Overview of Galaxy Formation Theory

1. Basics

1.1. General Story

Galaxies form through the nonlinear gravitational collapse of dark matter halos. Baryons participate in this collapse with the dark matter, but unlike dark matter are able to radiate energy away and therefore sink deeper into the potential well. They cannot radiate away angular momentum, and they experience gas pressure; this combination leads ultimately to the formation of rotating gas disks. In this process, pockets of the gas cool, becoming neutral or molecular. Inside the molecular regions, individual stars collapse and cool. Through stellar processes, these stars enrich the interstellar medium and return energy to it through feedback. At the centers of the galaxies, black holes form and grow, and also exert feedback on the gas. These feedback effects may help determine the star formation rates of the forming galaxies. As these events occur, galaxies experience further accretion and major mergers. Although there is not an agreed-upon quantitative understanding of this whole process, there is strong evidence that the dark matter halos with masses near $10^{12}\ M_\odot$.

Although this basic story is known, many of its details are unclear and it is not known if the story is complete. The current theories of galaxy formation do a good job reproducing the gross properties of galaxies near $z \sim 0$. It remains to be seen if they consistently predict the differences in the galaxy population at higher redshift, more detailed properties of galaxies such as their mass profiles, stellar kinematics, and internal chemical patterns, and their gaseous environments.

1.2. Halos and Galaxies

An overarching challenge for physical galaxy formation theory can be expressed as the relationship between halo mass and stellar mass of galaxies. Approximately, this relationship expresses how effiently the matter that fell into a halo was converted into star.

This relationship can be observationally constrained through *abundance matching*. Abundance matching assumes that halo mass and stellar mass are at least approximately monotonically related. For any halo mass M_h for which the number density of halos above that mass is predicted to be $\Phi_h(> M_h)$, one can find a corresponding stellar mass M_* for which the real universe is observed to have $\Phi_*(> M_*) = \Phi(> M_h)$, and conclude that halos of mass M_h host galaxies with stellar mass M_*. There are numerous refinements of this technique that account for scatter in the relationship between M_h and M_*, that use subhalos rather than halos, that use circular velocity instead of M_h.

The other basic tool for understanding this relationship is the *halo occupation distribution* model. This model connects N-body predictions of dark matter halo properties to observed galaxy

properties through $P(> M_*|M_h)$, the probability distribution of the number of galaxies above stellar mass M_* within a halo of mass M_h. Many refinements of this model exist, the most important of which is the distinction between *central galaxies* in a halo and *satellite galaxies*. The framework also can be used to study galaxies of different classes or properties, and as a function of different environments around each halo.

Wechsler & Tinker (2018) review the literature of the galaxy-halo connection, including abundance matching and halo occupation distributions. The major conclusion from these studies is that M_*/M_h peaks at $M_h \sim 10^{12}$. This ratio rises linearly or even more steeply between 10^{10} and 10^{12} $M_\odot$ and declines as $\sim M_h^{-1/2}$ at higher masses. These conclusions are validated by using them to predict observations of weak lensing and galaxy correlation functions. They can be further used to study the scatter in the relationship between M_h and M_*, how central and satellite galaxies differ in their star formation histories, and how these properties depend on environment.

These results outline a basic challenge of galaxy formation theories, which is to explain how the efficiency of star formation depends on halo mass.

1.3. Physical Processes

Theoretical models seek to explain galaxy-halo relation and the other properties of galaxies from, most ambitiously, a first-principles approach beginning with pattern of matter fluctuations at the time of recombination. Somerville & Davé (2015) outline the major physical processes at play:

- Gravity, which drives gravitational growth leading to the collapse and clustering of dark matter halos.
- Hydrodynamics, which controls the flow of the baryons and produces shocks.
- Thermal processes, which control the cooling of gas and thus how it will flow into galaxies.
- Star formation, which in the context of galaxy formation models means how cool gas proceeds to fragment and form dense cloud cores that lead to individual stars, and which may affect the subsequent evolution of the galaxy through feedback due to stellar winds and supernovae.
- Black hole formation and growth, which occurs due to gas inflow to the very centers of galaxies, and which may affect the subsequent evolution of the galaxy through AGN feedback.
- Nucleosynthesis, which leads to chemical enrichment of interstellar gas, which affects the thermal processes because of the importance of metal cooling.
- Radiative transfer, which can heat and cool gas, as well as affect the observed nature of the galaxies.

