The Dynamics, Structure & History of Galaxies
ASP Conference Series, Vol. nnn, 2002
G. S. Da Costa & E. M. Sadler, eds

Formation of Galactic Disks

S. Michael Fall
Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

Abstract. We review progress in understanding the formation of galactic disks in the standard cosmogonic scenario involving gravitational clustering of baryons and dark matter and dissipative collapse of the baryons. This scenario accounts remarkably well for the observed properties of galactic disks if they have retained most of the specific angular momentum they acquired by tidal torques. Early simulations, which included cooling of the gas but not star formation and the associated feedback, indicated instead that most of the angular momentum of the baryons would be transferred to the dark matter. Recent simulations indicate that this angular-momentum problem can be solved partially, and in some cases entirely, by feedback and other effects.

1. Introduction

Two key ingredients in the formation of galactic disks are dissipation and rotation. Dissipation by radiative cooling causes the gas in a protogalaxy to collapse inward; rotation then halts the collapse in the directions perpendicular but not parallel to the overall angular-momentum vector, resulting in a thin centrifugally supported disk. Other processes likely to play some role in the formation and subsequent evolution of galactic disks include dynamical friction, internal torques, kinematic viscosity, mergers, and star formation and the associated heating and stirring of the gas (feedback). In this article, we consider the general cosmogonic scenario in which the extended halos of galaxies form hierarchically by the gravitational clustering of non-dissipative dark matter, and the luminous components of galaxies form by a combination of the gravitational clustering and dissipative collapse of baryons, as proposed by White & Rees (1978). The formation of galactic disks and the origin of their rotation in this scenario were first studied by Fall & Efstathiou (1980). Most of the currently popular models of galaxy formation, including all variants of the cold and warm dark matter (CDM and WDM) models, are specific versions of this general scenario.

2. Tidal Torques

The formation of objects by gravitational clustering automatically endows them with some angular momentum (Peebles 1969). This is because the (proto)objects initially have irregular shapes and hence non-zero quadrupole moments. They therefore exert tidal torques on each other, with random strengths and directions. The objects acquire most of their angular momentum in the translinear regime, when their density contrasts are appreciable ($\delta\rho/\rho \sim 1$), but before they reach their maximum sizes and begin to contract. The rotation induced by tidal torques is usually quantified in terms of the dimensionless spin parameter

$$\lambda \equiv J|E|^{1/2}G^{-1}M^{-5/2}, \tag{1}$$

where J, E, and M are the total angular momentum, energy, and mass of the object, and G is the gravitational constant. Rapidly rotating objects, such as disks, have $\lambda \sim 1$, whereas slowly rotating objects, such as spheroids, have $\lambda \ll 1$.

Cosmological N-body simulations have revealed that the distribution of spins induced by gravitational clustering alone (i.e., without dissipation) is approximately lognormal:

$$p(\lambda)d\lambda \propto \exp\left\{-\frac{1}{2}\left[\frac{\ln(\lambda/\lambda_m)}{\sigma(\ln\lambda)}\right]^2\right\}d\ln\lambda, \tag{2}$$

with

$$\lambda_m \approx 0.05 \qquad \text{and} \qquad \sigma(\ln\lambda) \approx 0.5. \tag{3}$$

Most objects rotate slowly although there is a wide range of spins. This distribution appears to be nearly universal in the sense that it has little or no dependence on the cosmological parameters, the initial spectrum of density perturbation, or the masses and densities of the objects (Barnes & Efstathiou 1987; Zurek, Quinn, & Salmon 1988; Warren et al. 1992; Cole & Lacey 1996). These results are very useful, even if they are not yet fully understood theoretically.

3. Analytical Collapse Model

We can relate the factor by which the baryons collapse in the radial direction before they reach centrifugal balance in a disk to the initial rotation of a protogalaxy as follows (Fall 1983). For simplicity, we consider a galaxy with a luminous disk (D) and a dark halo (H) but not a luminous spheroid. Furthermore, we assume that the baryons destined to become the disk receive the same tidal torques as the dark matter before much dissipation occurs and that during and after the collapse the total specific angular momentum of each component is conserved:

$$J_D/M_D = J_H/M_H. \tag{4}$$

In other words, we assume for now that angular momentum is not transferred between the disk and the halo, or if it is, that this is accompanied by the transfer of enough mass (outflow or inflow) that equation (4) is still satisfied.

