiral galaxy NGC 1345.

The bulk of the stars in a spiral galaxy are located either close to a single plane (the galactic plane) in more or less conventional circular orbits around the center of the galaxy (the Galactic Center), or in a spheroidal galactic bulge around the galactic core.

However, some stars inhabit a *spheroidal halo* or *galactic spheroid*, a type of galactic halo. The orbital behaviour of these stars is disputed, but they may exhibit retrograde and/or highly inclined orbits, or not move in regular orbits at all. Halo stars may be acquired from small galaxies which fall into and merge with the spiral galaxy—for example, the Sagittarius Dwarf Spheroidal Galaxy is in the process of merging with the Milky Way and observations show that some stars in the halo of the Milky Way have been acquired from it.

NGC 428 is a barred spiral galaxy, located approximately 48 million light-years away from Earth in the constellation of Cetus.

Unlike the galactic disc, the halo seems to be free of dust, and in further contrast, stars in the galactic halo are of Population II, much older and with much lower metallicity than their Population I cousins in the galactic disc (but similar to those in the galactic bulge). The galactic halo also contains many globular clusters.

The motion of halo stars does bring them through the disc on occasion, and a number of small red dwarfs close to the Sun are thought to belong to the galactic halo, for example Kapteyn's Star and Groombridge 1830. Due to their irregular movement around the center of the galaxy, these stars often display unusually high proper motion.

Oldest Spiral Galaxy

The oldest spiral galaxy on file is BX442. At eleven billion years old, it is more than two billion years older than any previous discovery. Researchers think the galaxy's shape is caused by the gravitational influence of a companion dwarf galaxy. Computer models based on that assumption indicate that BX442's spiral structure will last about 100 million years.

Related

In June 2019, citizen scientists through Galaxy Zoo reported that the usual Hubble classification, particularly concerning spiral galaxies, may not be supported, and may need updating.

Origin of the Spiral Structure

Spiral galaxy NGC 6384 taken by Hubble Space Telescope.

The pioneer of studies of the rotation of the Galaxy and the formation of the spiral arms was Bertil Lindblad in 1925. He realized that the idea of stars arranged permanently in a spiral shape was untenable. Since the angular speed of rotation of the galactic disk varies with distance from the centre of the galaxy (via a standard solar system type of gravitational model), a radial arm (like a spoke) would quickly become curved as the galaxy rotates. The arm would, after a few galactic rotations, become increasingly curved and wind around the galaxy ever tighter. This is called the *winding problem*. Measurements in the late 1960s showed that the orbital velocity of stars in spiral galaxies with respect to their distance from the galactic center is indeed higher than expected from Newtonian dynamics but still cannot explain the stability of the spiral structure.

Since the 1970s, there have been two leading hypotheses or models for the spiral structures of galaxies:

- Star formation caused by density waves in the galactic disk of the galaxy.

- The stochastic self-propagating star formation model (sspsf model) – star formation caused by shock waves in the interstellar medium. The shock waves are caused by the stellar winds and supernovae from recent previous star formation, leading to self-propagating and self-sustaining star formation. Spiral structure then arises from differential rotation of the galaxy's disk.

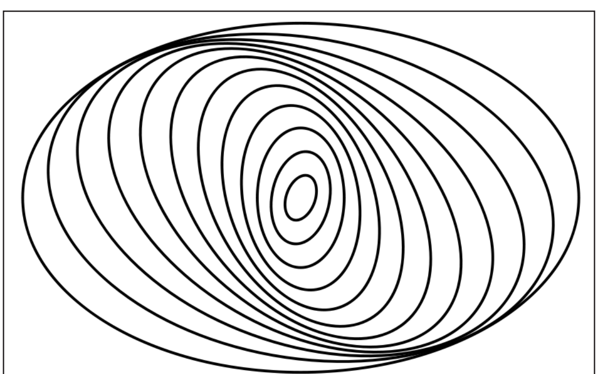
These different hypotheses are not mutually exclusive, as they may explain different types of spiral arms.

The spiral galaxy NGC 1084, home of five supernovae.

Density Wave Model

Bertil Lindblad proposed that the arms represent regions of enhanced density (density waves) that rotate more slowly than the galaxy's stars and gas. As gas enters a density wave, it gets squeezed and makes new stars, some of which are short-lived blue stars that light the arms.

Historical Theory of Lin and Shu



Exaggerated diagram illustrating Lin and Shu's explanation of spiral arms in terms of slightly elliptical orbits.

The first acceptable theory for the spiral structure was devised by C. C. Lin and Frank Shu in 1964, attempting to explain the large-scale structure of spirals in terms of a small-amplitude wave propagating with fixed angular velocity, that revolves around the galaxy at a speed different from that of the galaxy's gas and stars. They suggested that the spiral arms were manifestations of spiral density waves – they assumed that the stars travel in slightly elliptical orbits, and that the orientations of their orbits is correlated i.e. the ellipses vary in their orientation (one to another) in a smooth way with increasing distance from the galactic center. This is illustrated in the diagram above. It is clear that the elliptical orbits come close together in certain areas to give the effect of arms. Stars therefore do not remain forever in the position, but pass through the arms as they travel in their orbits.

Star Formation Caused by Density Waves

The following hypotheses exist for star formation caused by density waves:

- As gas clouds move into the density wave, the local mass density increases. Since the criteria for cloud collapse (the Jeans instability) depends on density, a higher density makes it more likely for clouds to collapse and form stars.
- As the compression wave goes through, it triggers star formation on the leading edge of the spiral arms.
- As clouds get swept up by the spiral arms, they collide with one another and drive shock waves through the gas, which in turn causes the gas to collapse and form stars.

More young Stars in Spiral Arms

The bright galaxy NGC 3810 demonstrates classical spiral structure.

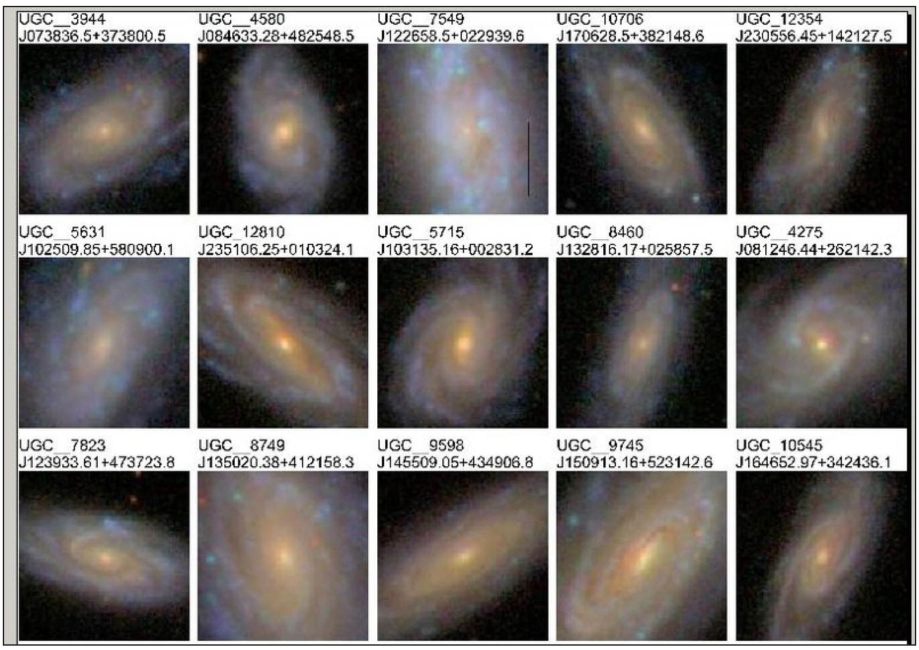
Spiral arms appear visually brighter because they contain both young stars and more massive and luminous stars than the rest of the galaxy. As massive stars evolve far more quickly, their demise tends to leave a darker background of fainter stars immediately behind the density waves. This make the density waves much more prominent.

Spiral arms simply appear to pass through the older established stars as they travel in their galactic orbits, so they also do not necessarily follow the arms. As stars move through an arm, the space velocity of each stellar system is modified by the gravitational force of the local higher density. Also the newly created stars do not remain forever fixed in the position within the spiral arms, where the average space velocity returns to normal after the stars depart on the other side of the arm.

Gravitationally Aligned Orbits

Charles Francis and Erik Anderson showed from observations of motions of over 20,000 local stars (within 300 parsecs) that stars do move along spiral arms, and described how mutual gravity between stars causes orbits to align on logarithmic spirals. When the theory is applied to gas, collisions between gas clouds generate the molecular clouds in which new stars form, and evolution towards grand-design bisymmetric spirals is explained.

Distribution of Stars in Spirals



The similar distribution of stars in Spirals.

The stars in spirals are distributed in thin disks radial with intensity profiles such that

$$I(R) = I_0 e^{-R/h},$$

with h being the disk scale-length; I_0 is the central value; it is useful to define: $R_{opt} = 3.2h$ as the size of the stellar disk, whose luminosity is:

$$L_{tot} = 2\pi I_0 h^2.$$

The spiral galaxies light profiles, in terms of the coordinate R/h, do not depend on galaxy luminosity.

Spiral Nebula

Before it was understood that spiral galaxies existed outside of our Milky Way galaxy, they were often referred to as spiral nebulae. The question of whether such objects were separate galaxies independent of the Milky Way, or a type of nebula existing within our own galaxy, was the subject of the Great Debate of 1920, between Heber Curtis of Lick Observatory and Harlow Shapley of Mt. Wilson Observatory. Beginning in 1923, Edwin Hubble observed Cepheid variables in several spiral nebulae, including the so-called "Andromeda Nebula", proving that they are, in fact, entire galaxies outside our own. The term spiral nebula has since fallen out of use.

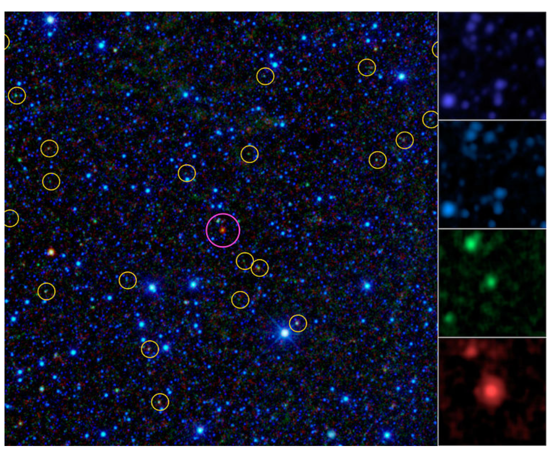
Faint Blue Galaxy

The faint blue galaxy (F.B.G.) problem in astrophysics first arose with observations starting in 1978 that there were more galaxies with a bolometric magnitude > 22 than then-current theory predicted. Galaxies can appear faint because they are small or because they are far away. Neither explanation, nor any combination, initially matched the observations. The distribution of these galaxies has since been found to be consistent with cosmic inflation, measurements of the cosmic microwave background, and a non-zero cosmological constant, that is, with the existence of the now-accepted dark energy. It thus serves as a confirmation of supernova observations requiring dark energy.

A second problem arose in 1988, with even deeper observations showing a much greater excess of faint galaxies. These are now interpreted as dwarf galaxies experiencing large bursts of stellar formation, resulting in blue light from young, massive stars. Thus F.B.G.s are extremely bright for their size and distance.

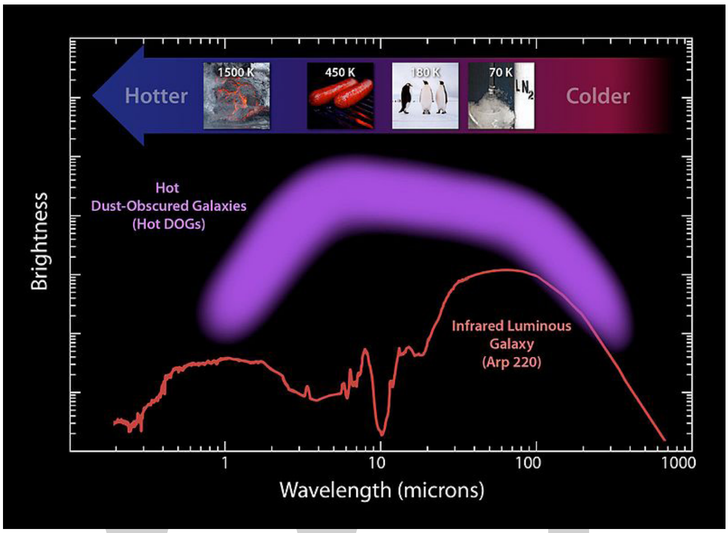
Most F.B.G.s appear between red-shift 0.5 and 2. It is believed that they disappear as separate objects by merger with other galaxies.

Hot, Dust-obscured Galaxy



A hot, dust-obscured galaxy.

A hot, dust-obscured galaxy, or hot DOG, is a rare type of quasar. The central black hole of such a galaxy emits vast amounts of radiation which heats the infalling dust and gas, releasing infrared light at a rate about 1,000 times as much as the Milky Way, making these some of the most luminous galaxies in the universe. However, the density of the surrounding dust is so great that most of that light is obscured. Their average temperatures range from 60 to 120 K (−213 to −153 °C; −352 to −244 °F), significantly higher than an average galaxy's temperature of 30 to 40 K (−243 to −233 °C; −406 to −388 °F). They also appear to concentrate a much higher proportion of their galactic mass in the central black hole than is observed in normal galaxies.



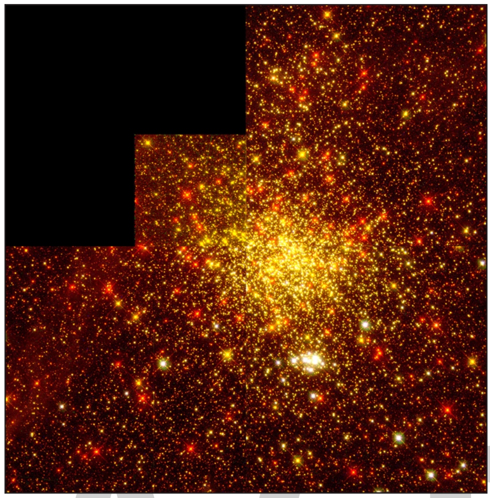
Luminosity plot for hot DOGs and a prototypical luminous infrared galaxy.

Researchers believe that hot DOGs may represent a phase of galactic evolution where the central black hole is capturing material at a rate faster than new stars are forming, yet the radiation pressure from that rapid absorption is pushing away much of that surrounding material. The black hole will eventually clear its area of influence of the excessive dust and gas, rendering it a regular, visible galaxy.

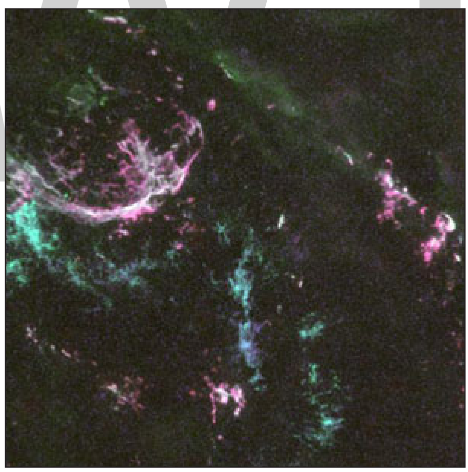
These objects were first detected by the Wide-field Infrared Survey Explorer (WISE) in 2010, and only one out of every 3,000 quasars observed by WISE are of this type. Wu, refer to these galaxies as "W1W2-dropouts" because they are faint or invisible in WISE's W1 (3.4 μm) and W2 (4.6 μm) detection bands.

Magellanic Cloud

Magellanic Cloud is either of two satellite galaxies of the Milky Way Galaxy, the vast star system of which Earth is a minor component. These companion galaxies were named for the Portuguese navigator Ferdinand Magellan, whose crew discovered them during the first voyage around the world. The Magellanic Clouds were recognized early in the 20th century as companion objects to the Milky Way Galaxy. When American astronomer Edwin Hubble established the extragalactic nature of what are now called galaxies, it became plain that the Magellanic Clouds had to be separate systems.



Globular cluster NGC 1850 in the Large Magellanic Cloud. Most of the cluster consists of yellow stars; the bright white stars are members of a second, open cluster about 200 light-years beyond NGC.

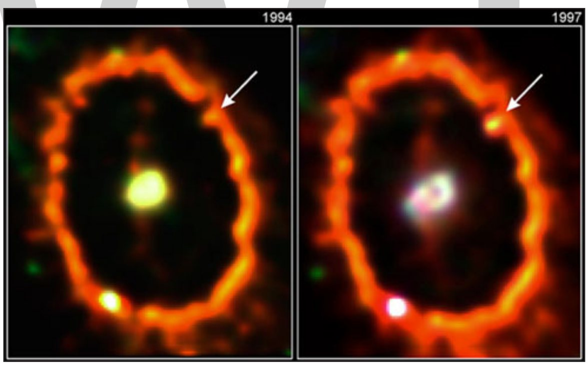


N132D, remnants of a supernova in the Large Magellanic Cloud.

The Magellanic Clouds are irregular galaxies that share a gaseous envelope and lie about 22° apart in the sky near the south celestial pole. One of them, the Large Magellanic Cloud (LMC), is a luminous patch about 5° in diameter, and the other, the Small Magellanic Cloud (SMC), measures less than 2° across. The Magellanic Clouds are visible to the unaided eye in the Southern Hemisphere, but they cannot be observed from most northern latitudes. The LMC is about 160,000 light-years from Earth, and the SMC lies 190,000 light-years away. The LMC and SMC are 14,000 and 7,000 light-years in diameter, respectively—smaller than the Milky Way Galaxy, which is about 140,000 light-years across.

Small Magellanic Cloud: Infant stars in the Small Magellanic Cloud.

The Magellanic Clouds were formed at about the same time as the Milky Way Galaxy, approximately 13 billion years ago. They are presently captured in orbits around the Milky Way Galaxy and have experienced several tidal encounters with each other and with the Galaxy. They contain numerous young stars and star clusters, as well as some much older stars. One of these star clusters contains R136a1, the most massive star known, with a mass 265 times that of the Sun.

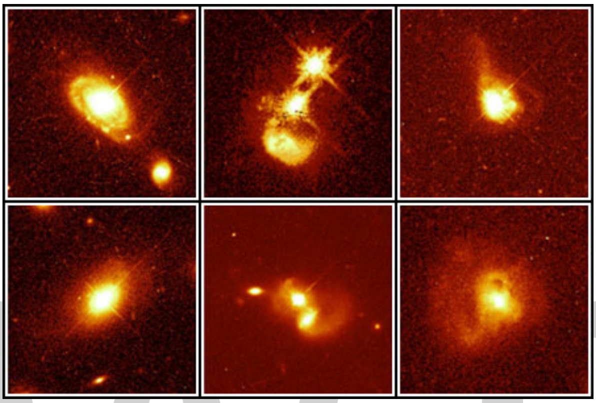


A knot in the central ring of Supernova 1987A, as observed by the Hubble Space Telescope in 1994 (left) and 1997 (right). The knot is caused by the collision of the supernova's blast wave with a slower-moving ring of matter it had ejected earlier. The bright spot on the lower left is an unrelated star.

The Magellanic Clouds serve as excellent laboratories for the study of very active stellar formation and evolution. For example, the Tarantula Nebula (also called 30 Doradus) is an immense ionized-hydrogen region that contains many young, hot stars. The total mass of 30 Doradus is about one million solar masses, and its diameter is 550 light-years, making it the largest region of ionized gas in the entire Local Group of galaxies. With the Hubble Space Telescope it is possible for astronomers to study the kinds of stars, star clusters, and nebulae that previously could be observed in great detail only in the Milky Way Galaxy.

Quasar

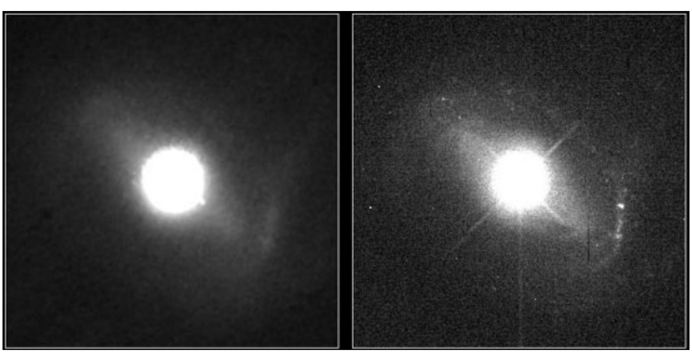
Quasar is an astronomical object of very high luminosity found in the centres of some galaxies and powered by gas spiraling at high velocity into an extremely large black hole. The brightest quasars can outshine all of the stars in the galaxies in which they reside, which makes them visible even at distances of billions of light-years. Quasars are among the most distant and luminous objects known.



Six quasar host galaxies, as observed by the Hubble Space Telescope. Shown are apparently normal, solitary galaxies (left), colliding galaxies (centre), and merging galaxies (right).