In galaxy formation theories, gas cooling, inflow, and feedback play critical roles, and as noted below can only be modeled through *subgrid* physics—i.e. not from first principles. What simulations can predict from first principles are the effects of gravity on nonlinear collapse, the cooling of gas on large scales, and the flows of gas. However, these first principles calculations are still limited by resolution, which may lead to qualitatively important errors even on large scales.

1.4. Methodology

Simulations of galaxy formation utilize N-body simulations, hydrodynamic simulations, and prescriptions for subgrid physics, or rely on approximations to those simulations known as it semi-analytic models.

As used in the cosmological literature, the term *N-body simulation* refers to a purely gravitational simulation of the collisionless Boltzmann equation, approximated with particle methods. These simulations nominally simulate pure cold dark matter models, with no baryons. Traditional N-body simulations do not solve the collisionless Boltzmann equation fully, but use particles to sample the density field and study the gravitational interactions of the particles. Because the number density of particles in the simulation is far below the number density of expected cold dark matter particles, if the gravitational interactions were solved exactly they would lead to unrealistically short two-body relaxation times in the simulations. Therefore, the gravitational force law needs to be *softened* at some length.

To follow the evolution of baryons, N-body simulations are not sufficient, because baryonic gas is a fluid, and also can radiate and cool. Furthermore, a fraction of the baryonic gas can form stars (which are dense enough that they behave like collisionless particles). Cosmological *hydrodynamic simulations* include these effects, but are correspondingly more expensive. Hydrodynamic simulations come in two basic types: particle- and mesh-based. *Smoothed particle hydrodynamics* methods (Springel 2010) treat fluids as a set of particles carrying information on the gas state. Mesh-based methods treat fluids as continuous fields. From the point of view of galaxy formation, both types of methods have developed sufficiently to agree with each other, and the challenges in simulations are not thought to be dominated by this methodological choice.

In both hydrodynamic and N-body simulations, there is a need for a large dynamic range, which drives us to consider high resolution. On the other hand, at late times much of the volume is nearly empty and does not require as detailed simulation. Therefore, simulators are driven to consider *adaptive resolution* techniques to focus computational resources (grid cells and particles) in the regions of greatest interest. Another technique is to run a large volume simulation and choose a subset of regions to rerun with higher resolution as a *zoom-in simulation*, with initial and boundary conditions set by the original simulation.

Cosmological-scale hydrodynamic simulations do not model the physics of the interstellar medium, star formation, and feedback directly. The very highest resolution simulations reach

about 100 pc in resolution, and therefore barely resolve gaseous galactic disks. Therefore typically the radiative cooling models used are those appropriate for diffuse, collisionally ionized gas; with this cooling curve dense gas is stable at 10^4 K, but the additional cooling processes on smaller-than-resolved scales that lead to molecular cloud and star formation are not included. Nor are the subsequent processes that lead to chemical evolution and feedback from stars.

The transition from dense, cold gas to stars, and subsequent feedback on the interstellar medium from stellar processes, is therefore is typically treated with *subgrid physics* modeling. Differences in the subgrid physics are the dominant source of disagreement between predictions for galaxy evolution from numerical simulations.