We approximate the disk by an exponential model with a scale radius α^{-1} (sometimes denoted by R_d) and the halo by a singular isothermal sphere with a circular velocity v_c and a truncation radius r_t. Then, neglecting the self-gravity of the disk, we have

$$J_D/M_D = 2v_c\alpha^{-1}, \tag{5}$$

$$J_H/M_H = \sqrt{2}\lambda v_c r_t. \tag{6}$$

Equating these gives a very simple relation between the collapse factor αr_t of the baryons in the disk and the spin parameter λ of the halo:

$$\alpha r_t = \sqrt{2}/\lambda. \tag{7}$$

This implies $\alpha r_t \sim 30$ for $\lambda \sim 0.05$ and hence $r_t \sim 100$ kpc for a typical spiral galaxy like the Milky Way (with $\alpha^{-1} \approx 3$ kpc). The collapse factor is larger in halos with smaller spin parameters and smaller in halos with larger spin parameters. The collapse factor would also be larger if the baryons were to lose some of their specific angular momentum.

Equation (7) is an excellent approximation even when the halo has a finite core radius and when the self-gravity of the disk is included (Fall & Efstathiou 1980; see their Fig. 3). It is also a good approximation if the halo has an NFW profile (Navarro, Frenk, & White 1996, 1997) that is later deformed by the (adiabatic) contraction of the disk within it (Mo, Mao, & White 1998). The radius r_t of the halo in equation (7) should be the one within which the baryons have collapsed onto the disk. This may be compared and contrasted with the radius r_{200} within which the mean density of the protogalaxy is 200 times the critical (closure) density. The latter is near the transition between the virialized and infalling parts of the halo (Cole & Lacey 1996). In recent work in this field, it has been customary to identify r_t with r_{200}. This assumption, however, does not have much physical justification beyond the constraint $r_t \lesssim r_{200}$. It is likely that r_t depends on cooling, heating, and other non-gravitational processes at least as much as it depends on the gravitational clustering that determines r_{200}. Thus, we have no guarantee that r_t and r_{200} will be equal or even proportional to each other. (Anyone who doubts this should compare the cooling and virial radii in clusters of galaxies.)

4. Scaling Relations

The properties of the halos that form in a hierarchy by gravitational clustering obey some simple scaling relations. We can combine these with the relation between the collapse factor and the spin parameter to derive the corresponding scaling relations for the properties of galactic disks as follows (Fall 1983). We approximate the relation between the typical masses M_H of the halos (within r_t) and their circular velocities v_c by a power law, $M_H \propto v_c^k$. The index k depends in general on the initial spectrum of density perturbations, the cosmological parameters, the range of masses considered, and the relation between r_t and r_{200}. Simple arguments based on the different formation times of objects with different mean densities give $k = 12/(1-n)$ for $(\delta\rho/\rho)_{\rm rms} \propto M^{-(3+n)/6}$ (White & Rees 1978). This implies $k \approx 4$ for the effective index $n \approx -2$ of the CDM spectrum on galactic scales (Blumenthal et al. 1984). N-body simulations indicate instead $k \approx 3$ for the NFW halos in CDM cosmogonies (Navarro et al. 1997). This index is appropriate for $r_t = r_{200}$ and hence for perfectly synchronous formation. In view of the unknown initial spectrum (index n) and unknown relation between r_t and r_{200}, we regard k as a parameter with some theoretical uncertainty.

The scaling relation for the halos can be reexpressed in terms of the typical luminosities and circular velocities of the disks as

$$L_D \propto (f_D/\Upsilon_D) v_c^k, \tag{8}$$

where $f_D \equiv M_D/M_H$ is the ratio of masses of the disks and halos, and $\Upsilon_D \equiv M_D/L_D$ is the mass-to-light ratio in the disks. This may be compared with the observed Tully-Fisher (1977) relation, $L_D \propto v_c^l$, the index of which varies from $l \approx 3$ in the B band (0.44 μm) to $l \approx 4$ in the K band (2.2 μm) (Verheijen 2001). Most of this variation can be explained by a dependence of Υ_D on v_c, indicated by the observed correlation between the colors and the luminosities of galaxies. When this effect, including the mass of interstellar gas, is taken into account, the index of the baryonic Tully-Fisher relation is found to be $k = 3.5 \pm 0.4$ for $f_D = \text{const}$ (Bell & de Jong 2001).