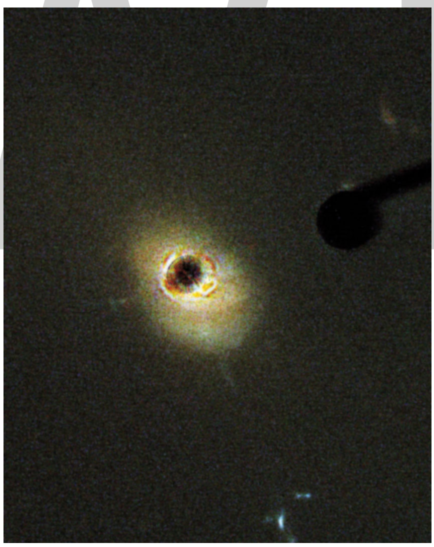
Discovery of Quasars

The term quasar derives from how these objects were originally discovered in the earliest radio surveys of the sky in the 1950s. Away from the plane of the Milky Way Galaxy, most radio sources were identified with otherwise normal-looking galaxies. Some radio sources, however, coincided with objects that appeared to be unusually blue stars, although photographs of some of these objects showed them to be embedded in faint, fuzzy halos. Because of their almost starlike appearance, they were dubbed “quasi-stellar radio sources,” which by 1964 had been shortened to “quasar”.



Quasar 1229+204, as observed by the Hubble Space Telescope. The image above shows that the quasar is surrounded by spiral arms characteristic of galaxies. The tremendous light generated by quasars and their great distance from Earth work to obscure the fainter galactic structures in which they are embedded. This quasar is apparently fueled by a collision between its host galaxy and a dwarf galaxy.

The optical spectra of the quasars presented a new mystery. Photographs taken of their spectra showed locations for emission lines at wavelengths that were at odds with all celestial sources then familiar to astronomers. The puzzle was solved by the Dutch American astronomer Maarten Schmidt, who in 1963 recognized that the pattern of emission lines in 3C 273, the brightest known quasar, could be understood as coming from hydrogen atoms that had a redshift (i.e., had their emission lines shifted toward longer, redder wavelengths by the expansion of the universe) of 0.158. In other words, the wavelength of each line was 1.158 times longer than the wavelength measured in the laboratory, where the source is at rest with respect to the observer. At a redshift of this magnitude, 3C 273 was placed by Hubble's law at a distance of slightly more than two billion light-years. This was a large, though not unprecedented, distance (bright clusters of galaxies had been identified at similar distances), but 3C 273 is about 100 times more luminous than the brightest individual galaxies in those clusters, and nothing so bright had been seen so far away.



Quasar: 3C 273, the brightest quasar. The black region at the centre of the image is blocking light from the central object, revealing the host galaxy of 3C 273.

An even bigger surprise was that continuing observations of quasars revealed that their brightness can vary significantly on timescales as short as a few days, meaning that the total size of the quasar cannot be more than a few light-days across. Since the quasar is so compact and so luminous, the radiation pressure inside the quasar must be huge; indeed, the only way a quasar can keep from blowing itself up with its own radiation is if it is very massive, at least a million solar masses if it is not to exceed the Eddington limit—the minimum mass at which the outward radiation pressure is balanced by the inward pull of gravity. Astronomers were faced with a conundrum: how could an object about the size of the solar system have a mass of about a million stars and outshine by 100 times a galaxy of a hundred billion stars?

The right answer—accretion by gravity onto supermassive black holes—was proposed shortly after Schmidt's discovery independently by Russian astronomers Yakov Zel'dovich and Igor Novikov and Austrian American astronomer Edwin Salpeter. The combination of high luminosities and small sizes was sufficiently unpalatable to some astronomers that alternative explanations were posited that did not require the quasars to be at the large distances implied by their redshifts. These alternative interpretations have been discredited, although a few adherents remain. For most astronomers, the redshift controversy was settled definitively in the early 1980s when American astronomer Todd Boroson and Canadian American astronomer John Beverly Oke showed that the fuzzy halos surrounding some quasars are actually starlight from the galaxy hosting the quasar and that these galaxies are at high redshifts.

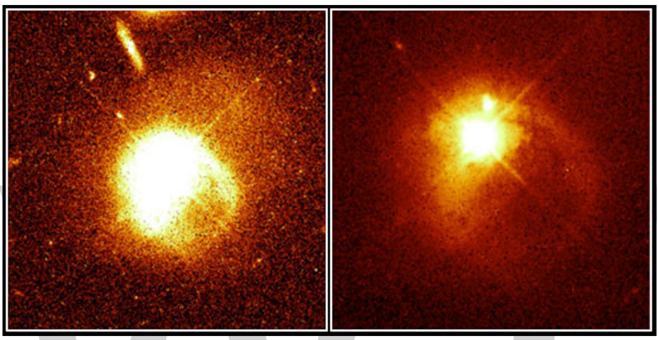
By 1965 it was recognized that quasars are part of a much larger population of unusually blue sources and that most of these are much weaker radio sources too faint to have been detected in the early radio surveys. This larger population, sharing all quasar properties except extreme radio luminosity, became known as "quasi-stellar objects" or simply QSOs. Since the early 1980s most astronomers have regarded QSOs as the high-luminosity variety of an even larger population of "active galactic nuclei," or AGNs.

Finding Quasars

Although the first quasars known were discovered as radio sources, it was quickly realized that quasars could be found more efficiently by looking for objects bluer than normal stars. This can be done with relatively high efficiency by photographing large areas of the sky through two or three different-coloured filters. The photographs are then compared to locate the unusually blue objects, whose nature is verified through subsequent spectroscopy. This remains the primary technique for finding quasars, although it has evolved over the years with the replacement of film by electronic charge-coupled devices (CCDs), the extension of the surveys to longer wavelengths in the infrared, and the addition of multiple filters that, in various combinations, are effective at isolating quasars at different redshifts. Quasars have also been discovered through other techniques, including searches for starlike sources whose brightness varies irregularly and X-ray surveys from space; indeed, a high level of X-ray emission is regarded by astronomers as a sure indicator of an accreting black-hole system.

Physical Structure of Quasars

Quasars and other AGNs are apparently powered by gravitational accretion onto supermassive black holes, where "supermassive" means from roughly a million to a few billion times the mass of the Sun. Supermassive black holes reside at the centres of many large galaxies. In about 5–10 percent of these galaxies, gas tumbles into the deep gravitational well of the black hole and is heated to incandescence as the gas particles pick up speed and pile up in a rapidly rotating "accretion disk" close to the horizon of the black hole. There is a maximum rate set by the Eddington limit at which a black hole can accrete matter before the heating of the infalling gas results in so much outward pressure from radiation that the accretion stops. What distinguishes an "active" galactic nucleus from other galactic nuclei (the 90–95 percent of large galaxies that are currently not quasars) is that the black hole in an active nucleus accretes a few solar masses of matter per year, which, if it is accreting at around 1 percent or more of the Eddington rate, is sufficient to account for a typical quasar with a total luminosity of about 10^{39} watts. (The Sun's luminosity is about 4×10^{26} watts).



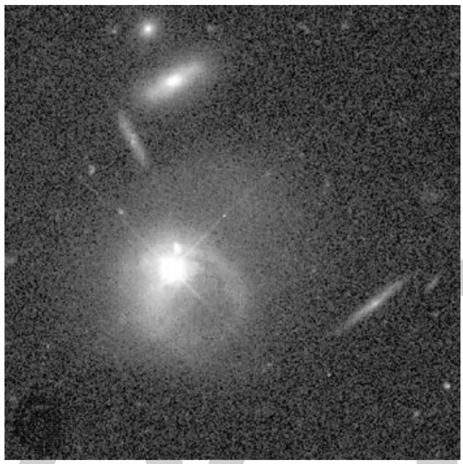
Quasar and its companion galaxy colliding, as observed by the Hubble Space Telescope. The image on the left shows a quasar and (pointing lower right) one of its spiral arms. In the image on the right, the companion galaxy is visible as a bright spot just above the quasar.

In addition to black holes and accretion disks, quasars have other remarkable features. Just beyond the accretion disk are clouds of gas that move at high velocities around the inner structure, absorbing high-energy radiation from the accretion disk and reprocessing it into the broad emission lines of hydrogen and ions of other atoms that are the signatures of quasar spectra. Farther from the black hole but still largely in the accretion disk plane are dust-laden gas clouds that can obscure the quasar itself. Some quasars are also observed to have radio jets, which are highly collimated beams of plasma propelled out along the rotation axis of the accretion disk at speeds often approaching that of light. These jets emit beams of radiation that can be observed at X-ray and radio wavelengths (and less often at optical wavelengths).

Because of this complex structure, the appearance of a quasar depends on the orientation of the rotation axis of the accretion disk relative to the observer's line of sight. Depending on this angle, different quasar components—the accretion disk, emission-line clouds, jets—appear to be more or less prominent. This results in a wide variety of observed phenomena from what are, in reality, physically similar sources.

Evolution of Quasars

The number density of quasars increases dramatically with redshift, which translates through Hubble's law to more quasars at larger distances. Because of the finite speed of light, when quasars are observed at great distances, they are observed as they were in the distant past. Thus, the increasing density of quasars with distance means that they were more common in the past than they are now. This trend increases until "look-back times" that correspond to around three billion years after the big bang, which occurred approximately 13.5 billion years ago. At earlier ages, the number density of quasars decreases sharply, corresponding to an era when the quasar population was still building up. The most distant, and thus earliest, quasars known were formed less than a billion years after the big bang.



The quasar PKS 2349–014 (bright central disk), several billion light-years from Earth. It appears to be merging with a companion galaxy.

Individual quasars appear as their central black holes begin to accrete gas at a high rate, possibly triggered by a merger with another galaxy, building up the mass of the central black hole. The current best estimate is that quasar activity is episodic, with individual episodes lasting around a million years and the total quasar lifetime lasting around 10 million years. At some point, quasar activity ceases completely, leaving behind the dormant massive black holes found in most massive galaxies. This "life cycle" appears to proceed most rapidly with the most-massive black holes, which become dormant earlier than less-massive black holes. Indeed, in the current universe the remaining AGN population is made up predominantly of lower-luminosity Seyfert galaxies with relatively small supermassive black holes.

In the present-day universe there is a close relationship between the mass of a black hole and the mass of its host galaxy. This is quite remarkable, since the central black hole accounts for only about 0.1 percent of the mass of the galaxy. It is believed that the intense radiation, mass outflows, and jets from the black hole during its active quasar phase are responsible. The radiation, outflows, and jets heat up and can even remove entirely the interstellar medium from the host galaxy. This loss of gas in the galaxy simultaneously shuts down star formation and chokes off the quasar's fuel supply, thus freezing both the mass in stars and the mass of the black hole.

References

- Galaxy, science: britannica.com, Retrieved 17 June, 2019
- Milky-Way-Galaxy, place: britannica.com, Retrieved 08 April, 2019
- Saxena A.; Marinello M.; Overzier R.A.; Best P.N.; et al. (2018). "Discovery of a radio galaxy at z = 5.72". Monthly Notices of the Royal Astronomical Society. 480: 2733. arXiv:1806.01191. Bibcode:2018MNRAS.480.2733S. doi:10.1093/mnras/sty1996
- Francis, Charles; Anderson, Erik (June 2014). "Two estimates of the distance to the Galactic Centre". Monthly Notices of the Royal Astronomical Society. 441 (2): 1105–1114. arXiv:1309.2629. Bibcode:2014MNRAS.441.1105F. doi:10.1093/mnras/stu631
- Andromeda-galaxy-m31- 15590: space.com, Retrieved 27 May, 2019
- Liu, Chengze; et al. (November 2015). "The Next Generation Virgo Cluster Survey. X: Properties of Ultracompact Dwarfs in The M87, M49 and M60 Regions". Astrophysical Journal. 812 (1): 34. arXiv:1508.07334. Bibcode:2015ApJ...812...34L. doi:10.1088/0004-637X/812/1/34
- Active-galaxies, science-objects: imagine.gsfc.nasa.gov, Retrieved 25 August, 2019
- Francis, C.; Anderson, E. (2009). "Galactic spiral structure". Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 465 (2111): 3425–3446. arXiv:0901.3503. Bibcode:2009RSPSA.465.3425F. doi:10.1098/rspa.2009.0036

Chapter 6

The Universe

The universe is a combination of space and time. It includes planets, stars, galaxies, and all other forms of matter and energy. Early universe, Friedmann universe, mass, shape, size and density of universe, Hubble's law, steady state theory, cosmological red-shift, etc. are some of the concepts studied to better understand our universe. All these concepts have been carefully analyzed in this chapter.

Universe is the whole cosmic system of matter and energy of which Earth, and therefore the human race, is a part. Humanity has travelled a long road since societies imagined Earth, the Sun, and the Moon as the main objects of creation, with the rest of the universe being formed almost as an afterthought. Today it is known that Earth is only a small ball of rock in a space of unimaginable vastness and that the birth of the solar system was probably only one event among many that occurred against the backdrop of an already mature universe. This humbling topic has unveiled a remarkable fact, one that endows the minutest particle in the universe with a rich and noble heritage: events that occurred in the first few minutes of the creation of the universe 13.7 billion years ago turn out to have had a profound influence on the birth, life, and death of galaxies, stars, and planets. Indeed, a line can be drawn from the forging of the matter of the universe in a primal "big bang" to the gathering on Earth of atoms versatile enough to serve as the basis of life.

The universe includes living things, planets, stars, galaxies, dust clouds, light, and even time. Before the birth of the Universe, time, space and matter did not exist. The Universe contains billions of galaxies, each containing millions or billions of stars. The space between the stars and galaxies is largely empty. However, even places far from stars and planets contain scattered particles of dust or a few hydrogen atoms per cubic centimetre. Space is also filled with radiation (e.g. light and heat), magnetic fields and high energy particles (e.g. cosmic rays).

The Universe is incredibly huge. It would take a modern jet fighter more than a million years to reach the nearest star to the Sun. Travelling at the speed of light (300,000 km per second), it would take 100,000 years to cross our Milky Way galaxy alone.

No one knows the exact size of the Universe, because we cannot see the edge – if there is one. All we do know is that the visible Universe is at least 93 billion light years across. (A light year is the distance light travels in one year – about 9 trillion km.)

The Universe has not always been the same size. Scientists believe it began in a Big

Bang, which took place nearly 14 billion years ago. Since then, the Universe has been expanding outward at very high speed. So the area of space we now see is billions of times bigger than it was when the Universe was very young. The galaxies are also moving further apart as the space between them expands.

Expansion of the Universe

The expansion of the universe is the increase in distance between any two given gravitationally unbound parts of the observable universe with time. It is an intrinsic expansion whereby *the scale of space itself changes*. The universe does not expand "into" anything and does not require space to exist "outside" it. Technically, neither space nor objects in space move. Instead it is the metric governing the size and geometry of spacetime itself that changes in scale. Although light and objects within space time cannot travel faster than the speed of light, this limitation does not restrict the metric itself. To an observer it appears that space is expanding and all but the nearest galaxies are receding into the distance.

During the inflationary epoch about 10^{-32} of a second after the Big Bang, the universe suddenly expanded, and its volume increased by a factor of at least 10^{78} (an expansion of distance by a factor of at least 10^{26} in each of the three dimensions), equivalent to expanding an object 1 nanometer (10^{-9} m, about half the width of a molecule of DNA) in length to one approximately 10.6 light years (about 10^{17} m or 62 trillion miles) long. A much slower and gradual expansion of space continued after this, until at around 9.8 billion years after the Big Bang (4 billion years ago) it began to gradually expand more quickly, and is still doing so.

The metric expansion of space is of a kind completely different from the expansions and explosions seen in daily life. It also seems to be a property of the universe as a whole rather than a phenomenon that applies just to one part of the universe or can be observed from "outside" it.

Metric expansion is a key feature of Big Bang cosmology, is modelled mathematically with the Friedmann-Lemaître-Robertson-Walker metric and is a generic property of the universe we inhabit. However, the model is valid only on large scales (roughly the scale of galaxy clusters and above), because gravitational attraction binds matter together strongly enough that metric expansion cannot be observed at this time, on a smaller scale. As such, the only galaxies receding from one another as a result of metric expansion are those separated by cosmologically relevant scales larger than the length scales associated with the gravitational collapse that are possible in the age of the universe given the matter density and average expansion rate.

Physicists have postulated the existence of dark energy, appearing as a cosmological constant in the simplest gravitational models, as a way to explain the acceleration. According to the simplest extrapolation of the currently-favoured cosmological model, the Lambda-CDM model, this acceleration becomes more dominant into the future. In June 2016, NASA and ESA scientists reported that the universe was found to be expanding 5% to 9% faster than thought earlier, based on studies using the Hubble Space Telescope.

While special relativity prohibits objects from moving faster than light with respect to a local reference frame where space time can be treated as flat and unchanging, it does not apply to situations where spacetime curvature or evolution in time become important. These situations are described by general relativity, which allows the separation between two distant objects to increase faster than the speed of light, although the definition of "separation" is different from that used in an inertial frame. This can be seen when observing distant galaxies more than the Hubble radius away from us (approximately 4.5 gigaparsecs or 14.7 billion light-years); these galaxies have a recession speed that is faster than the speed of light. Light that is emitted today from galaxies beyond the cosmological event horizon, about 5 gigaparsecs or 16 billion light-years, will never reach us, although we can still see the light that these galaxies emitted in the past. Because of the high rate of expansion, it is also possible for a distance between two objects to be greater than the value calculated by multiplying the speed of light by the age of the universe. These details are a frequent source of confusion among amateurs and even professional physicists. Due to the non-intuitive nature of the subject and what has been described by some as "careless" choices of wording, certain descriptions of the metric expansion of space and the misconceptions to which such descriptions can lead are an on-going subject of discussion within education and communication of scientific concepts.

In 1912, Vesto Slipher discovered that light from remote galaxies was redshifted, which was later interpreted as galaxies receding from the Earth. In 1922, Alexander Friedmann used Einstein field equations to provide theoretical evidence that the universe is expanding. In 1927, Georges Lemaître independently reached a similar conclusion to Friedmann on a theoretical basis, and also presented the first observational evidence for a linear relationship between distance to galaxies and their recessional velocity. Edwin Hubble observationally confirmed Lemaître's findings two years later. Assuming the cosmological principle, these findings would imply that all galaxies are moving away from each other.

Based on large quantities of experimental observation and theoretical work, the scientific consensus is that space itself is expanding, and that it expanded very rapidly within the first fraction of a second after the Big Bang. This kind of expansion is known as "metric expansion". In mathematics and physics, a "metric" means a measure of distance, and the term implies that the sense of distance within the universe is itself changing.

Overview of Metrics and Comoving Coordinates

To understand the metric expansion of the universe, it is helpful to discuss briefly what a metric is, and how metric expansion works.

A metric defines the concept of distance, by stating in mathematical terms how distances between two nearby points in space are measured, in terms of the coordinate system. Coordinate systems locate points in a space (of whatever number of dimensions) by assigning unique positions on a grid, known as coordinates, to each point. Latitude and longitude, and x-y graphs are common examples of coordinates. A metric is a formula which describes how a number known as "distance" is to be measured between two points.

It may seem obvious that distance is measured by a straight line, but in many cases it is not. For example, long haul aircraft travel along a curve known as a "great circle" and not a straight line, because that is a better metric for air travel. (A straight line would go through the earth). Another example is planning a car journey, where one might want the shortest journey in terms of travel time - in that case a straight line is a poor choice of metric because the shortest distance by road is not normally a straight line, and even the path nearest to a straight line will not necessarily be the quickest. A final example is the internet, where even for nearby towns, the quickest route for data can be via major connections that go across the country and back again. In this case the metric used will be the shortest time that data takes to travel between two points on the network.

In cosmology, we cannot use a ruler to measure metric expansion, because our ruler internal forces easily overcome the extremely slow expansion of space leaving the ruler intact. Also any objects on or near earth that we might measure are being held together or pushed apart by several forces which are far larger in their effects. So even if we could measure the tiny expansion that is still happening, we would not notice the change on a small scale or in everyday life. On a large intergalactic scale, we can use other tests of distance and these do show that space is expanding, even if a ruler on earth could not measure it.

The metric expansion of space is described using the mathematics of metric tensors. The coordinate system we use is called "comoving coordinates", a type of coordinate system which takes account of time as well as space and the speed of light, and allows us to incorporate the effects of both general and special relativity.

For example, consider the measurement of distance between two places on the surface of the Earth. This is a simple, familiar example of spherical geometry. Because the surface of the Earth is two-dimensional, points on the surface of the Earth can be specified by two coordinates — for example, the latitude and longitude. Specification of a metric requires that one first specify the coordinates used. In our simple example of the surface of the Earth, we could choose any kind of coordinate system we wish, for example latitude and longitude, or X-Y-Z Cartesian coordinates. Once we have chosen a specific coordinate system, the numerical values of the coordinates of any two points are uniquely determined, and based upon the properties of the space being discussed, the appropriate metric is mathematically established too. On the curved surface of the Earth, we can see this effect in long-haul airline flights where the distance between two points is measured based upon a great circle, rather than the straight line one might plot on a two-dimensional map of the Earth's surface. In general, such shortest-distance paths are called "geodesics". In Euclidean geometry, the geodesic is a straight line, while in non-Euclidean geometry such as on the Earth's surface, this is not the case. Indeed, even the shortest-distance great circle path is always longer than the Euclidean straight line path which passes through the interior of the Earth. The difference between the straight line path and the shortest-distance great circle path is due to the curvature of the Earth's surface. While there is always an effect due to this curvature, at short distances the effect is small enough to be unnoticeable.

