THE FORMATION OF GALAXIES

James E. Gunn
Princeton University Observatory, Princeton, NJ08544, USA.

Abstract: The presently fashionable ideas for galaxy formation are reviewed briefly, and it is concluded that the standard isothermal heirarchy fits the available data best. A simple infall picture is presented which explains many of the observed properties of disk galaxies.

Galaxy formation is an extremely active field of research at the present time, but if anything has grown more confused in the past few years. This confusion is largely due to the impetus lent by new data on the distribution of galaxies in space and the growth of grand unification theories (GUTs) in particle physics which, at least in their most straightforward applications, put severe constraints on the allowed forms of perturbations which might form galaxies. There are still basically two schools of thought on the subject, though the data look more and more as if some synthesis will ultimately describe Nature better than either.

It is generally agreed that galaxies form from density (and perhaps velocity) perturbations generated in a still-unknown way in the very early universe, and in particular those perturbations which survive until the epoch of the initial combination of the primeval plasma (decoupling) at a temperature of about 4000 K or a redshift of about 1500. The perturbation amplitudes required to form galaxies are of the order of a few percent at that epoch, and are large for low-density universes and for scenarios in which galaxies form very early. The disagreement centers around the nature of the perturbations in the earlier universe...are they adiabatic (i.e. accompanied by no change in the specific entropy, in which case $\Delta\rho = 3\Delta T$) or are they isothermal ($\Delta T = 0$)? The crucial difference for galaxy formation is that, as was first shown by Silk (13), adiabatic perturbations are damped by radiative viscosity for masses below 10^{13} to 10^{15} solar masses. If the universe is dominated by neutrinos of mass about 30 eV, as now seems quite possible, an analagous phenomenon occurs -- here the neutrinos simply run away from the density maxima because they have large random velocities, and the lower mass limit on surviving perturbations is set by the comoving distance they can travel once they uncouple, which in proportional to the inverse of their mass. For neutrino

E. Athanassoula (ed.), Internal Kinematics and Dynamics of Galaxies, 379–386.

https://doi.org/10.1017/S0074180900033325 Published online by Cambridge University Press

masses of interest, the cutoff masses are of the same order as the Silk mass. Since the neutrinos are generally believed once to have been in thermal equilibrium in the Universe (at times $\lesssim$1 sec, temperatures $\gtrsim 10^{10}$ Kelvin), any perturbations in the neutrino density will of necessity be adiabatic. Since GUTs are able to account for the synthesis of baryons by thermal processes as well, it would seem that the most likely sort of perturbations would be the adiabatic ones (see the Silk and Weinberg papers in (1)).

Both adiabatic pictures (the baryons-only one and the baryons-plus-neutrinos one) have negligible perturbation amplitudes on the scale of galaxies at decoupling, and so in these pictures clusters (or superclusters) form first, and galaxies form later by fragmentation processes. Gravitation plays an important role in these processes, to be sure, but the formation is NOT from collapse of perturbations present at decoupling. An excellent review is contained in ref (18).

The isothermal perturbations trade the mystery of their genesis for the simplicity of their subsequent behaviour - they remain essentially frozen in during the early evolution of the Universe and emerge from decoupling unscathed, and with negligible attendant velocity disturbances. It is possible, as shown by the work of Peebles and coworkers (see, for example, (11)) to account for the development of structure on all scales in the Universe today from a simple heirarchy of isothermal perturbations. The recent observations of Davis et al. (2) of the distribution of nearby galaxies and those of Kirschner et al. (8) pointing to the existence of very large voids in the galaxy distribution seem to indicate the existence of large amplitude disturbances on very large scales, as one would expect from the adiabatic picture, though on smaller scales ($\lesssim$3 Mpc, say) the heirarchical (isothermal) picture works quite well.

What tests are there? It would seem that the crucial question is that of the relative formation times of galaxies and clusters, and in my opinion essentially all the data on this question suggests that galaxies form very much earlier than clusters. The data take several forms, but among the most persuasive are the remarkable uniformity of the colors and spectra of ellipticals, both in the "field" and in rich clusters, synthesis models for which indicate great (and uniform) ages (see, for example, (7)); in the same vein is the data on very distant ellipticals seen at redshifts near unity, the spectra of which are consistent with ratios of ages then to now the same as the ratio of the ages of the Universe then to now. (This is a rather more sensitive test than the absolute age alone, since many uncertainties disappear in the ratio comparison (Gunn, in (1)). To me the most telling piece of evidence comes from the Galaxy and the Local Group. The Galactic globular cluster system is very old, and must correspond to a formation time for the bulge of the Galaxy at a redshift of at least about two, when the mean density of the Universe was thirty times greater than now. The density in the collapsing perturbation must have been at least three to five times (depending on the geometry) denser than the mean at turn-around, and that grows by at least a like factor during collapse. The