The central surface brightness of an exponential disk is given by $I_0 = \alpha^2 L_D/2\pi$. Using equations (7) and (8) and the relation $v_c^2 = GM_H/r_t$, we can rewrite this in the form

$$I_0 \propto \lambda^{-2} v_c^{l-2k+4} \propto v_c^{l-2k+4}, \tag{9}$$

where all of the dependence on f_D and Υ_D is contained in the factor v_c^{l-k}. The second proportionality holds because the spin parameters of the halos are statistically independent of their other properties. For $k \approx 3.5$, we expect $I_0 \approx \text{const}$ in the B band ($l \approx 3$) and $I_0 \propto v_c$ in the K band ($l \approx 4$). The first of these is the same as the original Freeman (1970) relation in the B band. It would be interesting to search for the predicted correlation in the K band.

The scaling relations above indicate how the typical luminosities and surface brightnesses of galactic disks depend on their circular velocities. The dispersions about these relations are determined in part by the dispersion in the spin parameter. We expect the L_D-v_c correlation to have relatively little scatter because it is independent of λ [equation (8)], while we expect the I_0-v_c correlation to have a great deal of scatter because it includes the factor λ^{-2} [equation (9)]. These trends are qualitatively consistent with the observed Tully-Fisher and Freeman relations. The distribution of I_0 at fixed v_c can be derived from equation (2) for $p(\lambda)$ and equation (7) for αr_t (Dalcanton, Spergel, & Summers 1997; Mo et al. 1998; Weil, Eke, & Efstathiou 1998). These results can then be combined with the luminosity function of galaxies to derive the joint distributions of I_0 and α^{-1} (Dalcanton et al. 1997) and L_D and α^{-1} (de Jong & Lacey 2000). In all these studies, the observed distributions are reproduced quite well provided the coefficient in equation (7) is reasonably close to $\sqrt{2}$, i.e., provided the disks have retained most of the specific angular momentum they acquired by tidal torques.

5. Origins of Exponential Disks

The results above are based on the assumption that the *total* specific angular momenta in the disks are the same as those in their halos [equation (4)]. We may refer to this as the *weak* form of the assumption of angular-momentum conservation. In addition, it is sometimes supposed that the *distributions* of specific angular momentum in the disks are the same as those in their halos. This is usually expressed in terms of the fractions of mass with specific angular momentum below $h = Rv_\phi$ in the two components:

$$M_D(h)/M_D = M_H(h)/M_H. \tag{10}$$

We may refer to this as the *strong* form of the assumption of angular-momentum conservation. It is analogous to Mestel's (1963) hypothesis that $M_D(h)$ would be conserved in the collapse of a one-component protogalaxy (made before dark halos were discovered). Note that equation (10) implies equation (4), but the converse is not true in general. It is possible to change $M(h)/M$ in either component without affecting the corresponding J/M.

Several authors have pointed out that the distribution $M_D(h)/M_D$ for an exponential disk with a flat rotation curve is similar to that for a uniform sphere with a constant angular velocity (Gunn 1982; van der Kruit 1987; Dalcanton et al. 1997). This in turn resembles the distribution $M_H(h)/M_H$ caused by tidal torques (Barnes & Efstathiou 1987; Quinn & Zurek 1988), although there has been a recent tendency to emphasize the differences rather than the similarities (Firmani & Avila-Reese 2000; Bullock et al. 2001; van den Bosch 2001). Thus, the strong form of the assumption of angular-momentum conservation provides a possible explanation for the fact that the radial profiles of most galactic disks are approximately exponential. The main concern here is that several processes could alter $M_D(h)/M_D$ and hence the radial profiles of the disks over their lifetimes. These include the torques exerted by non-axisymmetric features in the disks (bars and spiral arms), the non-conservation of angular momentum as gas flows though shocks in spiral arms, and the exchange of angular momentum in collisions between clouds, not to mention galactic outflows, fountains, and mergers. It seems likely that these processes play some role in determining the present distributions of specific angular momentum in galactic disks.