On plane maps, great circles of the Earth are mostly not shown as straight lines. Indeed, there is a seldom-used map projection, namely the gnomonic projection, where all great circles are shown as straight lines, but in this projection, the distance scale varies very much in different areas. There is no map projection in which the distance between any two points on Earth, measured along the great circle geodesics, is directly proportional to their distance on the map; such accuracy is possible only with a globe.

Metric Tensors

In differential geometry, the backbone mathematics for general relativity, a metric tensor can be defined which precisely characterizes the space being described by explaining the way distances should be measured in every possible direction. General relativity necessarily invokes a metric in four dimensions (one of time, three of space) because, in general, different reference frames will experience different intervals of time and space depending on the inertial frame. This means that the metric tensor in general relativity relates precisely how two events in spacetime are separated. A metric expansion occurs when the metric tensor changes with time (and, specifically, whenever the spatial part of the metric gets larger as time goes forward). This kind of expansion is different from all kinds of expansions and explosions commonly seen in nature in no small part because times and distances are not the same in all reference frames, but are instead subject to change. A useful visualization is to approach the subject rather than objects in a fixed "space" moving apart into "emptiness", as space itself growing between objects without any acceleration of the objects themselves. The space between objects shrinks or grows as the various geodesics converge or diverge.

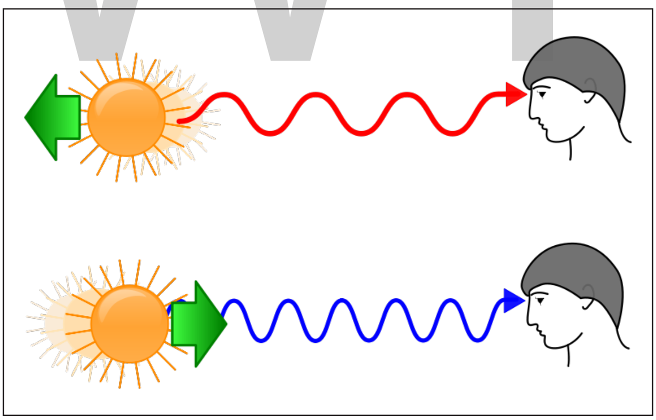
Because this expansion is caused by relative changes in the distance-defining metric, this expansion (and the resultant movement apart of objects) is not restricted by the speed of light upper bound of special relativity. Two reference frames that are globally separated can be moving apart faster than light without violating special relativity, although whenever two reference frames diverge from each other faster than the speed of light, there will be observable effects associated with such situations including the existence of various cosmological horizons.

Theory and observations suggest that very early in the history of the universe, there was an inflationary phase where the metric changed very rapidly, and that the remaining time-dependence of this metric is what we observe as the so-called Hubble expansion, the moving apart of all gravitationally unbound objects in the universe. The expanding universe is therefore a fundamental feature of the universe we inhabit — a universe fundamentally different from the static universe Albert Einstein first considered when he developed his gravitational theory.

Comoving Coordinates

In expanding space, proper distances are dynamical quantities which change with time. An easy way to correct for this is to use comoving coordinates which remove this feature and allow for a characterization of different locations in the universe without having to characterize the physics associated with metric expansion. In comoving coordinates, the distances between all objects are fixed and the instantaneous dynamics of matter and light are determined by the normal physics of gravity and electromagnetic radiation. Any time-evolution however must be accounted for by taking into account the Hubble law expansion in the appropriate equations in addition to any other effects that may be operating (gravity, dark energy, or curvature, for example). Cosmological simulations that run through significant fractions of the universe's history therefore must include such effects in order to make applicable predictions for observational cosmology.

Measurement of Expansion and Change of Rate of Expansion



When an object is receding, its light gets stretched (redshifted).

In principle, the expansion of the universe could be measured by taking a standard ruler and measuring the distance between two cosmologically distant points, waiting a certain time, and then measuring the distance again, but in practice, standard rulers are not easy to find on cosmological scales and the timescales over which a measurable expansion would be visible are too great to be observable even by multiple generations of humans. The expansion of space is measured indirectly. The theory of relativity predicts phenomena associated with the expansion, notably the redshift-versus-distance relationship known as Hubble's Law; functional forms for cosmological distance measurements that differ from what would be expected if space were not expanding; and an observable change in the matter and energy density of the universe seen at different lookback times.

The first measurement of the expansion of space came with Hubble's realization of the velocity vs. redshift relation. Most recently, by comparing the apparent brightness of distant standard candles to the redshift of their host galaxies, the expansion rate of the universe has been measured to be $H_o = 73.24 \pm 1.74$ (km/s)/Mpc. This means that for every million parsecs of distance from the observer, the light received from that distance is cosmologically redshifted by about 73 kilometres per second (160,000 mph). On the other hand, by assuming a cosmological model, e.g. Lambda-CDM model, one can infer the Hubble constant from the size of the largest fluctuations seen in the Cosmic Microwave Background. A higher Hubble constant would imply a smaller characteristic size of CMB fluctuations, and vice versa. The Planck collaboration measure the expansion rate this way and determine $H_o = 67.4 \pm 0.5$ (km/s)/Mpc. There is a disagreement between the two measurements, the distance ladder being model-independent and the CMB measurement depending on the fitted model, which hints at new physics beyond our standard cosmological models.

The Hubble parameter is not thought to be constant through time. There are dynamical forces acting on the particles in the universe which affect the expansion rate. It was earlier expected that the Hubble parameter would be decreasing as time went on due to the influence of gravitational interactions in the universe, and thus there is an additional observable quantity in the universe called the deceleration parameter which cosmologists expected to be directly related to the matter density of the universe. Surprisingly, the deceleration parameter was measured by two different groups to be less than zero (actually, consistent with –1) which implied that today the Hubble parameter is converging to a constant value as time goes on. Some cosmologists have whimsically called the effect associated with the "accelerating universe" the "cosmic jerk". The 2011 Nobel Prize in Physics was given for the discovery of this phenomenon.

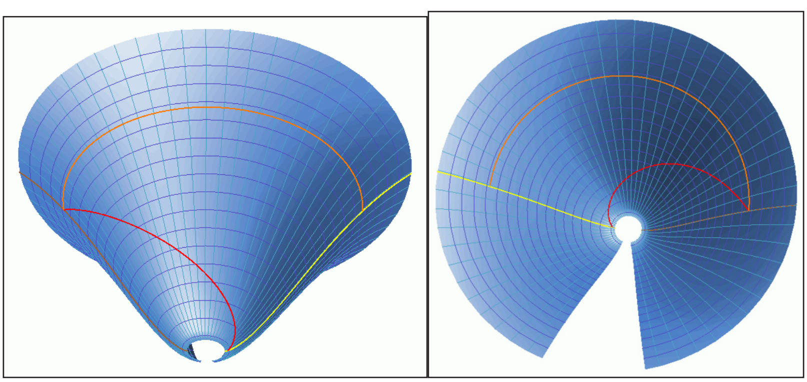
In October 2018, scientists presented a new third way (two earlier methods, one based on redshifts and another on the cosmic distance ladder, gave results that do not agree), using information from gravitational wave events (especially those involving the merger of neutron stars, like GW170817), of determining the Hubble Constant, essential in establishing the rate of expansion of the universe.

Measuring Distances in Expanding Space

At cosmological scales the present universe is geometrically flat, which is to say that the rules of Euclidean geometry associated with Euclid's fifth postulate hold, though in the past spacetime could have been highly curved. In part to accommodate such different geometries, the expansion of the universe is inherently general relativistic; it cannot be modeled with special relativity alone, though such models exist, they are at fundamental odds with the observed interaction between matter and spacetime seen in our universe.

The images below show two views of spacetime diagrams that show the large-scale geometry of the universe according to the ΛCDM cosmological model. Two of the dimensions of space are omitted, leaving one dimension of space (the dimension that grows as the cone gets larger) and one of time (the dimension that proceeds "up" the cone's surface). The narrow circular end of the diagram corresponds to a cosmological time of 700 million years after the big bang while the wide end is a cosmological time of 18 billion years, where one can see the beginning of the accelerating expansion as a splaying outward of the spacetime, a feature which eventually dominates in this model. The purple grid lines mark off cosmological time at intervals of one billion years from the big bang. The cyan grid lines mark off comoving distance at intervals of one billion light years in the present era (less in the past and more in the future). Note that the circular curling of the surface is an artifact of the embedding with no physical significance and is done purely to make the illustration viewable; space does not actually curl around on itself. (A similar effect can be seen in the tubular shape of the pseudosphere).

The brown line on the diagram is the worldline of the Earth (or, at earlier times, of the matter which condensed to form the Earth). The yellow line is the worldline of the most distant known quasar. The red line is the path of a light beam emitted by the quasar about 13 billion years ago and reaching the Earth in the present day. The orange line shows the present-day distance between the quasar and the Earth, about 28 billion light years, which is, notably, a larger distance than the age of the universe multiplied by the speed of light.



Two views of an isometric embedding of part of the visible universe over most of its history, showing how a light ray (red line) can travel an effective distance of 28 billion light years (orange line) in just 13 billion years of cosmological time.

According to the equivalence principle of general relativity, the rules of special relativity are locally valid in small regions of spacetime that are approximately flat. In particular, light always travels locally at the speed *c*; in our diagram, this means, according to the convention of constructing spacetime diagrams, that light beams always make an angle of 45° with the local grid lines. It does not follow, however, that light travels a distance *ct* in a time *t*, as the red worldline illustrates. While it always moves locally at *c*, its time in transit (about 13 billion years) is not related to the distance traveled in any simple way since the universe expands as the light beam traverses space and time. In fact the distance traveled is inherently ambiguous because of the changing scale of the universe. Nevertheless, we can single out two distances which appear to be physically meaningful: the distance between the Earth and the quasar when the light was emitted, and the distance between them in the present era (taking a slice of the cone along the dimension that we've declared to be the spatial dimension). The former distance is about 4 billion light years, much smaller than *ct* because the universe expanded as the light traveled the distance, the light had to "run against the treadmill" and therefore went farther than the initial separation between the Earth and the quasar. The latter distance (shown by the orange line) is about 28 billion light years, much larger than *ct*. If expansion could be instantaneously stopped today, it would take 28 billion years for light to travel between the Earth and the quasar while if the expansion had stopped at the earlier time, it would have taken only 4 billion years.

The light took much longer than 4 billion years to reach us though it was emitted from only 4 billion light years away, and, in fact, the light emitted towards the Earth was actually moving away from the Earth when it was first emitted, in the sense that the metric distance to the Earth increased with cosmological time for the first few billion years of its travel time, and also indicating that the expansion of space between the Earth and the quasar at the early time was faster than the speed of light. None of this surprising behavior originates from a special property of metric expansion, but simply from local principles of special relativity integrated over a curved surface.

Topology of Expanding Space

Over time, the space that makes up the universe is expanding. The words 'space' and 'universe', sometimes used interchangeably, have distinct meanings in this context. Here 'space' is a mathematical concept that stands for the three-dimensional manifold into which our respective positions are embedded while 'universe' refers to everything that exists including the matter and energy in space, the extra-dimensions that may be wrapped up in various strings, and the time through which various events take place. The expansion of space is in reference to this 3-D manifold only; that is, the description involves no structures such as extra dimensions or an exterior universe.

The ultimate topology of space is *a posteriori* — something which in principle must be observed — as there are no constraints that can simply be reasoned out (in other words there can not be any *a priori* constraints) on how the space in which we live is connected or whether it wraps around on itself as a compact space. Though certain cosmological models such as Gödel's universe even permit bizarre worldlines which intersect with themselves, ultimately the question as to whether we are in something like a "Pac-Man universe" where if traveling far enough in one direction would allow one to simply end up back in the same place like going all the way around the surface of a balloon (or a planet like the Earth) is an observational question which is constrained as measurable or non-measurable by the universe's global geometry. At present, observations are consistent with the universe being infinite in extent and simply connected, though we are limited in distinguishing between simple and more complicated proposals by cosmological horizons. The universe could be infinite in extent or it could be finite; but the evidence that leads to the inflationary model of the early universe also implies that the "total universe" is much larger than the observable universe, and so any edges or exotic geometries or topologies would not be directly observable as light has not reached scales on which such aspects of the universe, if they exist, are still allowed. For all intents and purposes, it is safe to assume that the universe is infinite in spatial extent, without edge or strange connectedness.

Regardless of the overall shape of the universe, the question of what the universe is expanding into is one which does not require an answer according to the theories which describe the expansion; the way we define space in our universe in no way requires additional exterior space into which it can expand since an expansion of an infinite expanse can happen without changing the infinite extent of the expanse. All that is certain is that the manifold of space in which we live simply has the property that the distances between objects are getting larger as time goes on. This only implies the simple observational consequences associated with the metric expansion explored below. No "outside" or embedding in hyperspace is required for an expansion to occur. The visualizations often seen of the universe growing as a bubble into nothingness are misleading in that respect. There is no reason to believe there is anything "outside" of the expanding universe into which the universe expands.

Even if the overall spatial extent is infinite and thus the universe cannot get any "larger", we still say that space is expanding because, locally, the characteristic distance between objects is increasing. As an infinite space grows, it remains infinite.

Density of Universe during Expansion

Despite being extremely dense when very young and during part of its early expansion - far denser than is usually required to form a black hole - the universe did not re collapse into a black hole. This is because commonly-used calculations for gravitational collapse are usually based upon objects of relatively constant size, such as stars, and do not apply to rapidly expanding space such as the Big Bang.

Effects of Expansion on Small Scales

The expansion of space is sometimes described as a force which acts to push objects apart. Though this is an accurate description of the effect of the cosmological constant, it is not an accurate picture of the phenomenon of expansion in general. For much of the universe's history the expansion has been due mainly to inertia. The matter in the very early universe was flying apart for unknown reasons (most likely as a result of cosmic inflation) and has simply continued to do so, though at an ever-decreasing rate due to the attractive effect of gravity.

In addition to slowing the overall expansion, gravity causes local clumping of matter into stars and galaxies. Once objects are formed and bound by gravity, they "drop out" of the expansion and do not subsequently expand under the influence of the cosmological metric, there being no force compelling them to do so.

There is no difference between the inertial expansion of the universe and the inertial separation of nearby objects in a vacuum; the former is simply a large-scale extrapolation of the latter.

Once objects are bound by gravity, they no longer recede from each other. Thus, the Andromeda galaxy, which is bound to the Milky Way galaxy, is actually falling *towards* us and is not expanding away. Within the Local Group, the gravitational interactions have changed the inertial patterns of objects such that there is no cosmological expansion taking place. Once one goes beyond the Local Group, the inertial expansion is measurable, though systematic gravitational effects imply that larger and larger parts of space will eventually fall out of the "Hubble Flow" and end up as bound, non-expanding objects up to the scales of superclusters of galaxies. We can predict such future events by knowing the precise way the Hubble Flow is changing as well as the masses of the objects to which we are being gravitationally pulled. Currently, the Local Group is being gravitationally pulled towards either the Shapley Supercluster or the "Great Attractor" with which, if dark energy were not acting, we would eventually merge and no longer see expand away from us after such a time.

A consequence of metric expansion being due to inertial motion is that a uniform local "explosion" of matter into a vacuum can be locally described by the FLRW geometry, the same geometry which describes the expansion of the universe as a whole and was also the basis for the simpler Milne universe which ignores the effects of gravity. In particular, general relativity predicts that light will move at the speed *c* with respect to the local motion of the exploding matter, a phenomenon analogous to frame dragging.

The situation changes somewhat with the introduction of dark energy or a cosmological constant. A cosmological constant due to a vacuum energy density has the effect of adding a repulsive force between objects which is proportional (not inversely proportional) to distance. Unlike inertia it actively "pulls" on objects which have clumped together under the influence of gravity, and even on individual atoms. However, this does not cause the objects to grow steadily or to disintegrate; unless they are very weakly bound, they will simply settle into an equilibrium state which is slightly (undetectably) larger than it would otherwise have been. As the universe expands and the matter in it thins, the gravitational attraction decreases (since it is proportional to the density), while the cosmological repulsion increases; thus the ultimate fate of the ΛCDM universe is a near vacuum expanding at an ever-increasing rate under the influence of the cosmological constant. However, the only locally visible effect of the accelerating expansion is the disappearance (by runaway redshift) of distant galaxies; gravitationally bound objects like the Milky Way do not expand and the Andromeda galaxy is moving fast enough towards us that it will still merge with the Milky Way in 3 billion years time, and it is also likely that the merged supergalaxy that forms will eventually fall in and merge with the nearby Virgo Cluster. However, galaxies lying farther away from this will recede away at ever-increasing speed and be redshifted out of our range of visibility.

Metric Expansion and Speed of Light

At the end of the early universe's inflationary period, all the matter and energy in the universe was set on an inertial trajectory consistent with the equivalence principle and Einstein's general theory of relativity and this is when the precise and regular form of the universe's expansion had its origin (that is, matter in the universe is separating because it was separating in the past due to the inflaton field) .

While special relativity prohibits objects from moving faster than light with respect to a local reference frame where spacetime can be treated as flat and unchanging, it does not apply to situations where spacetime curvature or evolution in time become important. These situations are described by general relativity, which allows the separation between two distant objects to increase faster than the speed of light, although the definition of "distance" here is somewhat different from that used in an inertial frame. The definition of distance used here is the summation or integration of local comoving distances, all done at constant local proper time. For example, galaxies that are more than the Hubble radius, approximately 4.5 gigaparsecs or 14.7 billion light-years, away from us have a recession speed that is faster than the speed of light. Visibility of these objects depends on the exact expansion history of the universe. Light that is emitted today from galaxies beyond the cosmological event horizon, about 5 gigaparsecs or 16 billion light-years, will never reach us, although we can still see the light that these galaxies emitted in the past.

Scale Factor

At a fundamental level, the expansion of the universe is a property of spatial measurement on the largest measurable scales of our universe. The distances between cosmologically relevant points increases as time passes leading to observable effects outlined below. This feature of the universe can be characterized by a single parameter that is called the scale factor which is a function of time and a single value for all of space at any instant (if the scale factor were a function of space, this would violate the cosmological principle). By convention, the scale factor is set to be unity at the present time and, because the universe is expanding, is smaller in the past and larger in the future.

Extrapolating back in time with certain cosmological models will yield a moment when the scale factor was zero; our current understanding of cosmology sets this time at 13.799 ± 0.021 billion years ago. If the universe continues to expand forever, the scale factor will approach infinity in the future. In principle, there is no reason that the expansion of the universe must be monotonic and there are models where at some time in the future the scale factor decreases with an attendant contraction of space rather than an expansion.

Other Conceptual Models of Expansion

The expansion of space is often illustrated with conceptual models which show only the size of space at a particular time, leaving the dimension of time implicit.

In the "ant on a rubber rope model" one imagines an ant (idealized as pointlike) crawling at a constant speed on a perfectly elastic rope which is constantly stretching. If we stretch the rope in accordance with the ΛCDM scale factor and think of the ant's speed as the speed of light, then this analogy is numerically accurate.

In the "rubber sheet model" one replaces the rope with a flat two-dimensional rubber sheet which expands uniformly in all directions. The addition of a second spatial dimension raises the possibility of showing local perturbations of the spatial geometry by local curvature in the sheet.

In the "balloon model" the flat sheet is replaced by a spherical balloon which is inflated from an initial size of zero (representing the big bang). A balloon has positive Gaussian curvature while observations suggest that the real universe is spatially flat, but this inconsistency can be eliminated by making the balloon very large so that it is locally flat to within the limits of observation. This analogy is potentially confusing since it wrongly suggests that the big bang took place at the center of the balloon. In fact points off the surface of the balloon have no meaning, even if they were occupied by the balloon at an earlier time.

In the "raisin bread model" one imagines a loaf of raisin bread expanding in the oven. The loaf (space) expands as a whole, but the raisins (gravitationally bound objects) do not expand; they merely grow farther away from each other.

Observational Evidence

Theoretical cosmologists developing models of the universe have drawn upon a small number of reasonable assumptions in their work. These workings have led to models in which the metric expansion of space is a likely feature of the universe. Chief among the underlying principles that result in models including metric expansion as a feature are:

- The Cosmological Principle which demands that the universe looks the same way in all directions (isotropic) and has roughly the same smooth mixture of material (homogeneous).

- The Copernican Principle which demands that no place in the universe is preferred (that is, the universe has no "starting point").