https://doi.org/10.1017/S0074180900033325 Published online by Cambridge University Press

density contrast then grows at least as fast as $(1+z)^{-2/3}$. Thus the cluster/supercluster which collapsed to fragment and form the Galaxy in the adiabatic picture must be a structure of some 10^{13} to 10^{15} solar masses with a density some thirty times the mean, and there exists no such candidate structure. The local supercluster is the only aggregate in which the Galaxy is embedded which is massive enough, but its mean density interior to the Galaxy is only about two and a half times that of the surrounding, and it can at most be barely turning around around NOW. This situation is further discussed by Dekel in this volume.

Thus it would appear that individual galaxies and small groups are very unlikely to have had their origins in adiabatic pancakes. There are other scenarios (see, for example, Ostriker, in (1), which still require isothermal seeds) but the most straightforward is to fall back on the isothermal heirarchy. (It must in fairness be remarked that if one wishes the large-scale structure to be adiabatic and the small-scale to be isothermal, a remarkable coincidence in initial conditions is required in order that the amplitudes are comparable at decoupling - but there is no reason why the structure cannot ALL be isothermal with a suitable power spectrum.) We must only await a plausible mechanism for generating the isothermal perturbations.

Let us review briefly the growth of isothermal perturbations after z=1500. A perturbation (which we shall assume roughly spherical for simplicity) is characterized by a dimensionless amplitude $\delta(r)$, which is the ratio of the mean density interior to the critical density at that epoch. For reasonable ($\geqslant 0.1$) values of Ω, the mean density differs from the critical density by an amount which is small compared to the perturbations of interest for the formation of galaxies. The collapse time t_c for the material interior to r is simply related to $\delta(r)$, as is the ratio of the maximum expansion radius to the initial radius:

$$t_c = \frac{\pi}{H_i}\,\delta^{-3/2} \qquad\qquad R_m = R_i\,\delta^{-1}$$

If the RMS density perturbation is a power law in the contained mass, as is often assumed, one can make some simple predictions for the resulting structures. Let us assume that

$$\delta_{RMS} \;\alpha\; M^{-\frac{1}{2}-\frac{n}{6}}$$

(Here n=0 is white noise, and negative n corresponds to more power at large scales). If the further assumption is made that $\delta(r)$ is a stationary (in space) stochastic process with random phase and with the above spectrum to scales much smaller than those of interest, it follows that the distribution of densities is approximately normal. On thing to notice immediately is that galaxies, clusters, etc. form around peaks in the density, and it a tautological property of peaks that the surrounding mean density is lower than the peak value. Since the collapse time varies as the inverse three-halves power of the amplitude, it follows that late infall of material is a general property of this picture.

https://doi.org/10.1017/S0074180900033325 Published online by Cambridge University Press

If there is no dissipation in the subsequent evolution through maximum expansion and collapse, the potential energy at maximum expansion is converted to potential and kinetic in the virial ratio in the final stationary configuration, (corresponding to a factor of two decrease of gravitational radius) and the collapse time is related by a constant ratio to the final dynamical time. Thus the maximum radius and collapse time are simple functions of the mass and velocity dispersion in the final structure, and since the RMS perturbation amplitude is related to the mass by the heirarchy, one can predict the mean radii and velocity dispersions as functions of mass (see, for example, (11)).

Now it is certainly NOT true that galaxies form without dissipation; that is an obvious enough statement for spirals, but equally good arguments can be made for ellipticals based on density contrast with their surroundings (see, e.g. (17)). It IS reasonable to assume that the dark halos of galaxies form dissipationlessly. It can be argued (Gunn, in (1)) that one can relate simply the rotational velocity V_c of spirals and the velocity dispersion σ_* of the stars in ellipticals to the velocity dispersion σ_H in their halos; $\sigma_H = \sqrt{3/2}\sigma_* = \sqrt{1/2}V_c$. For galaxies with both disks and dynamically important bulges, the last half of this relation can be checked empirically, and the fact that those same galaxies have flat rotation curves demands the satisfaction of the whole relation. A long series of arguments ((6), (15); Faber, Gunn in (1)) suggests that the ratio of dark "halo" stuff to ordinary matter is very nearly constant on (galaxy+halo) scales and larger, and that this ratio is about fifteen to one.

