Fred Hoyle: Contributions to the Theory of Galaxy Formation †,

By G. EFSTATHIOU

Institute of Astronomy, Madingley Road, Cambridge, CB3 OHA, UK.

I review two fundamental contributions that Fred Hoyle made to the theory of galaxy formation. Hoyle was the first to propose that protogalaxies acquired their angular momentum via tidal torques from neighbouring perturbations during a period of gravitational instability. To my knowledge, he was also the first to suggest that the masses of galaxies could be explained by the requirement that primordial gas clouds cool radiatively on a suitable timescale. Tidal torques and cooling arguments play a central role in the modern theory of galaxy formation. It is a measure of Hoyle’s breadth and inventiveness that he recognised the importance of these processes at such an early stage in the history of the subject.

1. Introduction

I will begin by quoting from an obituary of Sir Fred Hoyle written by Leon Mestel (2001):

‘Fred Hoyle was the astrophysicist par excellence, and much else. He wrote technical papers on an astonishingly wide range of astronomical topics, his most important work permanently widening our vistas and influencing strongly the direction of future research.’

The ‘much else’ referred to by Leon in this quote includes Hoyle’s contributions to the popularisation of science, prolific science writing, his creation and directorship of the Institute of Astronomy at Cambridge and his often visionary work for the U.K. Science Research Council.

At this meeting our focus has been on Hoyle’s astronomical research. Everyone would agree that Hoyle’s outstanding scientific achievements were in developing the theory of stellar evolution and in understanding the origin of the chemical elements. As appropriate, these topics have taken the central stage at this meeting. Hoyle’s contributions to galaxy formation, the subject of this article, are less well known. I have chosen this topic to illustrate the ‘astonishingly wide range’ of Hoyle’s research referred to by Leon Mestel. I will argue that Hoyle was the first to understand the masses of galaxies and why galaxies rotate. These are important results in their own right and now form an essential part of the modern theory of galaxy formation. Many astronomers would be proud to list these two results (which are not even mentioned in any of Hoyle’s obituaries) as their towering achievements!

Section 2 discusses the origin of galactic angular momentum and is based on Hoyle’s (1949) contribution to the quaintly named volume ‘Problems of Cosmical Aerodynamics’ edited by Burgers and van de Hulst. This volume summarises the proceedings of the equally quaintly named ‘Symposium on the Motion of Gaseous Masses of Cosmical Dimensions’ held in Paris in 1949. Section 3 discusses the origin of galactic masses. This is based on Hoyle’s well known paper ‘On the Fragmentation of Gas Clouds into Galaxies and Stars’ (Hoyle 1953). Curiously, most of the citations to this paper refer to the sections on star formation (hierarchical, opacity limited, fragmentation), yet this paper provides

2. Origin of Galactic Rotation

Consider the situation shown in Figure 1. Here we have an ellipsoid separated by distance \( r \) from an object of mass \( M \). Let us assume that the ellipsoid is a homogeneous oblate spheroid of density \( \rho \), mass \( M_G \), semi-major axis \( a \) and semi-minor axis \( b \). The magnitude of the torque on the spheroid is \( (r \gg a) \)

\[
\Gamma = \frac{3GM}{2r^3}(\sin 2\theta)\rho \int (x^2 - z^2) d^3x
\]

\[
= \frac{3GMQ}{4r^3} \sin 2\theta, \quad Q = \frac{2}{5}M_G(a^2 - b^2),
\]

(2.1)

where \( Q \) is the quadrupole moment of the spheroid.

How does Hoyle apply equation (2.1)? First, we need to estimate the timescale over which the torque acts. For this, Hoyle argues that the torque will act on a timescale comparable to the collapse timescale of a protogalaxy \( (\sim (3/4\pi G \rho)^{1/2}) \), where for \( \rho \) Hoyle adopts the value \( 10^{-27} g/cm^3 \). This was Hoyle’s estimate of the present mean matter density of the Universe and is high, presumably because estimates of the Hubble constant were so high at the time (see for example the discussion of the age discrepancy of evolving world models in Bondi and Gold 1948). We can re-interpret Hoyle’s estimate as the mean density of the Universe at the redshift at which the protogalaxy begins to form, \((1 + z_f) \sim 4(\Omega_m h^2)^{-1/3} \), which nowadays we would regard as a reasonable number.

