1.31: Galaxy Formation

After recombination, baryons fall into the gravitational potential wells provided by dark matter halos. As hydrogen gas falls into halos, it gains kinetic energy and heats up to a temperate roughly determined by the halo mass; for a certain range of halo masses, cooling mechanisms allow the gas to cool, condense near the center of the halo, and begin the process of star and galaxy formation. The first generations of stars and galaxies flood the universe with high-energy radiation, reionizing neutral hydrogen in the intergalactic medium. A rich set of feedback processes then shape the evolution of galaxy populations over subsequent epochs.

Like nonlinear dark matter structure formation, galaxy formation and evolution are complex processes that generally require numerical simulations to model accurately. This Chapter provides a broad overview of galaxy formation theory and the connection between galaxies and dark matter halos, which are important tools for cosmological analyses based on galaxy populations.

How do Galaxies Form?

After recombination, most of the baryons in the universe are in the form of neutral hydrogen. During this epoch, dark matter overdensities are growing ($\delta \propto a$) and collapsing into dark matter halos. Consider a region in which hydrogen gas is gravitationally attracted towards the center of a halo; as it falls in, it heats up to a *virial temperature*

$$T_{\mathrm{vir}} \approx 10^5 \mathrm{~K}\left(\frac{V_{\mathrm{vir}}}{100 \mathrm{~km} \mathrm{~s}^{-1}}\right)^2, \tag{1.31.1}$$

where V_{vir} is the *virial velocity* of the halo, which is set by its mass and density profile. For example, a halo of mass $10^{12}\ M_{\odot}$ ($10^9\ M_{\odot}$) roughly has $V_{\text{vir}} \approx 200 \mathrm{~km~s}^{-1}$ ($V_{\text{vir}} \approx 20 \mathrm{~km~s}^{-1}$). In this way, the halo's gravitational potential sets the temperature of the infalling gas.

Gravitational acceleration isn't the only process that sets the gas temperature. In particular, hydrogen undergoes electromagnetic interactions including *collisional excitation*, *ionization*, *recombination*, and *bremsstrahlung* with photons and free electrons, which largely act to cool the gas at these early times. The net effect of these processes is captured by the *cooling function* $\Lambda(T)$, which is defined to set the characteristic cooling time via

$$t_{\text{cool}} = \frac{3 n k_B T}{2 n_{\mathrm{H}}^2 \Lambda(T)}, \tag{1.31.2}$$

where n is the total number density of gas particles and n_{H} is the number density of neutral hydrogen atoms in the gas. If the cooling time is shorter than the free-fall time, the gas cools more quickly than it can gravitationally equilibrate and condenses to the center of the halo.

Let's estimate the minimum virial velocity necessary to efficiently cool atomic hydrogen. Other than bremsstrahlung, the cooling processes mentioned above all involve electron transitions between hydrogen energy levels, the lowest of which are separated by 13.6 eV. Thus, the gas must be warmer than roughly 10 eV to initiate these processes, corresponding to $T_{\text{vir,min}} \approx 10^4 \mathrm{~K}$ and $V_{\text{vir,min}} \approx 10 \mathrm{~km~s}^{-1}$. This threshold is known as the *atomic cooling limit*, and it corresponds to a halo mass of roughly $10^8\ M_{\odot}$ today. Problem 32.1 explores how the cooling time depends on redshift, gas density, and gas composition.

Box 1.31.1

Exercise 32.1.1: Estimate the minimum halo mass necessary to efficiently cool molecular hydrogen (H_2). To convert from virial velocity to halo mass, you can combine the estimates of this relation above with $V_{\text{vir}} \propto M^{1/3}$.

As a result of the cooling processes described above, sufficiently dense regions of cold gas clump into *giant molecular clouds* (GMCs) that host star formation. The dynamic range of the star formation process is remarkable: gas accretes into halos on scales of hundreds of kiloparsecs (roughly corresponding to the size of a typical halo), condenses into GMCs with characteristic sizes of tens of parsecs, and forms into stars with sizes of $\approx 10^{-7}$ parsecs. Importantly, star formation in GMCs is highly inefficient, with typical star formation timescales that are orders of magnitude larger than the corresponding free-fall times.

The first generations of stars form in dark matter halos about 100 million years after the Big Bang, transitioning the universe from post-recombination "dark ages" to the epoch of "cosmic dawn." These stars eventually die, exploding in supernovae and flooding the universe with high-energy radiation, which heats the gas in halos, inhibiting star formation. This radiation also reionizes hydrogen in the intergalactic medium, transitioning the universe from a dark, opaque neutral hydrogen gas to a transparent, ionized plasma. Several independent lines of evidence, including the optical depth of the cosmic microwave background and the absorption spectra of high-redshift quasars, indicate that the universe is fully reionized by $z \approx 6$, roughly 1 billion years after the Big Bang.

Simulation of the end of the "cosmic dark ages," the formation of the first stars and galaxies during "cosmic dawn," and the epoch of reionization. Dark regions consist of neutral hydrogen, which fills the universe after recombination. Bright regions show ionized bubbles emanating from the first generations of stars as they explode in supernovae. These ionized bubbles expand to fill the entire simulation volume by the time reionization is complete, about 1 billion years after the Big Bang.