Scientists have tested carefully whether these assumptions are valid and borne out by observation. Observational cosmologists have discovered evidence — very strong in some cases — that supports these assumptions, and as a result, metric expansion of space is considered by cosmologists to be an observed feature on the basis that although we cannot see it directly, scientists have tested the properties of the universe and observation provides compelling confirmation. Sources of this confidence and confirmation include:

- Hubble demonstrated that all galaxies and distant astronomical objects were moving away from us, as predicted by a universal expansion. Using the redshift of their electromagnetic spectra to determine the distance and speed of remote objects in space, he showed that all objects are moving away from us, and that their speed is proportional to their distance, a feature of metric expansion. Further studies have since shown the expansion to be highly isotropic and homogeneous, that is, it does not seem to have a special point as a "center", but appears universal and independent of any fixed central point.

- In studies of large-scale structure of the cosmos taken from redshift surveys a so-called "End of Greatness" was discovered at the largest scales of the universe. Until these scales were surveyed, the universe appeared "lumpy" with clumps of galaxy clusters, superclusters and filaments which were anything but isotropic and homogeneous. This lumpiness disappears into a smooth distribution of galaxies at the largest scales.

- The isotropic distribution across the sky of distant gamma-ray bursts and supernovae is another confirmation of the Cosmological Principle.

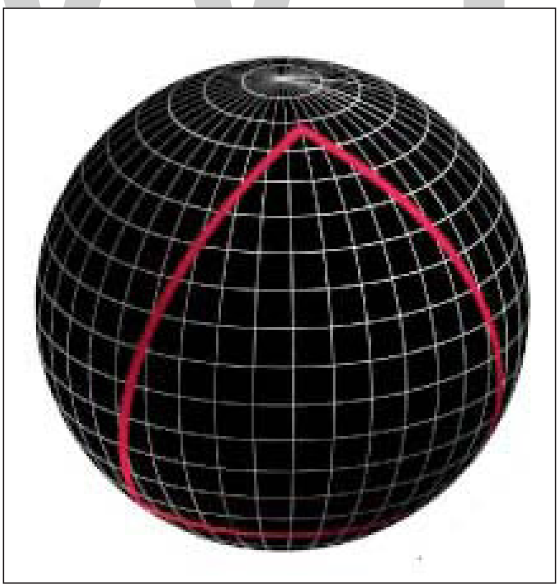
- The Copernican Principle was not truly tested on a cosmological scale until measurements of the effects of the cosmic microwave background radiation on the dynamics of distant astrophysical systems were made. A group of astronomers at the European Southern Observatory noticed, by measuring the temperature of a distant intergalactic cloud in thermal equilibrium with the cosmic microwave background, that the radiation from the Big Bang was demonstrably warmer at earlier times. Uniform cooling of the cosmic microwave background over billions of years is strong and direct observational evidence for metric expansion.

Taken together, these phenomena overwhelmingly support models that rely on space expanding through a change in metric. It was not until the discovery in the year 2000 of direct observational evidence for the changing temperature of the cosmic microwave background that more bizarre constructions could be ruled out. Until that time, it was based purely on an assumption that the universe did not behave as one with the Milky Way sitting at the middle of a fixed-metric with a universal explosion of galaxies in all directions (as seen in, for example, an early model proposed by Milne). Yet before this evidence, many rejected the Milne viewpoint based on the mediocrity principle.

More direct results of the expansion, such as change of redshift, distance, flux, angular position and the angular size of astronomical objects, have not been detected yet due to smallness of these effects. Change of the redshift or the flux could be observed by Square Kilometre Array or Extremely Large Telescope in the mid-2030s.

Shape of the Universe

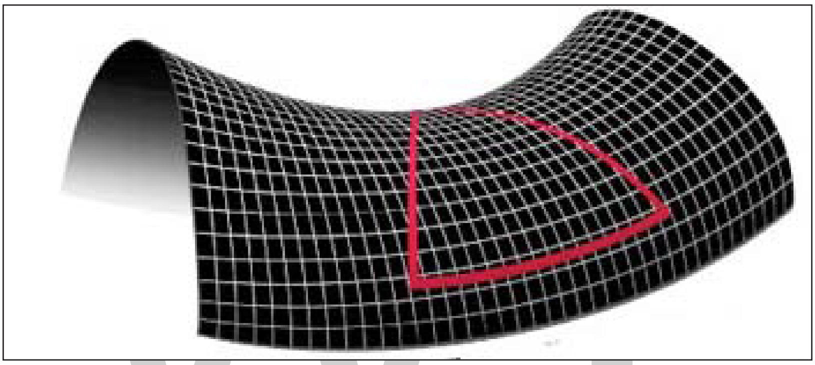
According to Einstein's general theory of relativity massive objects curve space (see here to find out more). If we assume that the Universe is isotropic (looks the same, on average, in every direction in the sky) and homogeneous (looks the same, on average, at every point), and that there is none of that mysterious substance called *dark energy*, then general relativity tells us that there are only three possibilities for the shape of the Universe, depending on how much matter there is within it.



The surface of a sphere is finite, but it doesn't have an edge. The angles of a spherical triangle add up to more than 180 degrees. If the Universe has positive curvature, then its geometry is the 3D analogue of a sphere.

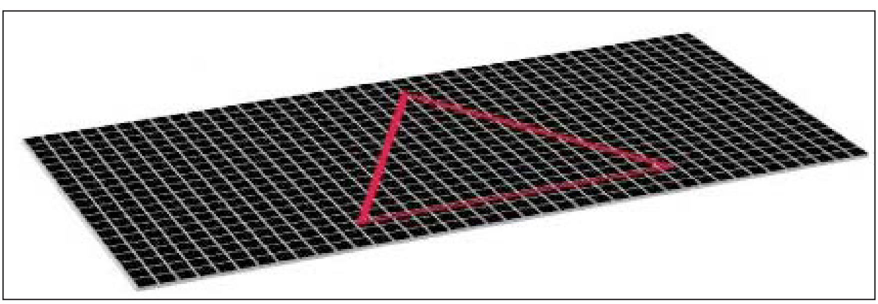
The first possibility is that the density of matter (the average amount of matter per unit volume) is so high; it curves the Universe around on itself to form the 3D analogue of a sphere. This 3D sphere is hard to imagine, but it's easily described mathematically and comes with a positive value for its curvature. If the Universe is indeed positively curved, then it is finite, just like the surface of a sphere is. What is more the gravitational pull exerted by all the mass will eventually stop it from expanding (which it is currently doing) and cause it to contract. Eventually it will end in a big crunch.

One thing that allows us to see whether a given space is positively curved is to draw triangles. If we'd like to travel between two points on a sphere along the shortest route, then we should travel along a *great circle*: A circle that passes through two opposite points on the sphere. Great circles are the analogues of straight lines on the flat plane. If we draw a triangle whose sides are made from pieces of great circles, we will find that, unlike in flat 2D space, its angles add up to more than 180 degrees.



The surface of a saddle. The angles of a triangle drawn on this saddle add up to less than 180 degrees. If the Universe has negative curvature, then its geometry is the 3D analogue of an infinite saddle.

Another possibility is that the density of matter is so low that the curvature of the Universe is negative. In this case, the Universe is the 3D analogue of the surface of a saddle. The angles of a triangle add up to less than 180 degrees and the Universe is infinite. It's hard to visualise this 3D *hyperbolic space*, but again it's easy to describe it mathematically. As for the ultimate fate of the Universe, there's not enough matter to significantly slow the rate of expansion: ultimately, the Universe will die in a cold and lonely *big freeze*.



A piece of the flat plane. The angles of a triangle drawn on this plane add up to exactly 180 degrees. If the Universe has zero curvature, then its geometry is the ordinary 3D space we learn about at school.

Finally, it could be that there's just enough matter for the Universe to have zero curvature. Angles of triangles add up to exactly 180 degrees and the Universe is infinite. If this is the case then it may be that matter can slow the rate of expansion, but it'll never be able to bring it down to zero. The Universe will also expire in a big freeze.

Observations of the cosmic microwave background (the left-over radiation from the Big Bang) suggest that the Universe is indeed flat, or at least very nearly so. This would imply eternal expansion, at least if the assumptions above are true.

The first is that general relativity is true. The theory has been tested many times and has always stood up to the tests admirably, so physicists on the whole take it to be true. Another assumption is that the Universe is isotropic, that is, as we look in the various directions in the sky, no direction stands out over an other, they are all the same, on average. This idea may seem strange at first, since the world definitely looks different in different directions. However, this assumption refers to the average look over large volumes of space. Based on current observations, this average really is the same everywhere, so the Universe is indeed likely to be isotropic.

Mass, Size and Density of the Universe

The mass, size, and density of the universe involve very big and very small numbers with large numbers of zeros in front of or behind the decimal point. Scientists write such numbers using a special code that keeps them short. When they write a number like 6e7 then the number after the "e" indicates by how many places the decimal point must be shifted to the right. For instance, 6e7 is equal to a 6 with 7 zeros behind it, or 60,000,000, and 6.1e7 is equal to 61,000,000. If the number after the "e" is negative, then the decimal point is shifted to the left. For instance, 6e-3 is equal to a 6 with 3 zeros in front of it (and the decimal point after the first zero), or 0.006, and 6.1e-3 is equal to 0.0061.

The mass density of visible matter (i.e., galaxies) in the Universe is estimated at 3e-28 kg/m^3 (3e-31 times the mass density of water). The radius of the visible Universe is estimated at 1.7e26 m (18 thousand million lightyears) plus or minus 20 percent or so. This yields a total mass of the visible matter of about 6e51 kg (1.3e52 lb), which is equivalent to the weight of 4e78 hydrogen atoms. Since nine out of ten atoms and ions in the Universe are in the form of hydrogen, this is a reasonable estimate for the number of atoms in the Universe (based on the visible galaxies only). Maybe a correction factor of the order of 2 has to be applied to account for the warping of space on very large scales.

However, there is considerable uncertainty about the mass density of all matter (visible and invisible) and energy (through Einstein's E = mc^2 equation). When one studies the movement of matter in and around galaxies, then it appears that up to about 10 times more mass is pulling at the matter (through its gravity) than is accounted for in the visible stars. This is the "missing-mass" problem. If this factor of ten holds throughout the Universe, then the total mass in the Universe would be about 6e52 kg. If the missing mass were mostly in the form of hydrogen atoms (which is not at all clear) then the number of atoms would be about 4e79.

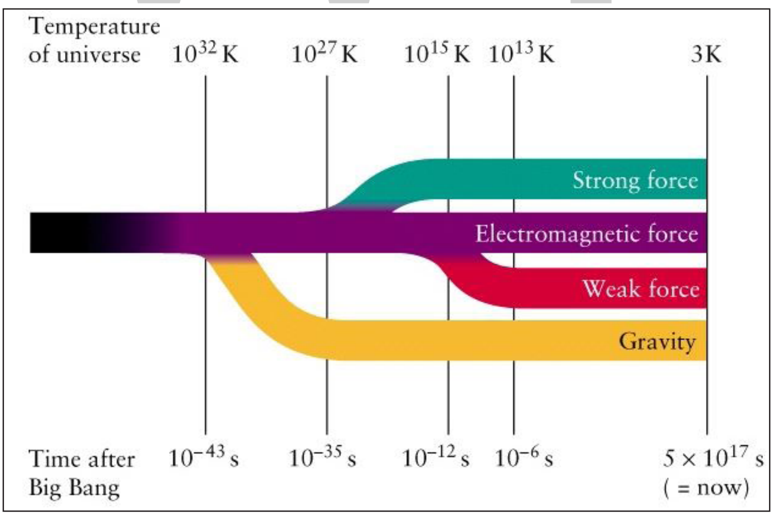
A currently popular theory of the formation of the Universe (the so-called Inflation Theory) predicts that the mass density of the Universe should be close to the so-called critical density that separates an open universe that always grows from a closed universe that ultimately collapses again. This critical mass density is currently equal to 6e-27 kg/m^3. If the Universe is at the critical density, then the total mass of the Universe is closer to 1e53 kg, and the number of atoms (assuming that most of the mass is in the form of hydrogen atoms) about 6e79.

It seems, then, that the number of atoms in the Universe is at least about 4e78, but perhaps as many as 6e79. That is, 10^{79} atoms.

The Early Universe

The Planck Epoch

The Plank Epoch encompasses the time period from 0 to 10^-43 seconds. This extremely small unit of time is aptly referred to as a "Plank Time". Not much is truly known about this period of time, however some very interesting hypothesis have been made.



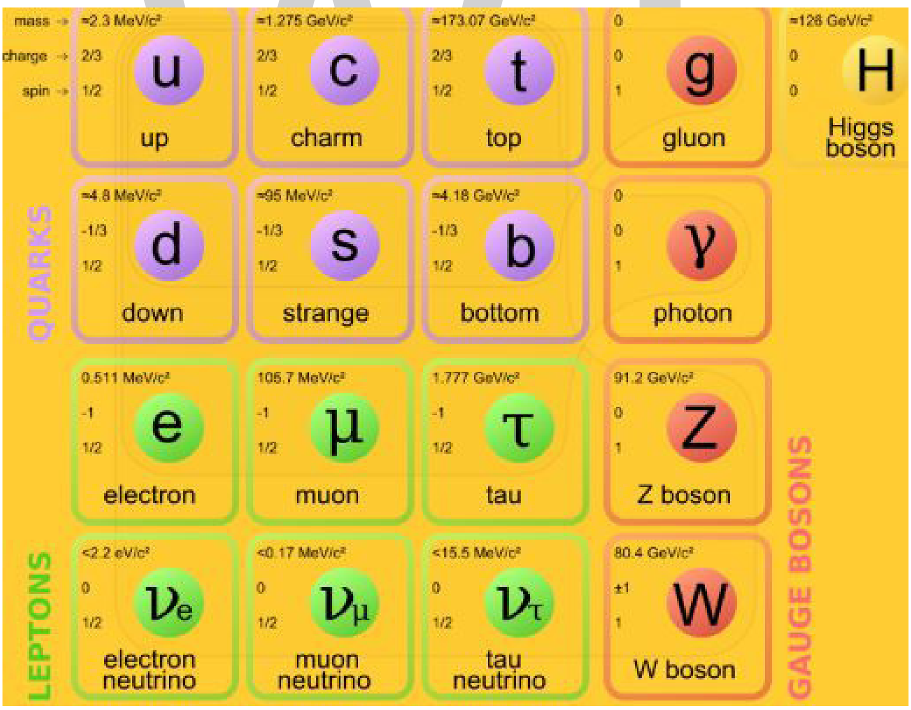
Einstein's theory of relativity proposes that a gravitational singularity may have existed. In a gravitational singularity, even the laws of quantum physics break down and the four fundamental forces (strong nuclear, weak nuclear, electromagnetic and gravity) could be unified as one. This is an extremely odd concept to consider. It also ties into the so-called 'Theory of Everything' which states that at high enough energy levels, even gravity will combine back into one unified force with the other three.

During the Plank Era, our universe was only 10^-35 meters wide (very small) and 10^32 degrees celsius (very hot).

Grand Unification Epoch

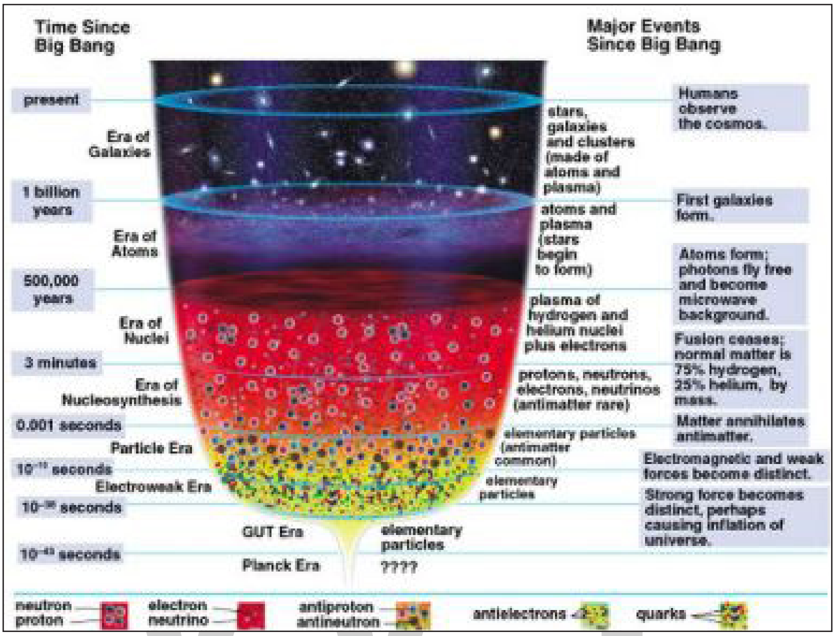
The Grand Unification Epoch took place from 10^-43 seconds to 10^-36 seconds after our universe was born. Quantum theory allows us to form a clearer picture of this Epoch compared to the mysterious Plank Epoch.

During the Grand Unification Epoch, the universe was still extremely hot and incomprehensibly small. However, it had cooled down enough to allow the force of gravity to separate from the other three fundamental forces. The unification of the strong nuclear, weak nuclear, and electromagnetic force that existed during this period of time is referred to as the electronuclear force. However, the splitting off of gravity from the electronuclear force wasn't the only milestone of this epoch- this is also when the first elementary particles began to form.



Elementary Particles: Elementary Particles are particles which have no substructure- i.e. they are the simplest form of matter possible. Elementary particles are the building blocks of electrons, neutrons, protons and more! Currently, there are 17 elementary particles that have been confirmed- the unconfirmed "gravitron"is still in the theoretical category. There are 12 "matter" elementary particles and 5 "force carrier" particles.

Fermions: These are the matter elementary particles are what make up the physical part of subatomic particles and are referred to as fermions. The two categories of elementary fermions are quarks and leptons. Quarks combine to form particles known as Hadrons, which make up the famous neutrons and protons; Leptons form electrons and other fundamental particles.



Bosons: The 5 force carrier particles mediate the interactions between the weak magnetic, strong magnetic, and electromagnetic forces. Bosons are the fundamental reason for the attractions/reactions we view as forces.

Hardon Epoch

Hadron Epoch 10^{-6} to 1 Second

Space was still expanding rapidly after the creation. The Universe was about the size of our Solar System today. The quark-gluon plasma was getting larger and its quarks were being pulled further apart. Now the strong force carried by gluons revealed an exotic and important property.

As quarks moved further apart, the gluon elastic force began to increase, and increase very rapidly. When the distance between any two quarks was more than 10^{-15} meters, the gluon elastic broke and produced two new quarks, one at each end of the break.

Not only was the space between quarks increasing as space expanded; they were also moving more slowly. The temperature dropped to around 10 thousand million degrees Celsius. Soon they did not have enough energy to break the gluon elastic, and each quark found itself permanently tied to one or more of its neighbours.

The Hadron Epoch started about 10^{-6} seconds after the creation and ended about 1 second later.

Hadrons

Quarks combined with each other to produce other particles which we find in the world around us today. Up and down quarks grouped together in threes to produce protons and neutrons. A quark combined with an anti-quark to produce a meson. To summarise:

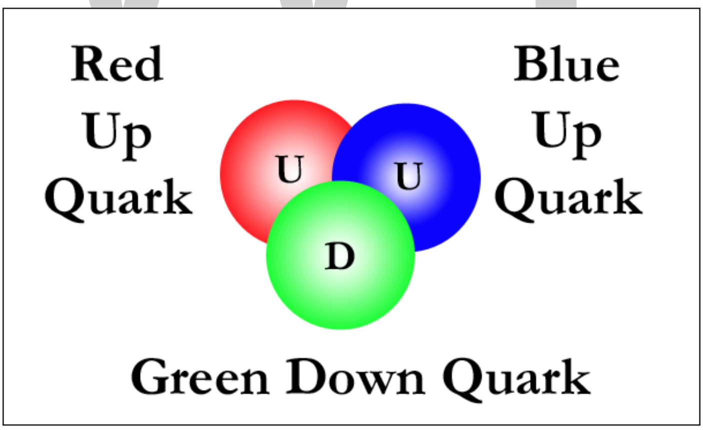
- Three quarks make a proton or a neutron. Particles of this type are called baryons.
- A quark and antiquark make a meson.

Both of these types, baryons and mesons, are grouped together under the name of hadrons. All hadrons are therefore made of quarks.

Proton

Of all the particles we have met in the young Universe so far, the proton was one of the first which would last down to the present day.

Protons are made of three quarks. There must be two up quarks and one down quark to make a proton.



A proton consists of three quarks.

Each quark has a property called color charge and now we will see why this property is given this name. Although the property is nothing like the colors we see with our eyes, but the name was given in analogy with color mixing.

To make white light you mix red, green and blue light. In a similar way, to make a proton you need three quarks each with a different sort of color charge, as shown in the diagram. In fact the naming of the colors is arbitrary.

Protons carry something called a positive electric charge. This will be very important later in our story because, as we will see, it will enable a proton to capture another particle (called an electron) and so make a very interesting object called an atom.

The antiproton is an antimatter particle which has the same mass as a proton but carries a negative electric charge.