Subgrid physics needs to account for the multi-phase nature of the interstellar medium; this multi-phase nature means that it does not necessarily behave as a simple gasr In most cosmological simulations this is taken into account in terms of empirically constrained effective equations of state and rules for the production of molecule gas. To model star formation, typically simulators assume that gas above some threshold density and that is converging ($\nabla \cdot \vec{v} < 0$), will form stars at a rate set by the free-fall time:

$$\dot{\rho}_* = \frac{\epsilon_* \rho_{\rm gas}}{t_{\rm ff}} \tag{1}$$

Since $t_{\rm ff} \propto \rho^{-1/2}$, this formulation has a Kennicutt-Schmidt-type relationship built into it. ϵ_* is observed in molecular clouds to be relative small (~ 0.01) probably due to turbulence well below the simulation resolution scales, in molecular clouds and generated by stellar processes.

The subgrid physics also must account for the feedback from stellar and AGN processes, both of which have been observed to cause outflows and heating under the right circumstances. The stellar processes of greatest effect should be supernovae, which should drive blast waves into the interstellar medium, which both imparts momentum and heats the gas. There is not a good theoretical understanding that predicts bottom-up how these effects combine to drive a wind, and there is not an agreed-upon way to implement an effective low resolution subgrid model which behaves as bottom-up models would predict. Generally the parameters are tuned with a wind velocity $v_{\rm wind}$ and a mass loading factor $\eta = \dot{M}_{\rm wind}/{\rm SFR}$ that scales with the galaxy velocity dispersion (i.e. depth of the potential well) as $\eta \propto \sigma^{-\alpha}$ with $\alpha \sim 1$–2. $\alpha \sim 1$ is appropriate for *momentum-driven winds*, in which the gas cools quickly relative to the dynamical time, and $\alpha \sim 2$ is appropriate for *energy-driven winds*, in which the outflow is driven by thermal heating that does not quickly radiate away.

Accreting supermassive black holes also can provide feedback. *Radiative mode* feedback (associated with radiatively efficient accretion onto the black hole) heats the gas up and ionizes it, and can drive winds through radiation pressure (for example, coupling to atomic lines, electrons, or dust). Again, it is unclear how this feedback should act exactly, or whether it is better approximated by a momentum-driven or energy-driven model. *Jet mode* or *radio mode* feedback is provided by the AGN jet, most often present without the radiatively efficient accretion. The kinetic energy in the jets are higher than the bolometric luminosity of the AGN. Although there are

hot bubbles observed in X-rays associated with some jets, which indicate enough heating to offset cooling, the exact way the energy is coupled remains unclear. To model all of this, simulations follow parametrized models for black hole accretion and growth, depending on the gas density and artificially seeding haloes with black holes.

All of the processes described above can also be followed with *semianalytic models* (or SAMs), which have the advantage of being faster and therefore allow more rapid testing of hypotheses and fitting of parameters to observables. Semianalytic models start typically with halo merger histories from N-body simulations (or more rarely these days, from excursion set modeling). The merger histories are then used to infer gas accretion and galaxy merger histories. Within that context, a similar set of rules for star formation, feedback, and quenching is applied.

Within any of these methods, nucleosynthesis along with stellar mass loss and supernovae lead to chemical evolution. The enrichment is modulated by pristine gas infall, and also chemically enriched outflows (enriched by the very supernovae that cause the outflow).

1.5. Features of Theoretical Predictions

The implications of the comparison of the theoretical models with observations is in flux, and much is not understood. However, there are a number of results which have proven robust over many years and are useful to understand, as they have shaped our understanding of the ingredients needed for galaxy formation

The cooling of gas in halos as they formed is predicted to be highly efficient, especially when metal cooling is accounted for, which greatly enhances the gas cooling rates at halo temperatures. If this cooling were not balanced by some process that prevented star formation, it would lead to far more star formation at all halo masses than observed. It is this *overcooling* problem, recognized even in the 1980s before the Cold Dark Matter was fully established, that motivates the extensive effort in modeling feedback.