Assuming this to be true, and using the semi-empirical mass-to-light ratios adopted by Faber and Gallagher (3), one can now plot a MASS-(either halo mass or "visible" mass) HALO VELOCITY DISPERSION diagram, in which one can draw, for example, lines of constant collapse time and the mean line for any heirarchy. The results of doing this for about fifty galaxies with well-determined dynamical properties is shown schematically in Figure 1, where also shown is the Coma Cluster (c), the mean line for an $n=-1$ heirarchy which has been placed correctly with respect to Coma at that mass scale, and two regions related to the cooling of material in the perturbation - the one vertical line to the right of which atomic cooling processes cannot cool the object on the collapse timescale (12), (17), and the $t_c = 3\times10^8$ yr line, above which any object can cool by Compton cooling against the microwave (at those epochs infrared) background in a time shorter than the collapse time. Galaxies should be found only above and to the left of the region in which structures cannot cool, and that is indeed where they obligingly lie. The ellipticals (black dots; spirals are open symbols) are systematically high, and indeed lie almost exclusively in the Compton-cooling dominated regime, many of them in the region where they cannot have cooled by atomic processes (note that MERGERS move points to the right and horizontally in the diagram, so these massive ellipticals cannot have been formed by merging lower-mass systems of any sort we see today).

This diagram suggests a simple picture for the formation of galaxies, which is discussed somewhat more fully in my paper in (1).

https://doi.org/10.1017/S0074180900033325 Published online by Cambridge University Press

First, ellipticals and the bulges of spirals are formed in the Compton-cooling regime by the densest parts of perturbations, which for the most tightly-bound ellipticals involves essentially all the mass. The lower-density outer parts both cool and collapse more slowly. The perturbation is given angular momentum by the action of tidal torques throughout its history, and acqires in the mean a dimensionless angular momentum $\lambda = J|E|^{\frac{1}{2}}M^{-\frac{5}{2}}G^{-1}$ of about 0.07. Fall and Efstathiou (4) have shown that that value is roughly correct for disk galaxies when account is taken of dissipation. If the perturbation is given roughly solid-body rotation by the torques (as one might expect from simple but possibly naive dimensional arguments), then a remarkable result follows. The angular momentum distribution of such a system, when translated into a disk density distribution with a flat rotation curve, results in a distribution of surface density which closely approximates an exponential over some four scale lengths. Furthermore, as successive shells fall in, the disk grows self-similarly with almost constant central surface density and always has the nearly exponential form. It is a significant fact that the bulge population of the Galaxy has a lower specific angular momentum than the disk (Gunn, (1)) which argues strongly against the hypothesis that the bulge and disk formed from the same stuff, separated only by star formation rate - it would appear that the bulge is a separate dynamical subsystem. In this picture the bulge formed early and the disk later, with the latter still forming.

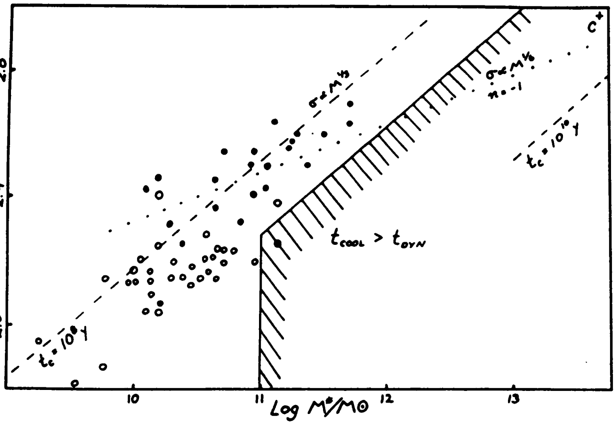


Figure 1. Halo velocity dispersion versus visible mass.

The picture predicts a definite evolutionary history for the disks of galaxies. If one looks at a fixed value of specific angular momentum (which does not correspond to fixed radius because the mass interior to the point in question grows slowly with time), the disk is at first of very low surface density and is vertically non-self-gravitating. The temperature of the gas will be about 15000 K and it will be mostly neutral. The scale height will be $H=\sigma_D r/V_c$, where σ_D is the equivalent one-dimensional velocity dispersion, about 10 km/s. As more material falls in, the disk becomes self-gravitating, and almost immediately becomes Jeans unstable (5,14). The only way to restore stability as more mass is added is for the gas to heat, but it cannot because of the cooling barrier. The mass motions driven become supersonic and rapid star form-

https://doi.org/10.1017/S0074180900033325 Published online by Cambridge University Press

ation occurs, after which the instability can heat the stars. Thus as the disk grows it always remains marginally unstable, at a fixed appropriate value of Q.