Protons on Soccearth

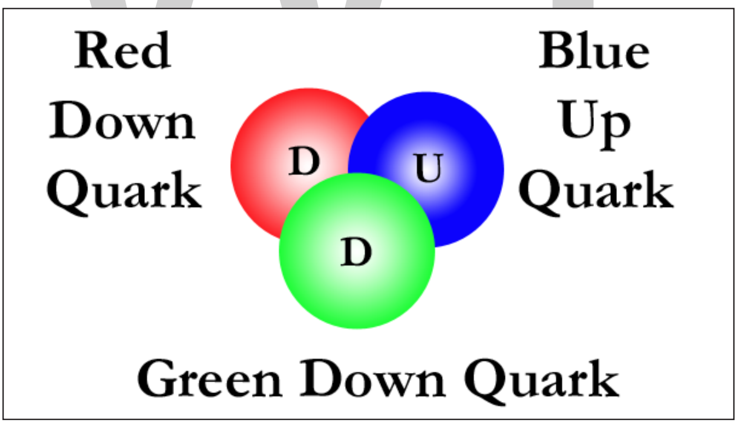
A proton is extremely small. One proton is only about $2*10^{-15}$ meters across (2 femtometers) and it weighs $1.673 *10^{-27}$ kilograms.

What would a proton look like on Soccearth, a soccer ball blown up to the size of the Earth? It would still be much too small to see, even under a microscope! You could only tell that it is there because it would attract an electron to make an atom.

Neutron

Another particle created during the Hadron Epoch was the neutron. The neutron was discovered in 1932 by the English physicist James Chadwick.

Like the proton, the neutron consisted of three quarks, one of each color charge. But, unlike the proton, the neutron contained two down quarks and one up.



Quarks inside a neutron.

The neutron is a little heavier than a proton. Its mass is nearly 1,840 times that of the electron. But, unlike the proton, the neutron has no electric charge. The neutron's name means neutral or without charge.

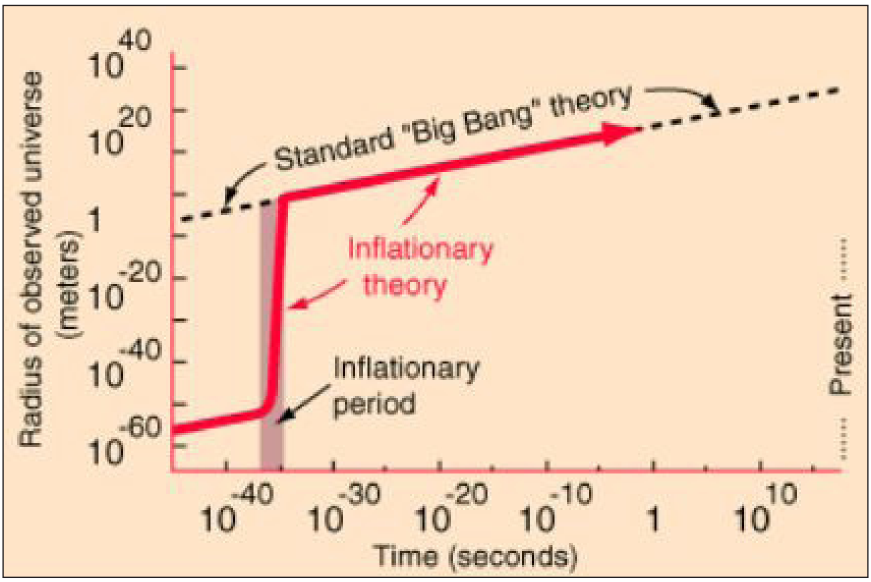
Free neutrons easily pass through atoms, because they have no electrical charge, and so they form highly penetrating radiation, interacting with matter only through collisions with atomic nuclei.

The neutron is important in forming a chain reaction in nuclear fission, as used in nuclear reactors and atom bombs. The absorption of neutrons by other nuclei, exposed to the high neutron densities in nuclear reactors, generates radioactive isotopes useful for a wide variety of purposes.

Inflationary Epoch

Cosmic Inflation

After Grand Unification things really started to get interesting. Enter the universe's Inflationary Epoch, which took place between 10^-36 to 10^-32 seconds after the big bang.



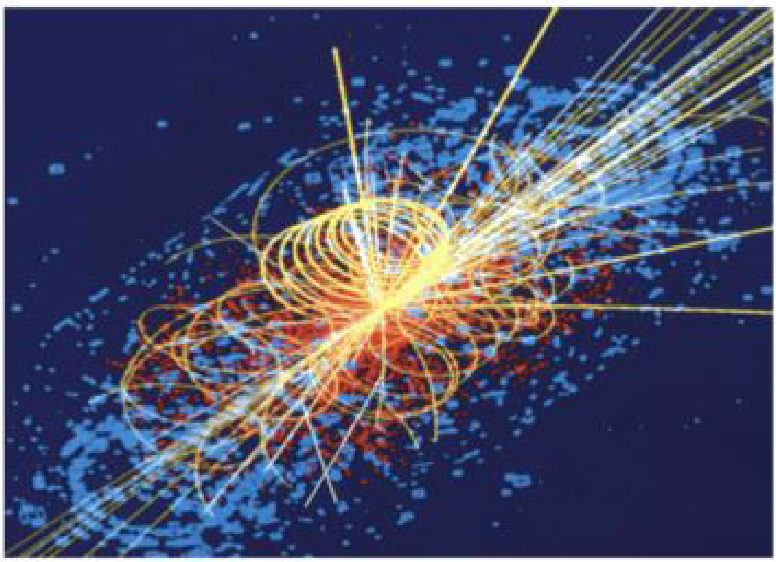
During the inflationary epic the strong nuclear force broke off from the weak nuclear and electromagnetic forces. This is thought to have triggered a rapid inflation of the universe into the empty space that surrounded it due to the enormous amount of energy released during this process. The universe began to expand exponentially during this time period, and with that the temperature began to drop quickly as well. To get an idea of just how small the universe was up to this period, ponder this: It is estimated the increase in size during the inflationary period was by a factor of around 10^26. However, this increase in size amounted to a final width of just 10 centimeters. It is extremely hard to comprehend a universe that was this small containing all the energy in the known universe, and almost gives notion that the concept of "size" is extremely subjective and almost meaningless.

Electroweak Epoch

The Electroweak Epoch lasted from 10^-36 to 10^-12 seconds. Note the slight overlap with the Inflationary Epoch. The Electroweak Epoch marks the beginning of the modern universe as we know it, due to the fact that all four fundamental forces will be distinct from one another by its end. We will also see the formation of the first particles that would give way to matter.

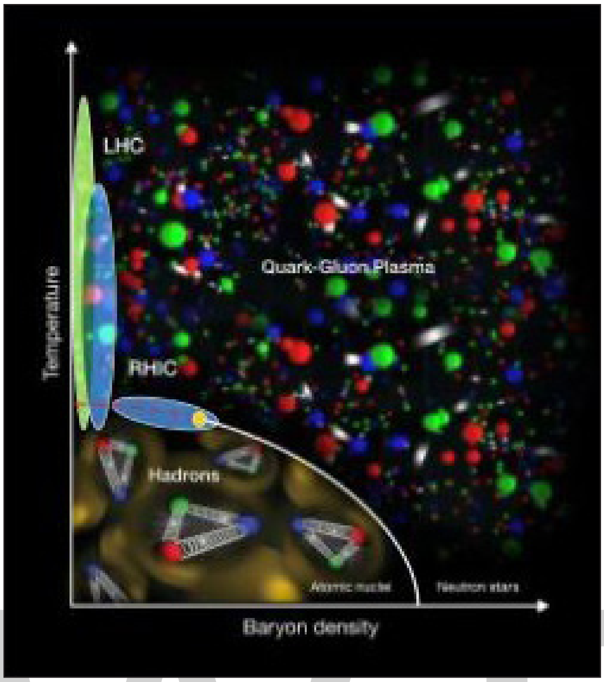
The separation of the strong nuclear force led to particle reactions that formed W, Z and Higgs boson particles. Remember, bosons are not the physical elementary particles, they are force carriers. This is a massive step towards physical matter becoming present in the universe, as the Higgs field can slow down other particles and impart mass on them, another very odd concept to fathom. This field is what allows our universe to support mass.

Towards this end of this era, right around 10^-12 seconds, the separation of the weak nuclear force and the electromagnetic force finally occurs. Now that the universe has cooled down enough to allow all four forces to act independently, and the laws of nature that we account for in classical physics begin take shape.



Quark Epoch

The Quark Epoch, lasting from 10^-12 to 10^-6 seconds, followed the Electroweak Epoch. Now that the universe has cooled down to about 10 quadrillion degrees and the four fundamental forces are separate, the universe had a higher degree of stability. It is here we begin to see the rapid formation of quarks and anti-quarks. These particles collide and annihilate one another on impact. In one of the more bizarre twist's of nature, a process known as baryogenesis caused a surplus of quarks to began to accumulate. Baryogenisis theorizes that for about one in every billion quark-antiquark collisions the quark was not annihilated. This assymetry of quarks to antiquarks is the reason why mass exists today, otherwise particles may have just collided and destroyed each other until the end of time.



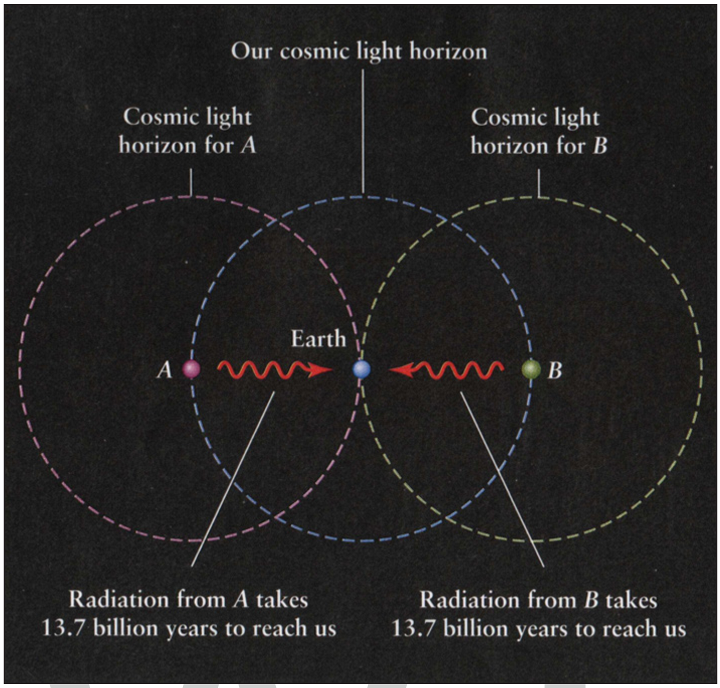
The build-up of quarks and the presence of gluons (a boson force-carrier particle) caused the consistency of the universe to be a densely populated quark-gluon "soup". It is named as such because observation of particle collisions have given physicist's reason to believe that the universe was a plasma at this point in time. Up until these observations, the prevailing theory was that the universe was a gas during the Quark Epoch.

The combination of quarks and gluons make-up protons and neutrons, but during the quark epoch they had too much energy to be confined inside them. However, the age of subatomic particles was on the horizon.

Inflation

According to the theory of inflation, the early Universe expanded exponentially fast for a fraction of a second after the Big Bang. Cosmologists introduced this idea in 1981 to solve several important problems in cosmology.

One of these problems is the horizon problem. Assume, for a moment, the Universe is not expanding. Now imagine a photon was released very early in the Universe and travelled freely until it hits the North Pole of the Earth. Now imagine another photon was released at the same time, but "opposite" to the first one. It will hit the Earth at the South Pole. Could these two photons exchange any information from the time when they are released? Clearly not, because the time required to send information from one photon to the other would be two times the age of the Universe. The photons are causally disconnected. They are outside of each other's horizon.

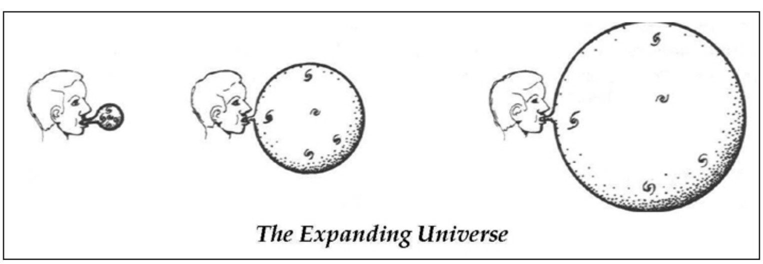


These photons could not have communicated with each other unless inflation took place during the very early Universe.

However, we observe that photons from opposite directions must have communicated somehow, because the cosmic microwave background radiation has almost exactly the same temperature in all directions over the sky.

This problem can be solved by the idea that the Universe *expanded* exponentially for a short time period after the Big Bang. Before this period of inflation, the entire Universe could have been in causal contact and equilibrate to a common temperature. Widely separated regions today were actually very close together in the early Universe, explaining why photons from these regions have (almost exactly) the same temperature.

A simple model for the expansion of the Universe is to consider the inflation of the balloon. A person at any point on the balloon might consider themselves to be at the centre of the expansion, as all neighbouring points are getting further away.



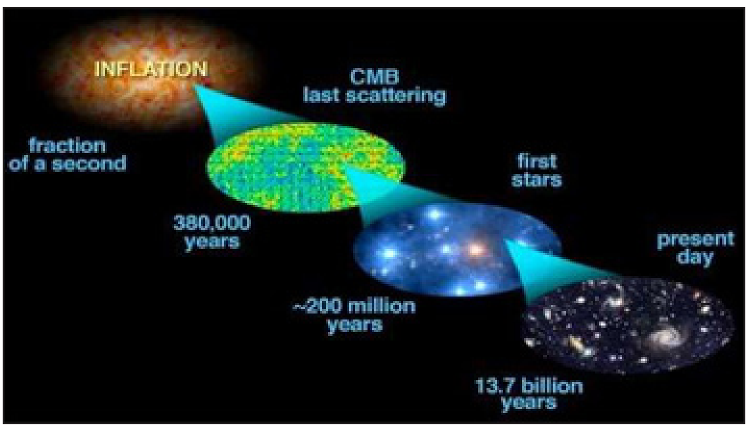
As the balloon inflates, the distances between objects on the surface of the balloon increases.

During inflation, the Universe expanded by a factor of about $e^{60}=10^{26}$. This number is a one followed by 26 zeros. It transcends normal political/economic discussions of inflation.

Quantum Fluctuations

Let's suppose that before inflating the balloon, we write a message on the surface of the balloon which is so tiny that you cannot read it. Inflating the balloon will make the message readable for you. This means that inflation acts as a microscope, which magnifies what was written on the initial balloon.

In a similar manner we are able to observe quantum fluctuations that were created at the beginning of inflation. The expansion of the Universe during the inflationary epoch serves as a huge microscope that magnifies quantum fluctuations, corresponding to a scale less than 10-28 cm, to cosmological distances. This leaves imprints in the cosmic microwave background radiation (hotter and colder regions) and in the distribution of galaxies.



Inflation works as a cosmic microscope to see the quantum fluctuations in the very early Universe.

Using classical physics, the evolution of the inflationary Universe is homogeneous - each spatial point evolves exactly the same way. However, quantum physics introduces some uncertainty in the initial conditions for the different spatial points.

These variations act as seeds for structure formation. After the inflationary period, when fluctuations are amplified, the density of matter will vary slightly from place to place in the Universe. Gravity will cause the more dense regions to start contracting, leading to the formation of galaxies.

Probing the Early Universe

The figure below shows how the image of quantum noise may appear imprinted on the cosmic microwave background. Red and blue denote hot and cold variations of the temperature, measured by the WMAP satellite over seven years. Comparing the statistics of the measured data with our theoretical calculations shows very good agreement.

Low Density Inflationary Universes

Universe of Critical Density

For quite some time it has been known that the mean density of our universe agrees with the critical density to within better than a factor of ten. Even with such large margin of error this agreement is remarkable. Establishing initial conditions so that the mean density remains close to the critical density for more than a fleeting moment is much like trying to balance a pencil on its point. A universe initially with slightly subcritical density rapidly becomes increasingly subcritical and soon virtually indistinguishable from an empty universe. Similarly, an ever so slightly supercritical universe rapidly collapses into a Big Crunch, never reaching the old age of our universe-somewhere around twelve billion years. To obtain a universe like ours seems to require fine tuning of the initial density to agree with the critical density to an accuracy around one part in 10^{60}.

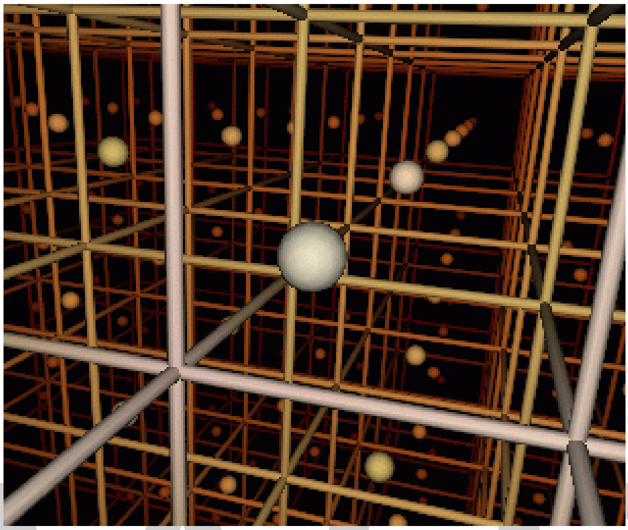
For a long time it was regarded simplest and aesthetically most pleasing to postulate that our universe is now of exactly critical density. The versions of inflation developed in the early 1980s provided a mechanism for setting the density of the universe near the critical density with nearly unlimited precision. For many years an exactly critical universe was touted as one of the few firm predictions of inflation.

Geometry and the Density of the Universe

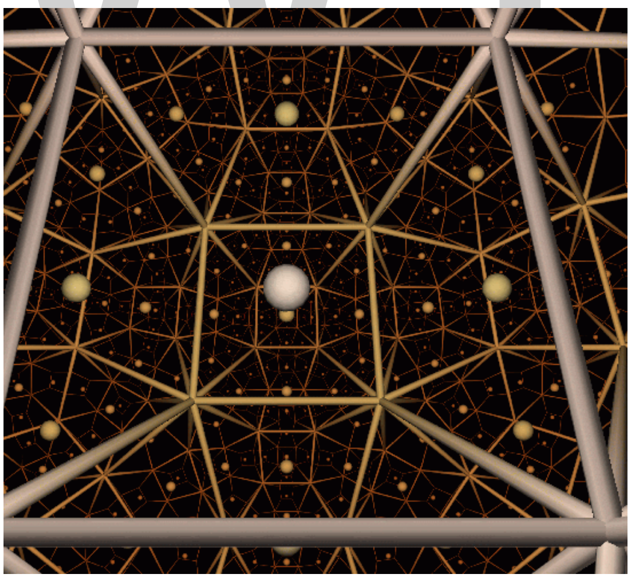
In Einstein's General Theory of Relativity, formulated in 1915, gravity is understood in terms of geometry rather than as just another ordinary force. Matter tells spacetime how to curve and the resulting spacetime curvature tells bodies how to move. For the special case of an expanding universe, idealized as filled with a uniform density of matter, a good approximation on large scales, General Relativity establishes an intimate connection between the density of the universe in comparison with the critical density and its geometry. A universe of critical density (at constant cosmic time) has the familiar Euclidean geometry so well known to us from every experience and from classical perspective as taught in art class. However, a universe of subcritical or supercritical density has a non-Euclidean geometry-hyperbolic if the density is subcritical, or spherical if the density is supercritical.

On small scales these different geometries are much alike. An ant on the surface of an apple might view its immediate surrounding as quite flat and might experience difficulty in figuring out that the apple is round. Likewise, if the curvature of the universe would become apparent only on scales beyond several billion light years we might be deceived into believing that its geometry is Euclidean. Only on large scales larger than the so-called curvature scale do the differences between the geometries become large effects.

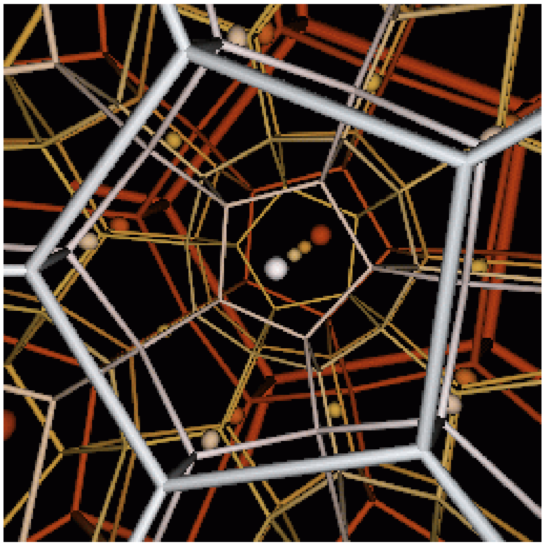
The following three plates illustrate the difference in perspective between the three possible geometries: A hyperbolic geometry, a Euclidean geometry, and a spherical geometry. In all three cases, space is divided into identical cells, whose edges are indicated by the rods. The balls within the cells are of identical size, and increasing distance is indicated by reddening.