In the simulations, the accretion mode of gas also is strongly affected by the details of the cooling. In particular, when the cooling time is short relative to the free-fall time, *cold mode accretion* occurs directly down to the disk of a galaxy; in simulations this tends to occur along streams fed by larger scale filametnts. When the cooling time is long relative to the free-fall time, a hot gas halo is formed that grows through *hot mode accretion*. Gas falling in forms shocks near the edge of the halo, and the system gradually cools. Generally larger mass halos ($> 10^{12} M_{\odot}$ experience hot mode accretion, and lower mass halos experience cold mode accretion.

In whatever manner gas cools, it leads to a decrease in its orbital energy. This leads the gas to fall to the center, but if angular momentum is conserved, it cannot fall all the way in. If we knew how far the gas could fall in (i.e. its angular momentum) we could predict the sizes of the disks that would form. We start by assuming the specific angular momentum of baryonic material is the same

as the dark matter, and it cannot be efficiently transported outwards. The net specific angular momentum $j_{\rm DM}$ of the dark matter in forming halos is produced by tidal torques on forming halos, and can be predicted from N-body simulations, and can be reexpressed as a fraction of the average specific angular momentum based on the virial quantities:

$$\lambda = \frac{j_{\rm DM}}{\sqrt{2} R_{\rm vir} v_{\rm vir}}. \tag{2}$$

Typically $\lambda \sim 0.03$–0.04 in simulations. One can easily calculate that for a flat rotation curve galaxies with an exponential disk of scale length R_d, the specific angular momentum is $j_* = 2 R_d v_{\rm flat}$. If we set $j_{\rm DM} = j_*$ then we can solve for R_d:

$$R_d = \frac{1}{\sqrt{2}} \lambda \left(\frac{v_{\rm vir}}{v_{\rm flat}} \right) R_{\rm vir} \tag{3}$$

Clearly for $R_{\rm vir} \sim 100$ kpc, R_d will be a few kpc, as we find for actual galaxies, indicating that angular momentum is roughly conserved.

The formation of elliptical galaxies and spheroids and explaining their scaling relations is more complex. The paradigm for a long time has been that major mergers of disk galaxies can form elliptical galaxies, which can grow by minor mergers with gas-rich or gas-poor dwarf galaxies. But simulations find it is hard to prevent the formation of a disk subsequently; this finding motivates the need to presume that gas is heated or removed prior to the mergers. In addition, it now appears possible that elliptical galaxies formed largely in situ, driven by violent disk instabilities that drove gas inward and formed a bulge; this would explain why ellipticals are present at high redshift ($z \sim 2$) even though the progenitor spirals that are required in the merger scenario are clearly very gas rich at those redshifts.

In these massive systems, AGN feedback is essential to explaining why their stellar-to-halo mass ratios are so low. In lower mass systems, the low ratios cannot be due to AGN feedback, but supernovae feedback may play a role. Dwarf galaxies appear to have had a nearly constant, very low efficiency star formation, but simulations with supernova feedback tend to have declining star formation with time, with much lower star formation rates today than observed.

Theoretical models also of course can predict the variation of properties with environment, the variation of metallicity with mass and within galaxies, and many other observable properties. This is a rich area of research with much remaining to be learned.

2. Key References

- *Physical Models of Galaxy Formation in a Cosmological Framework* (Somerville & Davé 2015)
- *The Connection Between Galaxies and Their Dark Matter Halos* (Wechsler & Tinker 2018)

3. Order-of-magnitude Exercises

1. For a typical Milky Way sized halo, how many particles must it be modeled with for the two-body relaxation time to be longer than the age of the universe.

4. Analytic Exercises

1. Angular momentum and size
2. Minor merger growth

5. Numerics and Data Exercises

1. Comparing halo and stellar mass function.
2. Dealing with a numerical simulation result

REFERENCES

Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51

Springel, V. 2010, ARA&A, 48, 391

Wechsler, R. H., & Tinker, J. L. 2018, ArXiv e-prints, arXiv:1804.03097