Lin & Pringle (1987) showed that the radial distributions of the stars in galactic disks would evolve toward exponential-like profiles if two conditions were met. (1) The net effect of the processes that redistribute angular momentum in the interstellar gas can be described by an effective viscosity $\nu_{\rm eff}$. (2) The associated timescale, $t_\nu = R^2/\nu_{\rm eff}$, is about the same as the timescale for star formation, $t_* = \Sigma_g/\dot{\Sigma}_s$. This is plausible because both the redistribution of angular momentum and the formation of stars may be regulated by instabilities in the disks, including bars and spiral arms, although we lack a full understanding of just how or even whether this would actually happen. The Lin-Pringle model has been explored further, including its chemical evolution, by several authors (Clarke 1989; Yoshii & Sommer-Larsen 1989; Sommer-Larsen & Yoshii 1989, 1990). This model raises the possibility that the ubiquitous exponential profile is the result of viscous processes operating in the disks after they formed rather than tidal torques acting on the galaxies while they formed. More likely, both are involved: the exponential profile is first established by external torques and is then reinforced by internal viscosity.

6. The Angular-Momentum Problem

About a decade ago, it became possible to simulate the hierarchical formation of galaxies with both dark matter and baryons by a combination of N-body and hydrodynamical models. The early simulations included the radiative cooling of the gas but not the formation of stars and the associated feedback (heating and stirring of the gas). The results were very different from the analytical models described above (Navarro & Benz 1991; Navarro & White 1994). The baryonic objects that formed in these simulations were ellipsoidal in shape and an order of magnitude smaller than galactic disks of the same scaled mass. Thus, they resembled the spheroids more than the disks of real galaxies. The reason for this is that the gas cooled quickly and collapsed into dense subunits within the halos. A combination of dynamical friction and gravitational torques within the halos then transferred most of the orbital angular momentum of the baryons to the dark matter, causing the subunits to sink toward the centers of the protogalaxies, where they then merged. (For a prescient discussion of how these processes might determine the different properties of galactic spheroids and disks, see Zurek et al. 1988.) The discrepancy between the baryonic objects produced in such simulations and real galactic disks has become known as the angular-momentum problem of galaxy formation. It is so severe that it is sometimes referred to as a crisis or catastrophe.

Steinmetz & Navarro (1999) have explored this and related issues with a series of high-resolution simulations. They have tried to alleviate the angular-momentum problem by including some feedback in their simulations but find this makes little difference to the outcome. This conclusion, however, depends on their particular prescription for feedback. Steinmetz & Navarro assume that the gas is heated only in the immediate vicinity of ongoing star formation. Then, since the gas is very dense at these locations, the energy input is quickly radiated away, before it can influence the motions or the thermodynamic state of the gas at other locations. Steinmetz & Navarro (1999) have also found that the zero-point of the relation between luminosity and circular velocity in their simulations differs significantly from that of the observed Tully-Fisher relation, especially in the Einstein-de Sitter cosmological model ($\Omega_M = 1$, $\Omega_\Lambda = 0$) with CDM. More recently, Eke, Navarro, & Steinmetz (2001) have shown that this problem is reduced substantially or solved completely in the concordance cosmological model ($\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$) with CDM or WDM.

7. Possible Solutions

It is now widely believed that the solution of the angular-momentum problem must be found in some extreme form of feedback that prevents the baryons from collapsing until after the violent relaxation in their halos is complete. The baryons would then collapse within relatively smooth halos, with little quadrupole coupling, and hence would retain most of their specific angular momentum, as was originally, although perhaps naively, envisaged (Fall & Efstathiou 1980). That this idea goes a long way toward solving the angular-momentum problem has been demonstrated explicitly in a number of recent simulations (Weil et al. 1998; Sommer-Larsen, Gelato, & Vedel 1999; Eke, Efstathiou, & Wright 2000; Thacker & Couchman 2001). In the Weil et al. and Eke et al. simulations, the gas evolves adiabatically until the redshift $z = 1$ and is then allowed to cool radiatively and collapse. In the Sommer-Larsen et al. and Thacker & Couchman simulations, the feedback is driven by local star formation but is extensive in both time and position (in contrast to the Steinmetz & Navarro prescription). The common result of these studies is that the disks retain much more of their specific angular momentum.