In the Euclidean geometry space is divided into cubes and one experiences the ordinary, familiar perspective: The apparent angular size of objects is proportional to the inverse of their distance.



Hyperbolic space shown here is tiled with regular dodecahedra. In Euclidean space such a regular tiling is impossible. The size of the cells is of the same order as the curvature scale. Although perspective for nearby objects in hyperbolic space is very nearly identical to Euclidean space, the apparent angular size of distant objects falls off much more rapidly, in fact exponentially, as can be seen in the figure.



The spherical space space shown here is tiled with regular dodecahedra. The geometry of spherical space resembles the surface of the earth except here a three-dimensional rather than two-dimensional sphere is being considered. Perspective in spherical space is peculiar. Increasingly distant objects first become smaller (as in Euclidean space), reach a minimum size, and finally become larger with increasing distance. This behavior is due to the focusing nature of the spherical geometry.

Geometry of our Universe

During the 1980s observations remained sufficiently crude so that a universe of critical density was quite plausible. But more recent observations have made it increasingly difficult to reconcile a critical universe with the observations.

It is known that in addition to the luminous matter seen in the form of stars the universe contains a large amount of "dark" matter, in particular in the halos around galaxies. The presence of this dark matter is inferred from its gravitational pull on the surrounding matter. Since the dark matter is distributed in a less clustered manner than the luminous matter, the apparent average density seems to increase as larger and larger scales are probed. For a long time it was hoped probing sufficiently large scales would uncover a critical density of dark matter.

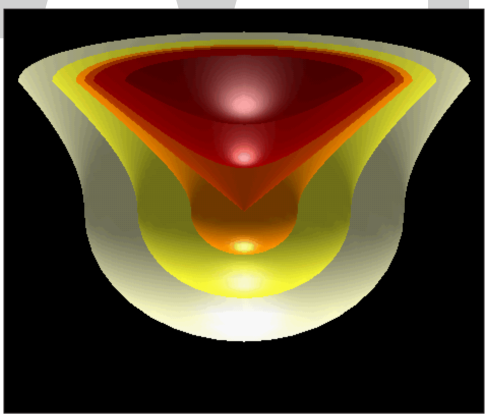
Today it seems unlikely that this hope will ever be realized. It is now possible to probe the average density of the universe on scales large enough to compromise a fair sample of the universe. We present the so-called "cluster baryon fraction" as one illustrative example of the strong evidence in favor of a universe of subcritical density. Rich clusters of galaxies are the largest gravitationally bound systems in the universe.

Using nuclear physics one can determine the baryon density of the universe. With the density of baryonic matter known, the total density can be determined from measuring the baryon fraction. The baryonic mass of a cluster can be determined by adding the masses of the constituent galaxies inferred from their light to the mass of the hot intracluster gas, which can be determined from X-ray observations of emission from the gas. The total mass can be determined by a variety of methods. The motions of the constituent galaxies allow one to determine the depth of the potential well and hence the total mass of the cluster. X-ray observations allow the same to be done with the gas, and gravitational lensing of background objects by the gravitational field of the cluster, resulting in the distortion in appearance of background galaxies, provides a completely independent check of the total mass.

These techniques, and a number of independent techniques as well, suggest a universe with approximately one third of the critical density. Although a universe of critical density cannot yet be ruled out definitively, the possibility of a critical universe now appears like quite a long shot.

Reconciling a Low Density Universe with Inflation

Inflation smooths the universe by postulating an early epoch of extremely rapid expansion during which whatever irregularities may have existed prior to inflation are virtually erased. In ordinary inflation, as developed by Guth, Linde, Albrecht, and Steinhardt, this smoothing flattens the universe as well, yielding a universe of critical density. In ordinary inflation, a critical universe could in principle be avoided by shortening the amount of inflation, but in that case the smoothness on large scales remains a mystery, causing inflation to lose most of its appeal.



The Creation of a Single Bubble Open Universe.

The vertical direction indicates time and the horizontal directions are spatial. The value of the inflaton field is constant on the various slices and the colors indicate the cooling down of the universe as one passes into the bubble interior. The bubble is expanding into the surrounding inflating spacetime stuck in the false vacuum. We live inside the bubble interior.

In single bubble open inflation there are two epochs of inflation. In inflation the rate of expansion is controlled by a scalar field, known as the inflaton field. The inflaton field wants to roll down the hill to the bottom and as the field descends the rate of expansion of the universe decreases, eventually ending the epoch of inflationary expansion. In open inflation the inflaton field at first remains stuck in a local minimum of the potential. While the field is stuck there, a first epoch of inflationary expansion takes place during which the universe is smoothed. In fact during this epoch the symmetry of the spacetime is so large that no particular time direction is preferred over any other.

According to classical physics, once stuck in the local minimum the inflaton field never escapes. However, quantum mechanics allows the field to tunnel through the barrier. This tunneling occurs through the nucleation of a bubble that subsequently expands, somewhat as an expanding bubble in a pot of boiling water.

Subsequently, the bubble expands at the speed of light. It cannot have any velocity other than the speed of light, for else a preferred time direction would be required to exist. The surfaces on the bubble interior on which the scalar field is constant have a hyperbolic spatial geometry, and these are the surfaces that we inside the bubble later perceive as surfaces of constant cosmic time. As one passes inside the bubble, the interior continues to inflate, creating a universe with a large curvature radius. Further inside the bubble the energy of the inflaton field is converted into ordinary matter and radiation, and the hyperbolic universe continues to expand and cool down.

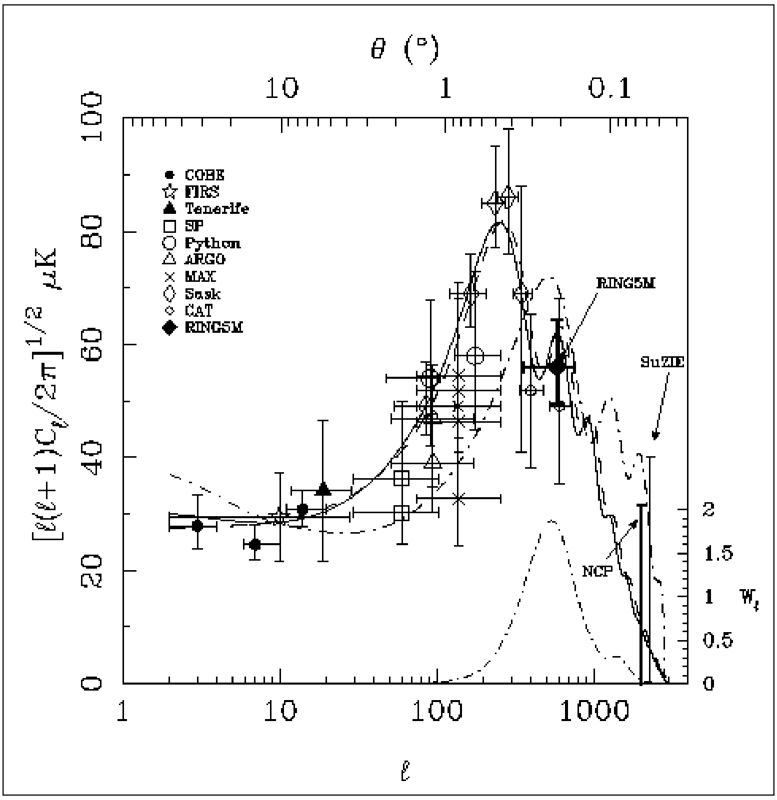
Testing Open Inflation

Plotted is the level of anisotropy as a function of angle and various measurements thereof. The curves indicate theoretical predictions for various models. The solid curve indicates a universe of critical density whereas the dot-dash-dot-dash curve indicates a low density universe. Note how the position of the first peak shifts to the right to smaller angular scales in the low density universe.

The best hope for testing open inflation derives from measuring the geometry of the universe, which can be determined through observing the ripples in the cosmic microwave background radiation.

The 3K cosmic microwave background radiation emanates from an epoch approximately three hundred thousand years after the Big Bang, when the universe was approximately one thousandth its present size. At this time the electrons, because of the cooling of the universe, combined with protons and other nuclei to form neutral hydrogen and other elements. Because of this change in composition from a highly ionized plasma to a neutral gas, the formerly opaque universe becomes virtually transparent. The non-uniformities in the microwave background provide a snapshot of the ripples at that time, which later developed into galaxies and the structure that we observe today.

Inflation in general, and open inflation on scales much shorter than the curvature scale, imprints essentially scale free fluctuations on the matter filling the universe. At recombination, however, the physics at that time, believed to be well understood, introduces a preferred scale of known length on which the first acoustic oscillations of the plasma occur. This scale is of known physical size, and from its angle subtended in the sky today, we can determine the geometry of the universe.



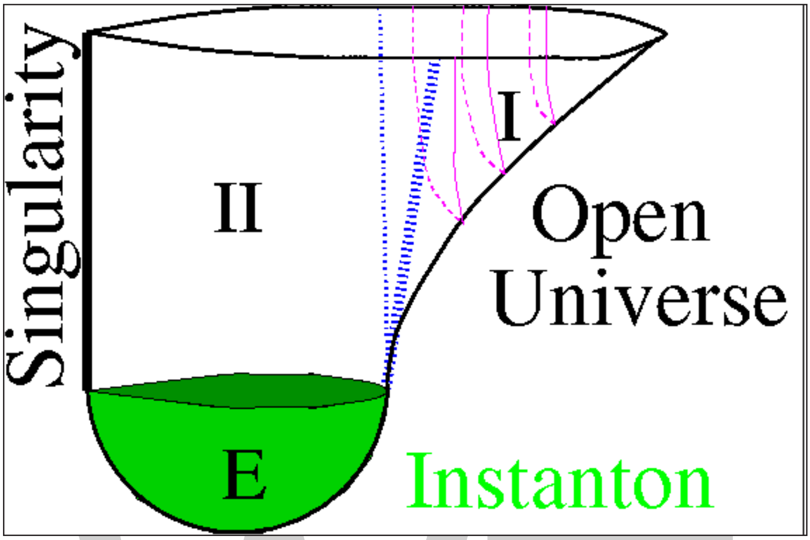
Microwave Anisotropy as a Function of Angle.

More General Open Inflation

The above models for open inflation provide a counter-example to the standard lore on inflation, but they rely upon the presence of a local minimum in the potential energy of the inflaton field. At our present level of understanding, we simply cannot tell whether this is what is predicted by a more fundamental theory such as M-theory or supergravity. But in the model theories for which we can calculate the inflaton potential energy, such local minima do not usually appear.

A bubble universe emanates from a Hawking-Turok instanton. The vertical direction indicates time and the horizontal directions are spatial. E indicates the Euclidean region, where time becomes spacelike, and I is the bubble interior. The heavy line to the left indicates the mild singularity occurring in these solutions.

Hawking and Turok realised that open inflation was in fact much more general, and could even occur in a theory where there is no local minimum in the inflaton potential energy. In fact, they showed that for essentially any potential energy function allowing inflation, an open universe similar to that obtained in the expanding bubble described above could be formed.



Hawking-Turok Instanton.

Hawking and Turok's calculation was performed in the framework of a proposal for the initial conditions made in 1983 by Hawking and James Hartle. They proposed that the initial condition for the universe should be that it possessed no initial boundary. One can picture the spacetime of an expanding universe as the surface of a cone, placed vertically with its sharp tip down. Time runs up the cone: Space runs around it. Time and space end at the sharp tip. The tip is 'singular' in mathematical terms and if this were a model of the universe we would find all our equations break down there. Instead, Hartle and Hawking proposed that the tip be rounded off.

This rounding off is only possible if the nature of spacetime changes in the vicinity of the tip. In effect, all directions must become 'horizontal' near the tip, which is to say that all directions are spacelike. This is just what we need to explain how time began. In effect the distinction between space and time is blurred and space is then rounded off.

The region where time becomes spacelike is technically termed the instanton region. Instantons are solutions to the equations of general relativity and matter (here, the inflaton field) which have four spacelike directions. Hawking and Turok showed that for essentially any theory which allows inflation, there is a family of instanton solutions each one of which describes the formation of an inflating open universe. The Hawking-Turok instantons do actually possess a singularity, but only at a single point. Unlike the singularity in the standard hot big bang, which is so severe that we cannot predict anything that happened in its presence, the singularity in the Hawking-Turok instantons is so mild that, as for the singularity in the electric field at the centre of a hydrogen atom, it does not affect our ability to make predictions.

Friedmann Universe

Friedmann Universe model was developed developed in 1922 by the Russian meteorologist and mathematician Aleksandr Friedmann. He believed that Albert Einstein's general theory of relativity required a theory of the universe in motion, as opposed to the static universe that scientists until then had proposed. He hypothesized a big bang followed by expansion, then contraction and an eventual big crunch. This model supposes a closed universe, but he also proposed similar solutions involving an open universe (which expands infinitely) or a flat universe (in which expansion continues infinitely but gradually approaches a rate of zero).

The Friedman Equation

If General Relativity (GR) is correct, the evolution of R with time, and hence the evolution of the universe, is given by the Friedman equation:

$$\left(\frac{1}{R}\frac{dR}{dt}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{R^2}$$

Here rho is the density of the universe (including the mass equivalent of any energy, according to E = m c ²), G is the Gravitational constant, c is the speed of light as always, and k is the curvature constant that appears in the Robertson-Walker metric.

The term on the left of the Friedman equation is the square of the Hubble parameter:

$$H \equiv \frac{1}{R}\frac{dR}{dt} = \frac{\dot{R}}{R}$$

The Friedman equation can be re-written in a very neat form if we define a critical density:

$$\rho_c = \frac{3H^2}{8\pi G}$$

and hence the density parameter:

$$\Omega = \rho / \rho_c$$

The Friedman equation now looks like:

$$kc^2 = H^2 R^2 (\Omega(R) - 1)$$

which shows the importance of the critical density: if the density is greater than this (Omega > 1), the right-hand side is positive, so we must have k = +1 and a closed universe. If Omega =1, we must have k = 0, and for smaller densities the universe is open (in the simplest topology). If k = 0, this last form reduces to the unhelpful 0 = 0. This is because in this case the scale factor R is not determined absolutely; instead one can choose it arbitrarily at some reference moment and the Friedman equation then determines how it changes with time.

The key to solving the equation is to find how the density (or equivalently Omega) depend on R. Imagine a box with co-moving observers at its corners: the amount of matter in the box stays the same as the universe expands, because the universe is homogeneous; all regions behave in the same way so, for instance, no region can have a net gain of matter with respect to its surrounding regions; any random motion of atoms (or galaxies) should cancel out on average. The volume of the box scales as R^3, so we expect the matter density to behave as:

$$\rho_m \propto R^{-3}$$

The energy density of radiation falls more quickly than this, because although the number of photons is preserved, each photon expands with the universe and hence loses energy because its frequency is falling. Running the universe back in time, the radiation energy density increases faster than the matter density and eventually overtakes it. However, this happens so close to the beginning of time (when R was about a thousandth of its present value) that we can ignore this short period and leave radiation out of the calculations.

But there is yet another component that we should worry about, the cosmological constant. This was put into GR by Einstein (although he later regretted it). It corresponds to an effective density that is independent of R, and might even be negative! As the matter density in the universe falls, the cosmological constant becomes more and more important, and eventually dominates. A positive cosmological constant corresponds to an effective negative pressure, equal and opposite to its energy density. Work done by expansion agains this "tension" creates the extra energy that allows the energy density to be constant when the volume of each bit of the universe is increasing. In GR the true source of gravity is not just mass (or density, for spread-out stuff) but:

$$\rho + 3P / c^2 = \rho_m + \rho_\Lambda + 3(-\rho_\Lambda c^2) / c^2 = \rho_m - 2\rho_\Lambda$$

Here the symbol Lambda, as is conventional, denotes the density corresponding to the cosmological constant: the important thing is that negative pressure counterbalances gravitational attraction - just the opposite of what you might think.

The Friedman equation is just a first-order differential equation and can be integrated (numerically, except for the simplest cases) to give the history of the universe in the form t(R). This function is fixed by just two parameters which can take a continuous range of values, and the curvature constant k= -1, 0, or +1. For the continuous parameters we could choose.

- The constant of proportionality in the equation for the matter density.
- The cosmological constant.

However, these are dimensional, which means that one of them simply defines the overall scale of the universe. It is generally best to work with dimensionless parameters.

Conventionally we use the density parameters at the present time:

$$\Omega_m = \rho_m / \rho_c, \quad \Omega_\Lambda = \rho_\Lambda / \rho_c, \quad \Omega = \Omega_m + \Omega_\Lambda$$

This choice has the nice feature that we don't have to worry about k explicitly, as it given by the sign of Omega-1. On the other hand, there is a drawback: there is nothing special about the present time, so we are inserting an extra scale (the Hubble time $t_H = 1 / H_o$, i.e. the inverse of the present-day Hubble parameter) into the equations, which complicates matters a bit. Notice that because H changes with time, so does the critical density, and so the density parameters (Omegas) change with time in a moderately complicated way.

Solutions to the Friedmann Equation

Matter

For a Universe with only matter, we have:

$$\frac{\dot{a}^2}{a^2} = H_0^2\,\Omega_m\,a^{-3}$$

$$\left(\frac{da}{dt}\right)^2 = H_0^2\,\Omega_m\,a^{-1}$$

$$a^{1/2}\,da = H_o\sqrt{\Omega_m}\,dt$$

Integrating and setting imposing a = 0 at t = 0, we obtain,

$$\frac{2}{3}a^{3/2} = H_o\sqrt{\Omega_m}\,t$$

$$a(t) = \left(\frac{3}{2}\sqrt{\Omega_m}H_0 t\right)^{3/2} \qquad (\text{Matter Domination})$$

For an Einstein-de Sitter (EdS) Universe, $\Omega_m = 1$ and the age of the Universe (t =today, with a = 1) is:

$$t = \frac{3}{2}H_0^{-1}$$

Radiation

For a Universe with only radiation, we have:

$$\frac{\dot{a}^2}{a^2} = H_0^2 \, \Omega_r \, a^{-4}$$

$$\left(\frac{da}{dt}\right)^2 = H_0^2 \, \Omega_r \, a^{-2}$$

$$a \, da = H_o \sqrt{\Omega_r} \, dt$$

Integrating, we obtain,

$$\frac{1}{2} a^2 = H_0 \sqrt{\Omega_r} \, t$$

Or,

$$a(t) = \left(2\sqrt{\Omega_r} \, H_0 t\right)^{1/2} \quad \text{(Radiation Domination)}$$

Notice the Universe grows slower than in matter domination. For $\Omega_r = 1$ we have the age is:

$$t = \frac{1}{2} H_0^{-1}$$

Cosmological Constant

For a Universe with only cosmological constant, we have:

$$\frac{\dot{a}^2}{a^2} = H_0^2 \, \Omega_\Lambda$$

$$\left(\frac{da}{dt}\right)^2 = H_0^2 \, \Omega_\Lambda \, a^{-2}$$

$$\frac{da}{a} = H_o \sqrt{\Omega_\Lambda} \, dt$$

Integrating, we obtain,

$$\ln(a) = H_0 \sqrt{\Omega_\Lambda} \, t + const$$

and we find,

$$a(t) \propto exp\left(\sqrt{\Omega_\Lambda} \, H_0 t\right) \quad \text{(Cosmological Constant Domination)}$$

Notice the Universe grows exponentially fast in this case. In this case we cannot set a = 0 initially.