How do Galaxies Evolve?

Early models and simulations of galaxy formation drastically overpredicted the number of galaxies as a function of galaxy luminosity, or the *galaxy luminosity function*, compared to observations. This issue, known as the "overcooling problem," is overcome by *feedback*---the effects of galaxy formation on the process itself. The supernova and photoionization heating mechanisms discussed above constitute one form of feedback. These mechanisms most strongly affect galaxies that inhabit halos with masses below $\approx 10^{12}\ M_{\odot}$ by heating gas to a significant fraction of these halos' virial temperatures. For halos with masses below $\approx 10^{10}\ M_{\odot}$, photoionization heating partially or completely shuts down subsequent star formation; we observe the relic of ancient star formation in these small halos as *dwarf galaxies* today.

The overcooling problem also applied to halos with masses above $\approx 10^{12}\ M_{\odot}$, which were predicted to form stars more efficiently than inferred from observations. Supernova feedback and photoionization heating do not significantly affect star formation in these halos due to their large virial temperatures. Instead, high-energy radiation from the matter orbiting supermassive black holes (SMBHs) at the centers of massive galaxies heats the surrounding gas. This mechanism, known as *active galactic nucleus (AGN) feedback*, helps bring the predicted luminosity function for the brightest galaxies into agreement with observations. Note that AGN feedback is ineffective in lower-mass systems because SMBH mass (and thus the amount of energy AGN feedback releases) decreases rapidly in smaller halos. The left panel of the figure below illustrates the resulting *galaxy--halo connection*; specifically, the relation between stellar mass and total halo mass, and the halo mass ranges in which various forms of feedback affect galaxy formation.

Galaxy properties, including (but not limited to) stellar mass, star formation rate, size, shape, color, and visual appearance (or *morphology*) evolve over cosmic time. The interplay between cooling, star formation, and feedback causes the global star formation rate in the universe to peak roughly 10 billion years after the Big Bang at $z \approx 2$, or "cosmic noon," as shown in the right panel below. Note that the rise and fall of AGN activity in the universe roughly coincides with the rise and fall of the global star formation rate.

Wechsler & Tinker 2018. *Right panel*: The average cosmic star formation rate per unit volume as a function of redshift (bottom axis) and lookback time (top axis). Points with errorbars show observational measurements, and the line shows a best-fitting star formation rate density model. Adapted from Madau & Dickinson 2014.

Galaxy morphology is historically classified according to the *Hubble sequence*, which delineates four major types of galaxies based on their visual appearance: elliptical, lenticular, spiral, and irregular. These classes of galaxies are arranged in a "tuning fork" sequence, which does **not** correspond to how galaxies typically evolve. In particular, although ellipticals and spirals are respectively referred to as "early-type" and "late-type" galaxies, modern observations indicate that the first generations of active, star-forming galaxies are predominantly spirals and irregulars. Mergers between these classes of galaxies yield ellipticals; this process often coincides with a shutdown in star formation, such that ellipticals are generally redder and less actively star-forming compared to spirals. Problem 32.2 explores how different galaxy morphologies arise and the relation between galaxy and halo evolution.

An updated version of the original Hubble sequence, which delineates four major types of galaxies: elliptical (E), lenticular (S0), spiral (S), and irregular (I). Note that galaxies do not generally evolve from left to right along the Hubble sequence. Instead, the universe is dominated by spiral and irregular galaxies at early times, which merge to form elliptical galaxies. Adapted from Kormendy & Bender 1996.

HOMEWORK Problems

Problem 1.31.1: Behavior of the cooling time.

Consider the behavior of the cooling time as a function of density, redshift, and gas composition:

a) Starting from Equation 1.31.2, argue that $t_{\text{cool}} \propto \rho_g^{-1}$, where ρ_g is the gas density. Qualitatively explain why an inverse dependence on gas density is reasonable. What does this dependence imply about the temperature profile of the gas as a function of radius from the center of a halo?

b) Use the result from part a) to show that $t_{\text{cool}} \propto a^3$. Qualitatively describe what consequences more efficient cooling at early times might have for galaxy formation.

c) Qualitatively describe how the presence of heavier elements affects the cooling time and minimum virial temperature necessary for efficient cooling. This is an important consideration because stars deposit heavier elements into the interstellar medium when they explode in supernovae.

Problem 1.31.2: Galaxy morphology and evolution.

This problem walks through a few additional aspects of galaxy morphology and evolution.

a) Using conservation of angular momentum, qualitatively describe why thin galactic disks form out of gas that falls into halos from the intergalactic medium.

b) When two spiral galaxies merge, qualitatively describe why the resulting stellar distribution is extended and elliptical.

c) Considering that small halos form first and merge together to build larger halos and how galaxy properties evolve as a result of mergers, qualitatively describe how you expect the stellar mass--halo mass relation to evolve over cosmic time.