Several other effects also help. The transfer of angular momentum is less severe in simulations with warm dark matter than in simulations with cold dark matter (Sommer-Larsen & Dolgov 2001). The reason for this is that the halos in WDM simulations have much less substructure and hence less dynamical friction and internal torques than the halos in CDM simulations. The transfer of angular momentum is also reduced in simulations with a cosmological constant (Eke et al. 2000). This happens because protogalaxies acquire their angular momenta earlier, before the gas cools, and then suffer fewer late mergers in the Λ-dominated simulations. It is also likely that some of the baryons in the inner parts of the protogalaxies, where the specific angular momentum is lowest, either formed the luminous spheroids or were expelled entirely in galactic winds (Eke et al. 2000).

The specific angular momenta of real galactic disks can be explained by a combination of extreme feedback and one or more of the other effects mentioned above. Does this mean the angular-momentum problem has been solved? Yes and no. Yes, because we now know that some mechanisms are capable of reconciling the simulations with observations. No, because we do not yet know whether these mechanisms operate in real galaxies. The issue of feedback is especially vexing because it probably depends on features of the interstellar medium on the scales of individual star-forming clouds, of order parsecs. (For a promising model of feedback, see Efstathiou 2000.) The energy released by young stars—in ionizing radiation, stellar winds, and supernova ejecta—is more than sufficient to alter radically the collapse of the baryons in protogalaxies. The key issues are where, when, and how this energy is deposited, in particular, whether it is absorbed within the clouds themselves or is distributed more widely within the protogalaxies. This, in turn, depends on the sizes and masses of the clouds, the locations of the stars within them, whether the clouds are porous, and so forth.

8. Perspective

The goal of the research reviewed here has been to find a physically consistent model that accounts for the observed properties of galactic disks. Two decades ago, it seemed as if we might be close to achieving this goal. At that time, the observed properties of galactic disks were explained remarkably well by an idealized model in which the baryons and dark matter in protogalaxies acquired the same specific angular momenta by tidal torques and then conserved them during and after the collapse of the baryons. A decade later we recognized that, if the gas in protogalaxies were to cool rapidly, most of the angular momentum of the baryons would be transferred to the dark matter. We now know that this angular-momentum problem can be solved by stellar feedback and other effects in principle, but we do not yet know how effective these processes are in practice. In another decade, with some luck, NGST will be in operation, and we should be able to observe the formation of galactic disks directly. It will be interesting to see how much theoretical and numerical progress has been made by then. The simulations will undoubtly improve, but it will be a formidable challenge to include all the relevant physics on all the relevant scales.

Appreciation. This article is dedicated to Ken Freeman at the celebration of his sixtieth birthday on Dunk Island. I am grateful to Ken for his friendship and collaboration over many years.

References

Barnes, J., & Efstathiou, G. 1987. ApJ, 319, 575

Bell, E. F., & de Jong, R. S. 2001, ApJ, 550, 212

Blumenthal, G. R., Faber, S. M., Primack, J. R., & Rees, M. J. 1984, Nature, 311, 517

Bullock, J. S., Dekel, A., Kolatt, T. S., Kravtsov, A. V., Klypin, A. A., Porciani, C., & Primack, J. R. 2001, ApJ, 555, 240

Clarke, C. J. 1989, MNRAS, 238, 283

Cole, S., & Lacey, C. 1996, MNRAS, 281, 716

Dalcanton, J. J., Spergel, D. N., & Summers, F. J. 1997, ApJ, 482, 659

de Jong, R. S., & Lacey, C. 2000, ApJ, 545, 781

Efstathiou, G. 2000, MNRAS, 317, 697

Eke, V., Efstathiou, G., & Wright, L. 2000, MNRAS, 315, L18

Eke, V. R., Navarro, J. F., & Steinmetz, M. 2001, ApJ, 554, 114

Fall, S. M. 1983, in IAU Symp. 100, Internal Kinematics and Dynamics of Galaxies, ed. E. Athanassoula (Dordrecht: Reidel), 391