Curvature Dominated

For a Universe with only curvature, we have,

$$\frac{\dot{a}^2}{a^2} = H_0^2\,\Omega_k a^{-2}$$

$$\left(\frac{da}{dt}\right)^2 = H_0^2\,\Omega_k$$

$$da = H_o\sqrt{\Omega_k}\,dt$$

Integrating, we obtain,

$$a(t) = H_0\sqrt{\Omega_k}\,t \quad \text{(Curvature Domination)}$$

with $\Omega_k = 1$ the age is:

$$t = H_0^{-1}$$

Matter + Curvature

For a Universe with both matter and non-zero curvature, we have:

$$\frac{\dot{a}^2}{a^2} = H_0^2\left[\Omega_m a^{-3} + \Omega_k a^{-2}\right]$$

$$\left(\frac{da}{dt}\right)^2 = H_0^2\left[\Omega_m a^{-1} + \Omega_k\right]$$

Therefore,

$$\rightarrow H_0 dt = \frac{da}{\sqrt{\Omega_m a^{-1} + \Omega_k}}$$

$$= \frac{1}{\sqrt{\Omega_m}} \frac{a^{1/2}\,da}{\sqrt{1 + \left(\Omega_k / \Omega_m\right)a}}$$

It turns out that it is easier to first solve for the conformal time $d\eta = dt/a$. We have:

$$\eta = \int d\eta = \int \frac{dt}{a} = \frac{1}{H_0\sqrt{\Omega_m}} \int \frac{a^{-1/2}\,da}{\sqrt{1 + \left(\Omega_k / \Omega_m\right)a}}$$

Let us assume we have a closed universe, i.e. $k > 0$ and therefore $\Omega_k = -k / H_0^2 < 0$. Then let $u^2 = -\Omega_k / \Omega_m a$, so that $u = \sqrt{-\Omega_k / \Omega_m a}^{1/2}$ and $du = 1/2\sqrt{-\Omega_k / \Omega_m a}^{1/2}\, da$. We have:

$$
\begin{aligned}
\eta &= \frac{1}{H_0 \sqrt{\Omega_m}} 2\sqrt{\frac{\Omega_m}{\Omega_k}} \int \frac{du}{\sqrt{1-u^2}} \\
&= \frac{2}{H_0 \sqrt{-k}} \sin^{-1} u
\end{aligned}
$$

or inverting,

$$
\begin{aligned}
u &= sin(\theta / 2) \\
\theta &= H_0 \sqrt{-\Omega_k}\, \eta
\end{aligned}
$$

Under the same change of variables $(a \rightarrow u)$, $u^2\, du = 1/2(-\Omega_k / \Omega_m)^{3/2} a^{1/2}\, da$, and the equation for *t* becomes,

$$
\begin{aligned}
\mathrm{t} &= \frac{1}{H_0 \sqrt{\Omega_m}} \int \frac{a^{1/2}\, da}{\sqrt{1+(\Omega_k / \Omega_m)a}} \\
&= \frac{1}{H_0 \sqrt{\Omega_m}} 2\left(\frac{\Omega_m}{-\Omega_k}\right)^{3/2} \int \frac{u^2\, du}{\sqrt{1-u^2}} \\
&= \frac{2\Omega_m}{H_0 (-\Omega_k)^{3/2}} \int \frac{u^2\, du}{\sqrt{1-u^2}}
\end{aligned}
$$

Now changing $u = sin(\theta / 2)$, $du = cos(\theta / 2)\, d\theta / 2$, we have:

$$
\begin{aligned}
t &= \frac{2\Omega_m}{H_0 (-\Omega_k)^{3/2}} \int \frac{sin^2(\theta / 2) cos(\theta / 2)\, d\theta / 2}{\sqrt{1 - sin(\theta / 2)^2}} \\
&= \frac{\Omega_m}{H_0 (-\Omega_k)^{3/2}} \int sin^2(\theta / 2)\, d\theta
\end{aligned}
$$

Now using $\cos(\theta) = \cos^2(\theta) - \sin^2(\theta) = 1 - 2\sin^2(\theta / 2)$, we find:

$$
\begin{aligned}
t &= \frac{\Omega_m}{2H_0 (-\Omega_k)^{3/2}} \int 1 - cos(\theta)\, d\theta \\
&= \frac{\Omega_m}{2H_0 (-\Omega_k)^{3/2}} (\theta - sin(\theta))
\end{aligned}
$$

Finally, recall that $a = -(\Omega_m/\Omega_k)u^2 = -(\Omega_m/\Omega_k)\sin^2(\theta/2)$ so that we have a parametric solution for a cycloid:

$$a = \frac{\Omega_m}{-2\Omega_k}[1-cos(\theta)]$$

$$t = \frac{\Omega_m}{2H_0(-\Omega_k)^{3/2}}[\theta - sin(\theta)]$$

$$\theta = H_0\sqrt{-\Omega_k}\,\eta$$

Notice that for small values of θ, we have:

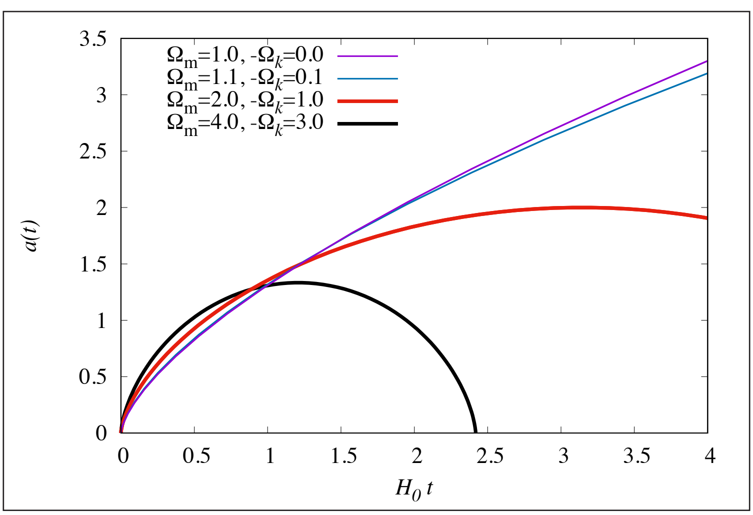
$$t \approx \frac{\Omega_m}{12H_0(-\Omega_k)^{3/2}}\theta^3 \rightarrow \theta = \left(\frac{12H_0}{\Omega_m}\right)^{1/3}(-\Omega_k)^{1/2}t^{1/3}$$

so that:

$$a \approx \frac{\Omega_m}{-4\Omega_k}\theta^2 = \frac{\Omega_m}{-8^{2/3}\Omega_k}\left(\frac{12H_0}{\Omega_m}\right)^{2/3}(-\Omega_k)t^{2/3} = \left(\frac{3}{2}\sqrt{\Omega_m}H_0 t\right)^{2/3}$$

Matter + Cosmological Constant

Figure shows scale factor a(t) as a function of H_0t (cosmic time normalized by the Hubble time H^{-1}_0) for a universe with only matter and curvature, with different values of Ω_m and $\Omega_k = 1 - \Omega_m$. Since $\Omega_k = k/H_0$, $\Omega_k < 0$ corresponds to a closed Universe (k > 0), which reaches a maximum turn-around scale factor $a_{ta} = \Omega_m/(\Omega_k)$ at time $H_0t_{ta} = (\pi/2)\,\Omega_m/(\Omega_k{}^{3/2})$. As $\Omega_k \rightarrow 0$, both a_{ta}, $t_{ta} \rightarrow \infty$ and the solution approaches that of a flat Universe without turn-around, i.e. a(t) = $\left(3/2\sqrt{\grave{U}_m}H_0\,t\right)^{2/3}$.



For a Universe with both matter and cosmological constant, we have:

$$\frac{\dot{a}^2}{a^2} = H_0^2\left[\Omega_m a^{-3} + \Omega_\Lambda\right]$$

$$\left(\frac{da}{dt}\right)^2 = H_0^2\left[\Omega_m a^{-1} + \Omega_\Lambda a^2\right]$$

Therefore,

$$\begin{aligned}
\rightarrow H_0 dt &= \frac{da}{\sqrt{\Omega_m a^{-1} + \Omega_\Lambda a^2}} \\
&= \frac{1}{\sqrt{\Omega_m}} \frac{a^{1/2}\, da}{\sqrt{1 + \left(\Omega_\Lambda / \Omega_m\right) a^3}} \quad \left(\text{Let } u^2 = \Omega_\Lambda / \Omega_m\, a^3\right) \\
&= \frac{2/3}{\sqrt{\Omega_\Lambda}} \frac{du}{\sqrt{1+u^2}}
\end{aligned}$$

Integrating, we obtain,

$$H_0\, t = \frac{2/3}{\sqrt{\Omega_\Lambda}} sinh^{-1}(u) = \frac{2/3}{\sqrt{\Omega_\Lambda}} sinh^{-1}\left(\sqrt{\frac{\Omega_\Lambda}{\Omega_m}} a^{3/2}\right)$$

and

$$a(t) = \left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{1/3} sinh^{2/3}\left(\frac{3\sqrt{\Omega_\Lambda} H_0}{2} t\right) \quad \left(\text{Matter} + \text{Cosmological Constant}\right)$$

which reduces to the matter dominated solution for small t:

$$a(t) \approx \left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{1/3} \left(\frac{3\sqrt{\Omega_\Lambda} H_0}{2}\right)^{2/3} t^{2/3} = \left(\frac{3}{2}\sqrt{\Omega_m} H_0 t\right)^{2/3}$$

and recovers the cosmological constant solution for large

$$t\left[\sinh\left(at\right) = \left(e^{at} - e^{-at}\right)/2 \rightarrow e^{at}/2\right]$$

$$a(t) \propto \left[\exp\left(\frac{3\sqrt{\Omega_\Lambda} H_0}{2} t\right)\right]^{2/3} = \exp\left(\sqrt{\Omega_\Lambda} H_0 t\right)$$

Photon Geodesics and Energy

Recall we defined the 4-momentum $P^{\alpha} = (E, p)$ for a massive particle as:

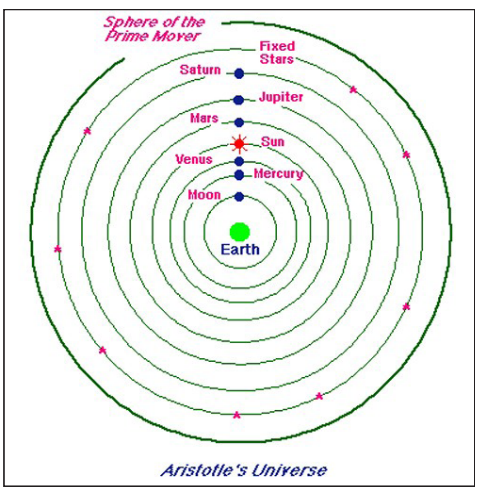
$$P^{\alpha} = m\frac{dx^{\alpha}}{d\tau}$$

But for a massless particle (e.g. a photon), both m = dτ = 0, so we need an alternative definition. We define it then with respect to a general implicit parameter λ along the particle trajectory:

$$P^{\alpha} = \frac{dx^{\alpha}}{d\lambda} = \left(\frac{dE}{d\lambda}, \frac{dp}{d\lambda}\right)$$

The Expanding Universe

Expanding Universe is the dynamic state of the extragalactic realm, the discovery of which transformed 20th-century cosmology. The development of general relativity and its application to cosmology by German-born physicist Albert Einstein, Dutch mathematician Willem de Sitter, and other theoreticians, along with the detection of extragalactic redshift (a shift to the longer wavelengths of light from galaxies beyond the Milky Way) by American astronomer Vesto Slipher, led to the realization in the 1920s that all galaxies are receding. American astronomer Edwin Hubble correlated these observations in mathematical form to provide evidence that the universe is expanding. The discovery of the 2.7 K cosmic microwave background radiation in 1965 by American physicists Arno Penzias and Robert Wilson was convincing evidence that the universe originated 13.8 billion years ago from a very dense and hot state in the big bang.



Geocentric universe of Aristotle and Ptolemy.

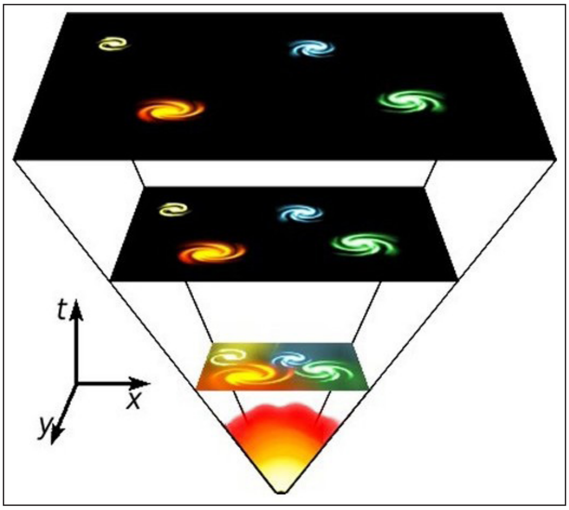
When Albert Einstein was formulating his ground-breaking theory of gravity in the early 20th Century, at a time when astronomers only really knew of the existence of our own galaxy, he necessarily used the simplifying assumption that the universe has the same gross properties in all parts, and that it looks roughly the same in every direction wherever in the universe an observer happens to be located. Like Sir Isaac Newton two hundred years before him, he assumed an infinite, static or "steady state" universe, with its stars suspended essentially motionless in a vast void.

However, when Einstein tried to apply his General Theory of Relativity to the universe as a whole, he realized that space-time as a whole must be warped and curved back on itself, which in itself would cause matter to move, shrinking uncontrollably under its own gravity. Thus, as early as 1917, Einstein and others realized that the equations of general relativity did not describe a static universe. However, he never quite came to terms with the idea of a dynamic, finite universe, and so he posited a mysterious counteracting force of cosmic repulsion (which he called the "cosmological constant") in order to maintain a stable, static universe. Adding additional and arbitrary terms to a theory is not something that scientists do lightly, and many people argued that it was an artificial and arbitrary construct and at best a stop-gap solution.

Up until that time, the assumption of a static universe had always been taken for granted. To put things into perspective, for most of history. It had been taken for granted that the static earth was the center of the entire universe, as Aristotle and Ptolemy had described. It was only in the mid-16th Century that Nicolaus Copernicus showed that we were not the center of the universe at all (or even of the Solar System for that matter!). It was as late as the beginning of the 20th Century that Jacobus Kapteyn's observations first suggested that the Sun was at the center of a spinning galaxy of stars making up the Milky Way. Then, in 1917, humanity suffered a further blow to its pride when Curtis Shapely revealed that we were not even the center of the galaxy, merely part of some unremarkable suburb of the Milky Way (although it was still assumed that the Milky Way was all there was).

Some years later, in 1925, the American astronomer Edwin Hubble stunned the scientific community by demonstrating that there was more to the universe than just our Milky Way galaxy and that there were, in fact, many separate islands of stars - thousands, perhaps millions of them, and many of them huge distances away from our own.

Then, in 1929, Hubble announced a further dramatic discovery which completely turned astronomy on its ear. With the benefit of improved telescopes, Hubble started to notice that the light coming from these galaxies was shifted a little towards the red end of the spectrum due to the Doppler effect (known as "redshift"), which indicated that the galaxies were moving away from us. After a detailed analysis of the redshifts of a special class of stars called Cepheids (which have specific properties making them useful as "standard candles" or distance markers), Hubble concluded that the galaxies and clusters of galaxies were in fact flying apart from each other at great speed, and that the universe was therefore definitively growing in size. In effect, all the galaxies we see are slightly red in color due to redshift.



Artist's impression of the "metric expansion" of the universe.

Hubble showed that, in our expanding universe, every galaxy is rushing away from us with a speed which is in direct proportion to its distance, known as Hubble's Law, so that a galaxy that is twice as far away as another is receding twice as fast, one ten times as far away if receding ten times as fast, etc. The law is usually stated as $v = H_o D$, where v is the velocity of recession, D is the distance of the galaxy from the observer and H_o is the Hubble constant which links them. The exact value of the Hubble constant itself has long been the subject of much controversy: Hubble's initial estimates were of the order of approximately 500 kilometers per second per megaparsec (equivalent to about 160 km/sec/million light years); the most recent best estimates, with the benefit of the Hubble Telescope and the WMAP probe, is around 72 kilometers per second per megaparsec. (It should perhaps be pointed out that the Hubble constant is technically a parameter, not a constant, because it will actually change over long periods of time.)

This expansion, usually referred to as the "metric expansion" of space, is a "broad-brush effect" in that individual galaxies themselves are not expanding, but the clusters of galaxies into which the matter of the universe has become divided are becoming more widely separated and more thinly spread throughout space. Thus, the universe is not expanding "outwards" into pre-existing space; space itself is expanding, defined by the relative separation of parts of the universe. Returning to the image of the expanding universe as a balloon inflating, if tiny dots are painted on the ballon to represent galaxies, then as the balloon expands so the distance between the dots increases, and the further apart the dots the faster they move apart. Another analogy often used (and maybe even clearer) is that of a raisin cake expanding as it bakes, so that the raisins (galaxies) gradually all move away from each other.

In such an expansion, then, the universe continues to look more or less the same from every galaxy, so the fact that we see all the galaxies receding from us does not necessarily mean that we are at the very center of the universe: Observers in all other galaxies would also see all the other galaxies flying away according to the same law, and the pattern of galactic dispersal would appear very much the same from anywhere in the cosmos.

The old model of a static universe, which had served since Sir Isaac Newton, was thus proved to be incontrovertibly false, but Hubble's discovery did more than just show that the universe was changing over time. If the galaxies were flying apart, then clearly, at some earlier time, the universe was smaller than at present. Following back logically, like a movie played in reverse, it must ultimately have had some beginning when it was very tiny indeed, an idea which gave rise to the theory of the Big Bang. Although now almost universally accepted, this theory of the beginnings of the universe was not immediately welcomed by everyone, and several strands of corroborating evidence were needed.

In the face of Hubble's evidence, Einstein was also forced to abandon his idea of a force of cosmic repulsion, calling it the "biggest blunder" he had ever made. But others, notably the Russian physicist Alexander Friedmann and the Belgian priest and physicist Georges Lemaître, had already used Einstein's own theory of proof that the universe was in fact in motion, either contracting or expanding. It is now recognized that Einstein's description of gravity as the curvature of space-time in his General Theory of Relativity was actually one of the first indications of a universe which had grown out of much humbler beginnings.

The Hubble's Law

The dominant motion in the universe is the smooth expansion known as Hubble's Law.

Recessional Velocity = Hubble's constant times distance,

$$V = H_o D$$

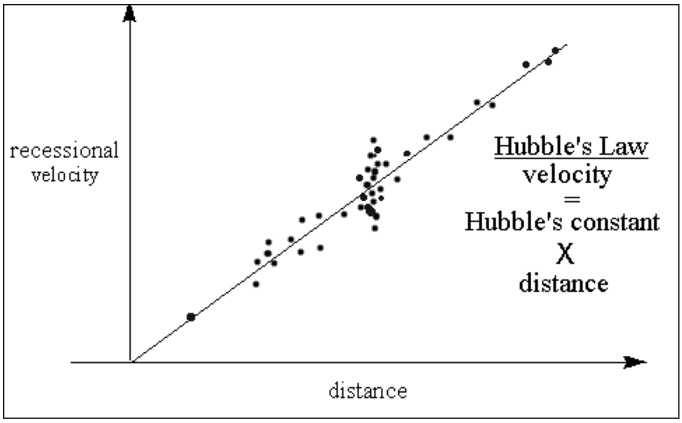
where,

V is the observed velocity of the galaxy away from us, usually in km/sec.

H is Hubble's "constant", in km/sec/Mpc.

D is the distance to the galaxy in Mpc.

In 1929, Hubble estimated the value of the expansion factor, now called the Hubble constant, to be about 500 km/sec/Mpc. Today the value is still rather uncertain, but is generally believed to be in the range of 45-90 km/sec/Mpc.



While in general galaxies follow the smooth expansion, the more distant ones moving faster away from us, other motions cause slight deviations from the line predicted by Hubble's Law. This diagram shows a typical plot of distance versus recessional velocity, with each point showing the relationship for an individual galaxy. In the example shown here, two things should be apparent:

- Few of the points fall exactly on the line. This is because all galaxies have some additional residual motion in addition to the pure expansion. This is referred to as the "cosmic velocity dispersion" or "cosmic scatter" and is probably due to the fact that the gas clouds that formed the galaxies all had some small additional motion of their own. The recessional velocity of a galaxy at a particular distance inferred from Hubble's law is called the "Hubble velocity".
- About in the middle of the diagram, there are a bunch of galaxies that appear to be at about the same distance but are spread out a lot in the velocity direction. This feature suggests the presence of a large cluster of galaxies, like the Virgo cluster. In addition to their "Hubble velocities", these galaxies have an extra velocity caused by their orbital motion around the center of the cluster. Because clusters of galaxies are very massive, this orbital velocity can be very large, more than 1000 km/s. Therefore in the vicinity of nearby clusters of galaxies, we cannot use Hubble's law to determine accurately the distance to the galaxy.

The Hubble Constant

The Hubble Constant (H_o) is one of the most important numbers in cosmology because it is needed to estimate the size and age of the universe. This long-sought number indicates the rate at which the universe is expanding, from the primordial "Big Bang."

The Hubble Constant can be used to determine the intrinsic brightness and masses of stars in nearby galaxies, examine those same properties in more distant galaxies and galaxy clusters, deduce the amount of dark matter present in the universe, obtain the scale size of faraway galaxy clusters, and serve as a test for theoretical cosmological models.

In 1929, American astronomer Edwin Hubble announced his discovery that galaxies, from all directions, appeared to be moving away from us. This phenomenon was observed as a displacement of known spectral lines towards the red-end of a galaxy's spectrum (when compared to the same spectral lines from a source on Earth). This redshift appeared to have a larger displacement for faint, presumably further, galaxies. Hence, the farther a galaxy, the faster it is receding from Earth.

The Hubble Constant can be stated as a simple mathematical expression, $H_o = v/d$, where v is the galaxy's radial outward velocity (in other words, motion along our line-of-sight), d is the galaxy's distance from earth, and H_o is the current value of the Hubble Constant.

However, obtaining a true value for H_o is very complicated. Astronomers need two measurements. First, spectroscopic observations reveal the galaxy's redshift, indicating its radial velocity. The second measurement, the most difficult value to determine, is the galaxy's precise distance from earth. Reliable "distance indicators," such as variable stars and supernovae, must be found in galaxies. The value of Ho itself must be cautiously derived from a sample of galaxies that are far enough away that motions due to local gravitational influences are negligibly small.