One can predict from this picture the velocity dispersion and scale height as a function of radius, and the latter is shown for the present Galaxy, with the local surface density taken to be 70 $M_{\odot}/pc^2$ and a rotational velocity of 220 km/s, in Figure 2. Note the remarkable constancy of the scale height, observed, of course, in external systems (9). The rapid ballooning of the disk occurs where the disk becomes non-self gravitating, and the neutral hydrogen in the disk in the Galaxy exhibits this behavior, as it may in other systems (Sancisi, this volume). This phenomenon is probably the best evidence for the dark halo material being roughly spherically distributed (6).

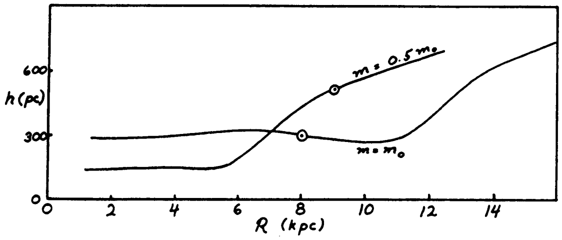


Figure 2. Predicted scale height for the present Galaxy and for an epoch when the disk was half as massive as now.

An added bonus of infall is that it explains the existence of spiral structure in a straightforward way. One would expect the fact that the disk is kept at the threshold of Jeans instability to keep it unstable to spiral modes as well, and I suggested that that would be the case in (1). Recently Carlberg and Sellwood have performed appropriate numerical experiments (described elsewhere in this volume) and the results are spectacularly positive. In addition, the spiral instability pumps the velocity dispersions of old stars, and the age-kinematics relations can probably be explained also, as they describe. Infall has been invoked for a long time, of course, to explain the age-metallicity relation (see, for example, (16)) and it appears to be as useful for dynamics.

Thus a large number of the properties of galaxies emerge in a natural way from the picture; on a rather more speculative level, the Hubble sequence finds a ready explanation. Galaxies which today have vigorous star formation are those for which the infall rates are high now, and those systems, on average, are ones which have long collapse times (systems with short collapse times have high infall rates at early times, but run out of material later). Thus the strong tendency for late-type systems to fall at the long-collapse-time end of the distribution in Figure 1 is explained. The correlation of bulge-to-disk ratio with Hubble type (defined in this instance by color or star formation rate) follows immediately - the bulge is a short-collapse-time subsystem, and perturbations which have long collapse times should, on average, have smaller such subsystems (if any at all) than denser ones. This picture is very schematic, but seems quite promising.

https://doi.org/10.1017/S0074180900033325 Published online by Cambridge University Press

This research was supported by the National Science Foundation of the U.S.

REFERENCES

1. Bruck, H.A., Coyne, G.V., Longair, M.S. 1981, Astrophysical Cosmology, Pontifica Academia Scientiarum, Citta del Vaticano.
2. Davis, M., Tonry, J., Huchra, J., and Latham, D. 1982, Ap. J. 253, 423.
3. Faber, S.M., and Gallagher, J.S. 1979, A.R.A.A. 17, 135.
4. Fall, S.M., and Efstathiou, G. 1980, M.N.R.A.S. 193, 189.
5. Goldreich, P., and Lynden-Bell, D. 1965, M.N.R.A.S. 130, 97.
6. Gunn, J.E. 1980, Phil. Trans. Roy. Soc. London A., 246, 313.
7. Gunn, J.E., Stryker, L.L., and Tinsley, B.M. 1981, Ap. J. 249, 48.
8. Kirschner, R.P., Oemler, A., Schechter, P.A., and Schectman, S. 1981, Ap. J. 243, L127.
9. van der Kruit, P., and Searle, L. 1981, Astron. Astrophys. 95, 105.
10. Peebles, P.J.E. 1969, ap. J. 155, 393.
11. Peebles, P.J.E., The Large-Scale Structure of the Universe, Princeton University Press.
12. Rees, M.J., and Ostriker, J.P. 1977, M.N.R.A.S. 179, 451.
13. Silk, J. 1967, Nature 215, 1155.
14. Toomre, A. 1964, Ap. J. 139, 1217.
15. Tremaine, S.D., and Gunn, J.E. 1979, Phys. Rev. Letters 42, 467.
16. Twarog, B.A. 1980, Ap. J. 242, 242.
17. White, S.D.M., and Rees, M.J. 1978, M.N.R.A.S. 183, 341.
18. Zel'dovich, Y., Einasto, J., and Shandarin, S. 1982, Nature, in press.