Fall, S. M., & Efstathiou, G. 1980, MNRAS, 193, 189

Firmani, C., & Avila-Reese, V. 2000, MNRAS, 315, 457

Freeman, K. C. 1970, ApJ, 160, 811

Gunn, J. E. 1982, in Astrophysical Cosmology, ed. H. A. Brück, G. V. Coyne, & M. S. Longair (Vatican: Pontificia Academia Scientiarum), 233

Lin, D. N. C., & Pringle, J. E. 1987, ApJ, 320, L87

Mestel, L. 1963, MNRAS, 126, 553

Mo, H. J., Mao, S., & White, S. D. M. 1998, MNRAS, 295, 319

Navarro, J. F., & Benz, W. 1991, ApJ, 380, 320

Navarro, J. F., & White, S. D. M. 1994, MNRAS, 267, 401

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493

Peebles, P. J. E. 1969, ApJ, 155, 393

Quinn, P. J., & Zurek, W. H. 1988, ApJ, 331, 1

Sommer-Larsen, J., & Dolgov, A. 2001, ApJ, 551, 608

Sommer-Larsen, J., & Yoshii, Y. 1989, MNRAS, 238, 133

Sommer-Larsen, J., & Yoshii, Y. 1990, MNRAS, 243, 468

Sommer-Larsen, J., Gelato, S., & Vedel, H. 1999, ApJ, 519, 501

Steinmetz, M., & Navarro, J. F. 1999, ApJ, 513, 555

Thacker, R. J., & Couchman, H. M. P. 2001, ApJ, 555, L17

Tully, R. B., & Fisher, J. R. 1977, A&A, 54, 661

van den Bosch, F. C. 2001, MNRAS, 327, 1334
van der Kruit, P. C. 1987, A&A, 173, 59
Verheijen, M. A. W. 2001, ApJ, 563, 694
Warren, M. S., Quinn, P. J., Salmon, J. K., & Zurek, W. H. 1992, ApJ, 399, 405
Weil, M. L., Eke, V. R., & Efstathiou, G. 1998, MNRAS, 300, 773
White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341
Yoshii, Y., & Sommer-Larsen, J. 1989, MNRAS, 236, 779
Zurek, W. H., Quinn, P. J., & Salmon, J. K. 1988, ApJ, 330, 519

Discussion

Quinn: With respect to the angular-momentum catastrophe, I think we all knew the simulations had problems since the cooling mechanisms (prescriptions) were not well defined physically and the heating was totally absent. Secondly, the loss of J/M by a factor ~ 10 is a great success in making ellipticals and spheroids, as pointed out by Zurek, Quinn, & Salmon (1988).
Fall: I tend to agree with both of your points. Thanks for reminding me of your paper with Zurek and Salmon. My impression is that the simulations without stellar feedback produce mainly spheroid-like objects, whereas those with extreme feedback produce mainly disk-like objects. Eventually, we should aim to produce both types of objects in the observed proportions in the same simulations. It would be nice to find a simple mechanism to do this.
Silk: The feedback models in which the gas is kept hot until $z \sim 1$ are going to have a problem in accounting for the observations of relatively massive disks at $z \sim 1$ and also for the apparent lack of evolution in the Tully-Fisher relation back to this redshift.
Fall: The kinds of observations you mention are potentially valuable constraints on, or input to, the models. However, my impression is that such comparisons are still very tentative. The samples of galaxies at high redshifts have not been selected in the same way as the nearby samples, and the photometric evolution of the disks remains quite uncertain.
Illingworth: Physically, why is it that Λ-dominated cosmologies help with the angular-momentum problem?
Fall: In Λ-dominated models, protogalaxies acquire their angular momenta earlier, before the gas cools, and are then disrupted by fewer late mergers than in matter-dominated models (see Eke et al. 2000).
Bosma: Are the angular-momentum and cuspy-halo problems related?
Fall: One might think so. Both problems seem to be aggravated by substructure within galactic halos. And for this reason, they should both be alleviated in models with warm dark matter. Recent simulations indicate, however, that WDM has more impact on the angular-momentum problem than it does on the cuspy-halo problem.