The units of the Hubble Constant are "kilometers per second per megaparsec." In other words, for each megaparsec of distance, the velocity of a distant object appears to increase by some value. (A megaparsec is 3.26 million light-years.) For example, if the Hubble Constant was determined to be 50 km/s/Mpc, a galaxy at 10 Mpc, would have a redshift corresponding to a radial velocity of 500 km/s.

The value of the Hubble Constant initially obtained by Edwin Hubble was around 500 km/s/Mpc, and has since been radically revised because initial assumptions about stars yielded underestimated distances.

For the past three decades, there have been two major lines of investigation into the Hubble Constant. One team, associated with Allan Sandage of the Carnegie Institutions, has derived a value for Ho around 50 km/s/Mpc. The other team, associated with Gerard DeVaucouleurs of the University of Texas, has obtained values that indicate Ho to be around 100 km/s/Mpc.

The Steady State Theory

In cosmology, the Steady State theory (also known as the Infinite Universe theory or continuous creation) is a model developed in 1948 by Fred Hoyle, Thomas Gold, Hermann Bondi and others as an alternative to the Big Bang theory (known, usually, as the standard cosmological model). In steady state views, new matter is continuously created as the universe expands, so that the perfect cosmological principle is adhered to. Although the model had a large number of supporters among cosmologists in the 1950s and 1960s, the number of supporters decreased markedly in the late 1960s with the discovery of the cosmic microwave background radiation, and today only a very small number of supporters remain. The key importance of the steady-state model is that as a competitor to the Big Bang, it was an impetus in generating some of the most important research in astrophysics, much of which ultimately ended up supporting the Big Bang theory.

The Steady State Theory of Bondi, Gold and Hoyle was inspired by the circular plot of the film Dead of Night they watched together. Theoretical calculations showed that a static universe was impossible under general relativity and observations by Edwin Hubble had shown that the universe was expanding. The steady state theory asserts that although the universe is expanding, it nevertheless does not change its look over time (the perfect cosmological principle); it has no beginning and no end.

The theory requires that new matter must be continuously created (mostly as hydrogen) to keep the average density of matter equal over time. The amount required is low and not directly detectable: roughly one solar mass of baryons per cubic megaparsec per year or roughly one hydrogen atom per cubic meter per billion years, with roughly five times as much dark matter. Such a creation rate would, however, cause observable effects on cosmological scales.

An aesthetically unattractive feature of the theory is that the postulated spontaneous new matter formation would presumably need to include deuterium, helium, and a small amount of lithium, as well as regular hydrogen, since no mechanism of nucleosynthesis in stars or by other processes accounts for the observed abundance of deuterium and helium-3. (In the Big Bang model, primordial deuterium is made directly after the "bang," before the existence of the first stars).

Chaotic inflation theory has many similarities with Steady State Theory, however on a much larger scale than originally envisaged.

C-field

Bondi and Gold proposed no mechanism for the creation of matter required by the Steady State Theory, but Hoyle proposed the existence of what he called the "C-field," where "C" stands for "Creation." The C-field has negative pressure, which enables it to drive the steady expansion of the cosmos, whilst also creating new matter, keeping the large-scale matter density approximately constant; in this respect the C-field is similar to the inflaton field used in cosmic inflation. For this reason Hoyle's conception of the steady state in 1948 incorporates many features that later emerged in both inflationary cosmology and the recently observed accelerating universe, which may be modeled in terms of a cosmological constant in Einstein's model of the universe.

The C-field and the notion of quasi-steady state universe also has some resemblance to chaotic inflation theory or eternal inflation which sometimes posits an infinite universe with neither beginning nor end, in which inflation operates continuously, on a scale beyond the observable universe, to create the matter of the cosmos. However, both steady state and quasi-steady state assert that the creation events of the universe (new hydrogen atoms in the steady state case) can be observed within the observable universe, whereas inflationary theories do not posit inflation as an ongoing process within the observable universe.

Quasi-steady State

Quasi-steady state cosmology (QSS) was proposed in 1993 by Fred Hoyle, Geoffrey Burbidge, and Jayant V. Narlikar as a new version of steady state ideas, intended to explain additional features unaccounted for in the initial proposal. The theory suggests pockets of creation occurring over time within the universe, sometimes referred to as minibangs, mini-creation events, or little bangs. After the observation of an accelerating universe, further modifications of the model were made. Mainstream cosmologists who have reviewed QSS have pointed out flaws and discrepancies with observations left unexplained by proponents.

Arguments for and against the Steady State Theory

There are a number of observations that astronomers have made to test cosmological theories, including both the steady–state and the big bang theory. Some of these cosmological observations are described below.

Evolution of the Universe

When astronomers look at the most distant objects in the universe, they are looking back in time. For example if one observes a quasar that is three billion light–years away, it has taken the light three billion years to get here, because a light–year is the distance light travels in one year when in a vacuum. Astronomers are therefore seeing the quasar as it looked three billion years ago. Quasars, the most distant objects known in the universe, are thought to be very active nuclei of distant galaxies. The nearest quasar is about a billion light–years away. The fact that astronomers do not see any quasar closer than a billion light years away suggests that quasars disappeared at least a billion years ago. The universe has changed with time. Several billion years ago, quasars existed; they no longer do. This observation provides evidence that the perfect cosmological principle is untrue, and therefore that the steady–state theory is incorrect. Note, however, that when the steady–state theory and the perfect cosmological principle were first suggested, scientists had not yet discovered quasars.

Expansion of the Universe

In his work measuring distances to galaxies, American astronomer Edwin Hubble, after whom the Hubble space telescope was named, noticed an interesting correlation. The more distant a galaxy is, the faster it is moving away from the Earth. This relationship is called the Hubble law. This relationship can be used to find the distances to additional galaxies, by measuring the speed of recession. More importantly, Hubble deduced the cause of this correlation. His result: the universe is expanding. To visualize this expansion, draw some galaxies on an ordinary balloon and blow it up. Notice how the galaxies move farther apart as the balloon expands. Measuring distances between the galaxies at the rates at which they move apart, would give a relationship similar to Hubble's law.

The expanding universe can be consistent with either the big bang or the steady–state theory. However in the steady–state theory, new matter must appear to fill in the gaps left by the expansion. Normally as the universe expands, the average distance between galaxies would increase as the density of the universe decreases. These evolutionary changes with time would violate the fundamental assumption behind the steady–state theory. Therefore, in the steady–state theory, hydrogen atoms appear out of empty space and collect to form new galaxies. With these new galaxies, the average distance between galaxies remains the same even in an expanding universe.

The Hubble plot also provides evidence that the universe changes with time. The slope of the Hubble plot gives scientists the rate at which the universe is expanding. If the universe is not evolving, this slope should remain the same even for very distant galaxies. The measurements are difficult, but the Hubble plot seems to curve upward for the most distant galaxies. The universe was expanding faster in the distant past, contrary to the prediction of the steady–state theory that the universe is not evolving.

Cosmic Background Radiation

In the mid 1960s, German–born American physicist Arno Allan Penzias and American physicist Robert Woodrow Wilson were working on a low noise (static) microwave antenna when they made an accidental discovery of cosmic significance. After doing everything possible to eliminate sources of noise, including cleaning out nesting pigeons and their waste, there was still a small noise component left. This weak noise did not vary with direction or with the time of day or year, because it was cosmic in origin. It also corresponded to a temperature of about 3K (where K = Kelvin;–518 °F;–270 °C, three degrees above absolute zero). This 3K cosmic background radiation turned out to be the leftover heat from the initial Big Bang that had been predicted by proponents of the Big Bang theory as early as the 1940s.

Because this cosmic background radiation was a prior prediction of the big bang theory, it provided strong evidence to support the big bang theory. Proponents of the steady–state theory have been unable to explain in detail how this background radiation could arise in a steady–state universe. The cosmic background radiation therefore gave the steady–state theory its most serious setback. Penzias and Wilson received the 1978 Nobel Prize in physics for their work.

Steady–state Theory

The steady–state theory was inspired at least in part by a 1940s movie entitled Dead of Night. The movie had four parts and a circular structure such that at the end the movie was the same as at the beginning. After seeing this movie in 1946, Austrian–born American astrophysicist Thomas Gold, Austrian–English mathematician and cosmologist Sir Hermann Bondi, and English astronomer Sir Fred Hoyle wondered if the universe might not be constructed the same way. The discussion that followed led ultimately to the steady–state theory.

In 1948, Bondi and Gold proposed extending the cosmological principle to the perfect cosmological principle, so that the universe looks the same at all times as well as at all locations. They then proposed the steady–state theory based on the new perfect cosmological principle. Because Hubble had already observed that the universe is expanding, Bondi and Gold proposed the continuous creation of matter. Hydrogen atoms created from nothing combine to form galaxies. In this manner, the average density of the universe remains the same as the universe expands. In the steady–state theory, the rate at which new matter is created must exactly balance the rate at which the universe is expanding. Otherwise, the average density of the universe will change and the universe will evolve, violating the perfect cosmological principle. To maintain the steady–state, in a cubic meter of space one hydrogen atom must appear out of nothing every five billion years. In a volume of space the size of the Earth, the amount of new matter created would amount to roughly a grain of dust in a million years. In the entire observable universe, roughly one new galaxy per year will form from these atoms. Bondi and Gold recognized that a new theory must be developed to explain how the hydrogen atoms formed out of nothing, but did not suggest a new theory.

In the same year, Hoyle proposed a modification of German–American physicist Albert Einstein's general theory of relativity. Hoyle worked independently of Bondi and Gold, but they did discuss the new theories. Hoyle's modification used a mathematical device to allow the creation of matter from nothing, as implied in general relativity. No experiments or observations have been made to justify or contradict this modification of general relativity.

There are a number of problems with the steady–state theory, but at the time the theory was proposed, there were also points in its favor.

The steady–state theory rests on the foundation of the perfect cosmological principle. Hence, any evidence that the universe evolves is evidence against the steady–state theory. The existence of quasars and the change in the expansion rate of the universe a few billion years in the past, are evidence against steady–state. This evidence for the evolution of the universe did not exist in 1948, when the steady–state theory originated. It became part of the cumulative weight of evidence, which had built up against the steady–state theory, by the mid 1960s. In gallant attempts to save the steady–state theory, its proponents, chiefly Hoyle and Indian astrophysicist Jayant Vishnu Narlikar, have argued that the universe can change over time periods of a few billion years without violating the perfect cosmological principle. Cosmologists (scientists that study the universe) must look at even longer time spans to see that these changes with time average out.

The cosmic background radiation is widely considered the final blow to the steady–state theory. Again, proponents of the steady–state theory have made gallant efforts to save their theory in the face of what most astronomers consider overwhelming evidence. They argue that the background radiation could be the cumulative radiation of a large number of radio sources that are too faint to detect individually. This scheme requires the existence of roughly 100 trillion (about 10,000 times the number of observable galaxies) such sources that are about one millionth as bright as the radio sources astronomers do detect. Few astronomers are willing to go to such great lengths to rescue the steady–state theory.

Another objection raised against the steady–state theory is that it violates one of the fundamental laws of physics as that law is currently understood. The law of conservation of matter and energy states that matter and energy are interchangeable and can change between forms, but the total amount of matter and energy in the universe must remain constant. It can be neither created nor destroyed. The steady–state theory requires continuous creation of matter in violation of this law. However, laws of science result from experimental evidence and are subject to change, not at mere whim, but as experimental results dictate. The rate at which matter is created in the steady–state theory is small enough that normally it would not have been noticed. Hence, scientists would not have discovered experimentally any conditions under which matter could be created or any modifications required in this law.

Were there ever any points in favor of the steady–state theory? When the steady–state model was first suggested, the best estimate of the age of the universe in the context of the big bang model was about two billion years. However, the Earth and solar system are about five billion years old. The oldest stars in the Milky Way galaxy are at least 10 to 12 billion years old. These age estimates present the obvious problem of a universe younger than the objects it contains. This problem is no longer so severe. Modern estimates for the age of the universe range from about 10 billion years to about 20 billion years, with a currently accepted average of about 13.7 billion years. In the steady–state theory the universe has always existed, so there are no problems presented by the ages of objects in the universe.

For some people there are philosophical or esthetic grounds for preferring the steady–state hypothesis over the big bang theory. The big bang theory has a moment of creation, which some people prefer for personal or theological reasons. Those who do not share this preference often favor the steady–state hypothesis. They prefer the grand sweep of a universe that has always existed to a universe that had a moment of creation and may, by inference, also have an end in some far distant future time.

Hubble Parameter and Scale Factor

Hubble's Constant with Fractional Rate of Change of Scale Factor

Here, we will relate the Hubble's Constant with fractional rate of Change of Scale Factor. We can write velocity in the following manner and simplify.

$$v = \frac{dr_p}{dt}$$

$$= \frac{d[a(t)r_c}{dt}$$

$$v = \frac{da}{dt} \times \frac{1}{a} \times (ar_c)$$

$$v = \frac{da}{dt} \times \frac{1}{a} \times r_c$$

Here, v is the recessional velocity, a is the scale factor and r_p is the proper distance between the galaxies. Hubble's Empirical Formula was of the nature:

$$v = H \times r_p$$

Thus, comparing the above two equations we obtain:

Hubble's Parameter = Fractional rate of change of the scale factor

$$H = da / dt \times 1 / a$$

This is not a constant since the scale factor is a function of time. Hence it is called the Hubble's parameter and not the Hubble's constant.

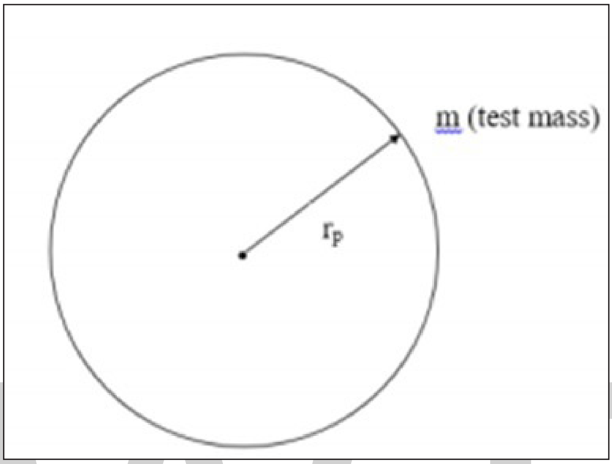
Empirically we write:

$$H = V / D$$

Thus, from this equation, we can infer that since D is increasing and V is a constant, then H reduces with the time and expansion of the universe.

Friedmann Equation in Conjunction with the Robertson-walker Model

Here, we will understand how the Friedmann Equation is used in conjunction with the Robertson-Walker model. To understand this, let us take the following image which has a test mass at distance r_p from body of mass M as an example.



Taking into consideration the above image, we can express force as:

$$F = G \times M \times \frac{m}{r_p^2}$$

Here, G is the universal gravitational constant and ρ is the matter density inside the observable universe.

Now, assuming uniform mass density within the sphere we can write:

$$M = \frac{4}{3} \times \pi \times r_p^3 \times \rho$$

Using these back in our force equation we get:

$$F = \frac{4}{3} \times \pi \times G \times r_p \times \rho \times m$$

Thus, we can write the potential energy and kinetic energy of the mass m as:

$$V = -\frac{4}{3} \times \pi \times G \times r_p^2 \times m \times \rho$$

$$K.E. = - \times m \times \frac{dr}{dt}$$

Using the Virial Theorem:

$$U = K.E + V$$

$$U = \frac{1}{2} \times m \times \left(\frac{dr_p}{dt}\right)^2 - \frac{4}{3} \times \pi \times G \times r_p^2 \times m \times \rho$$

But here, $r_p = ar_c$. So, we get:

$$U = \frac{1}{2} \times m \times \left(\frac{da}{dt}\right)^2 r_c^2 - \frac{4}{3} \times \pi \times G \times r_p^2 \times m \times \rho$$

On further simplification, we obtain the Friedmann equation,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3} \times G \times \rho + \frac{2U}{m} \times r_c^2 \times a^2$$

Here *U* is a constant. We also note that the universe we live in at present is dominated by matter, while the radiation energy density is very low.

Cosmological Redshift

Laboratory experiments here on Earth have determined that each element in the periodic table emits photons only at certain wavelengths (determined by the excitation state of the atoms). These photons are manifest as either emission or absorption lines in the spectrum of an astronomical object, and by measuring the position of these spectral lines, we can determine which elements are present in the object itself or along the line of sight.

However, when astronomers perform this analysis, they note that for most astronomical objects, the observed spectral lines are all shifted to longer (redder) wavelengths. This is known as 'cosmological redshift' (or more commonly just 'redshift') and is given by:

$$z = \frac{\lambda_{obs} - \lambda_{rest}}{\lambda_{rest}}$$

For relatively nearby objects, where z is the cosmological redshift, λ_{obs} is the observed wavelength and λ_{rest} is the emitted/absorbed wavelength.

Caused solely by the expansion of the Universe, the value of the cosmological redshift indicates the recession velocity of the object, or its distance. For small velocities (much less than the speed of light), cosmological redshift is related to recession velocity (v) through:

$$z \approx \frac{v}{c}$$

where, c the speed of light. At larger distances (higher redshifts), using the theory of general relativity gives a more accurate relation for recession velocities, which can be greater than the speed of light. Note this doesn't break the ultimate speed limit of c in Special Relativity as nothing is actually moving at that speed, rather the entire distance between the receding object and us is increasing. This is a complex formula requiring knowledge of the overall expansion history of the universe to calculate correctly but a simple recession velocity is given by multiplying the comoving distance (D) of the object by the Hubble parameter at that redshift (H) as:

$$z \approx \frac{HD}{v} - 1$$

Although cosmological redshift at first appears to be a similar effect to the more familiar Doppler shift, there is a distinction. In Doppler Shift, the wavelength of the emitted radiation depends on the motion of the object at the instant the photons are emitted. If the object is travelling towards us, the wavelength is shifted towards the blue end of the spectrum, if the object is travelling away from us, the wavelength is shifted towards the red end. In cosmological redshift, the wavelength at which the radiation is originally emitted is lengthened as it travels through (expanding) space. Cosmological redshift results from the expansion of space itself and not from the motion of an individual body.

For example, in a distant binary system it is theoretically possible to measure both a Doppler shift and a cosmological redshift. The Doppler shift would be determined by the motions of the individual stars in the binary – whether they were approaching or receding at the time the photons were emitted. The cosmological redshift would be determined by how far away the system was when the photons were emitted. The larger the distance to the system, the longer the emitted photons have travelled through expanding space and the higher the measured cosmological redshift.

References

- Universe, science: britannica.com, Retrieved 05 May, 2019
- Shape-and-fate-universe: plus.maths.org, Retrieved 10 February, 2019
- HadronEpoch: historyoftheuniverse.com, Retrieved 15 April, 2019

- Inflation-zero, outreach-origins: ctc.cam.ac.uk, Retrieved 08 August, 2019
- Friedman, cosmo: jb.man.ac.uk, Retrieved 28 January, 2019
- Hubbles-law: hosting.astro.cornell.edu, Retrieved 12 April, 2019
- Topics-bigbang-expanding: physicsoftheuniverse.com, Retrieved 16 May, 2019
- Steady-state-theory: newworldencyclopedia.org, Retrieved 05 June, 2019
- Cosmology-hubble-parameter-and-scale-factor, cosmology: tutorialspoint.com, Retrieved 04 March, 2019

Permissions

All chapters in this book are published with permission under the Creative Commons Attribution Share Alike License or equivalent. Every chapter published in this book has been scrutinized by our experts. Their significance has been extensively debated. The topics covered herein carry significant information for a comprehensive understanding. They may even be implemented as practical applications or may be referred to as a beginning point for further studies.

We would like to thank the editorial team for lending their expertise to make the book truly unique. They have played a crucial role in the development of this book. Without their invaluable contributions this book wouldn't have been possible. They have made vital efforts to compile up to date information on the varied aspects of this subject to make this book a valuable addition to the collection of many professionals and students.

This book was conceptualized with the vision of imparting up-to-date and integrated information in this field. To ensure the same, a matchless editorial board was set up. Every individual on the board went through rigorous rounds of assessment to prove their worth. After which they invested a large part of their time researching and compiling the most relevant data for our readers.

The editorial board has been involved in producing this book since its inception. They have spent rigorous hours researching and exploring the diverse topics which have resulted in the successful publishing of this book. They have passed on their knowledge of decades through this book. To expedite this challenging task, the publisher supported the team at every step. A small team of assistant editors was also appointed to further simplify the editing procedure and attain best results for the readers.

Apart from the editorial board, the designing team has also invested a significant amount of their time in understanding the subject and creating the most relevant covers. They scrutinized every image to scout for the most suitable representation of the subject and create an appropriate cover for the book.

The publishing team has been an ardent support to the editorial, designing and production team. Their endless efforts to recruit the best for this project, has resulted in the accomplishment of this book. They are a veteran in the field of academics and their pool of knowledge is as vast as their experience in printing. Their expertise and guidance has proved useful at every step. Their uncompromising quality standards have made this book an exceptional effort. Their encouragement from time to time has been an inspiration for everyone.

The publisher and the editorial board hope that this book will prove to be a valuable piece of knowledge for students, practitioners and scholars across the globe.

Index