DISCUSSION

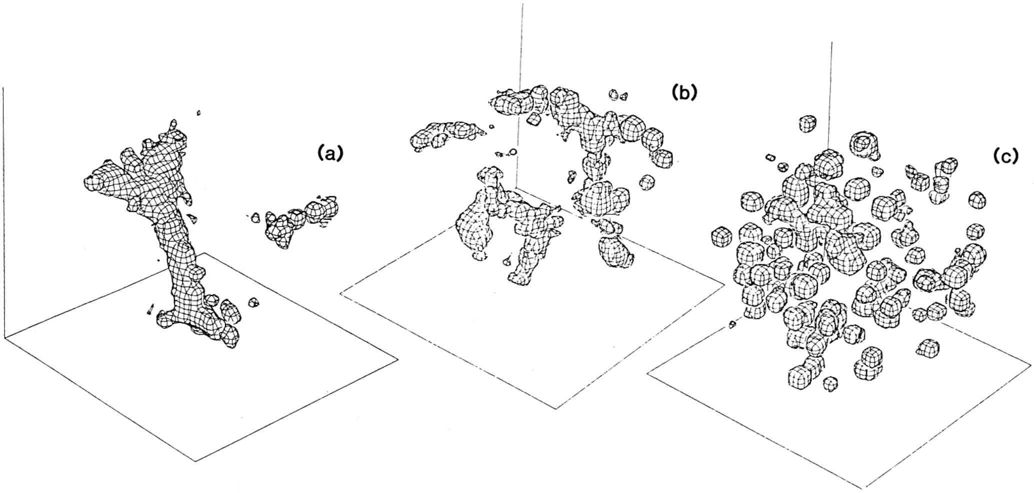
BLITZ : Wouldn't the star formation which takes place after the formation of the disk affect the evolutionary picture of the infall ? In our own galaxy, the massive star formation is a strong function of radius, and thus the effect on the interstellar medium via HII regions, stellar winds and supernovae is also likely to be a strong function of radius. Wouldn't you expect the dynamics of the ISM to seriously affect the simple infall picture ?

GUNN : All that the "simple infall picture" demands is that star formation more-or-less keep up with the infall, and the Jeans instability almost guarantees that this happen . Details will doubtless affect the quantitative chemical evolution, for example, and may or may not through transport effects, move the density distribution around in the disk. Qualitatively, however, it is difficult to see how the picture could be substantially changed unless the star formation activity were energetic enough to unbind the infalling matter.

WHITE : I would like to make a comment on what "pancakes" actually look like. Carlos Frenk, Marc Davis and I have run N-body simulations of the

https://doi.org/10.1017/S0074180900033325 Published online by Cambridge University Press

growth of clustering from random. Those initial conditions are with a short wavelength cut-off (the sort of initial condition advocated by Zel'dovich and his colleagues). We find that at times when the correlation properties of such models are similar to those of the observed galaxy distribution, they contain large coherent structures as much as ten times larger than the scale at which the two-point correlation is equal to unity. These "pancakes" are best displayed by equidensity contours of the smoothed particle distribution. They are primarily prolate objects and seem to link together at low density contrast in what may be a space-filling network. Such large-scale structure is not seen in model universes in which clustering has grown from "white noise" initial conditions even though they have a two-point correlation function quite similar to that of the pancake simulations.



Equidensity contours at 3 times the mean density in N-body simulations of clustering in an Einstein-de Sitter universe. (a) and (b) show the generic structure formed from random phase "adiabatic" initial conditions. In (a) the coherence length is such that only one "pancake" forms in the region simulated, while in (b) it is smaller,and several linked structures are apparent. (c) is the analogous structure in a simulation started from white noise "isothermal" initial conditions. The correlation function is unity on similar scales in (b) and (c). (From Frenk, White and Davis, in preparation).

GUNN : This is very interesting, but the crucial question is whether the universe looks like your pictures or not. It may be (or may not be) that it looks more complicated than n = -1 clustering models, but it is not clear either whether your models are closer than those to reality or not.

https://doi.org/10.1017/S0074180900033325 Published online by Cambridge University Press