Hoyle then assumes that the protogalaxy collapses until it is centrifugally supported, in which case the final angular velocity will be

\[
\Omega_f \approx \left( \frac{r^3}{3MGx} \right)^3 \left( \frac{4\pi G \rho}{3} \right)^{7/2}, \quad x = \frac{5Q}{4M_Ga^2} \sin 2\theta.
\]

(2.2)

To estimate the term \((r^3/M)\) in (2.2) Hoyle uses a clever argument. I quote directly from his paper

“We now reach the important step of interpreting the external gravitational field that produces the couple acting on the condensation. Instead of regarding this field as arising from a neighbouring galaxy, we notice that there are large scale
irregularities in the distribution of the internebular material. The existence of such irregularities probably exist also among the general field nebula, as is evidenced by the occurrence of peculiar velocities among these galaxies. Since the peculiar velocities average about 200 km s$^{-1}$ in the neighborhood of our own galaxy, this indicates $(GM/r)^{1/2} \approx 2 \times 10^7$ cm s$^{-1}$ in this neighborhood. In addition the observed over-all radii of the great nebular clusters are of order $10^6$ parsecs, which suggests a value of $r$ of order $3 \times 10^{24}$ cm.

Given how little was known about galaxy clustering at the time, this is a brilliant way of constraining the matter distribution and leads to Hoyle’s final estimate of

$$\Omega_f \approx \frac{10^{-16}}{x^3} \text{ s}^{-1},$$

which he argues is consistent with the characteristic angular velocities of the Milky Way and Andromeda ($\sim 10^{-15}$ s$^{-1}$) if the dimensionless number $x$ is of order 1/3 (a reasonable value).

What did the conference participants make of this argument? There is a wonderful discussion reported at the end of Hoyle’s contribution, which I have abstracted:

- **Heisenberg**: I had some difficulty understanding Dr. Hoyle’s argument... You start with an irregular cloud and then you ask why this irregular cloud gets an angular momentum. Now, I don’t think that any theory can give us an irregular cloud from the beginning without giving to it angular momentum at the same time.... If you have an irregular cloud, then it must have been produced by some kind of astronomical turbulence.

- **Hoyle**: I take it that Professor Heisenberg assumes turbulence to be present in the intergalactic medium arising from some unknown source. For my part I would regard the gravitational forces as the basic phenomena. The energy released by the gravitational contraction might possibly furnish turbulent motions and might lead to the appearance of eddies. I feel very doubtful, however, about the assumption that the intergalactic medium should already have turbulence from itself.

- **Heisenberg**: May I express my view in the following way. A cloud means turbulence and thus you should not start with a sphere without turbulence, since anything which is a cloud is turbulent by itself.

- **Batchelor**: [Who understood Kelvin’s circulation theorem] Why?

- **Heisenberg**: How can an irregular thing like a cloud have originated other than as a consequence of turbulent motion?

- **Hoyle**: A cloud can form in a more or less uniform medium through gravitational instability.

I can find nothing of significance on the tidal torque theory in the literature (apart from Sciama’s, 1955, application to galaxy formation in the Steady State theory) until Peebles’ important paper in 1969. In fact, the short abstract of Peebles’ paper almost echoes Hoyle’s last remark:

‘It is shown that the angular momentum of rotation of the Galaxy agrees in magnitude with the prediction of the gravitational instability picture for the rotation of the galaxies.’

In this paper, Peebles analysed the tidal torque theory using linear perturbation theory.

† By 1949 Heisenberg’s research interests had shifted to the study of turbulence. It is rumoured that Heisenberg said that he was looking forward to discussing quantum mechanics with other physicists in Heaven after he was dead. However, he thought that he would need a personal audience with God to understand turbulence.
and showed that it could account, roughly, for the angular momentum of the Milky Way. As part of my thesis work, I used N-body simulations to estimate the efficiency of the tidal torque mechanism (Efstathiou and Jones 1979). These simulations suggested that the dimensionless spin parameter $\lambda$ (roughly the ratio of rotational to kinetic energy) for a protogalaxy would have a low value of

$$\lambda = \frac{J}{|E|^{1/2}G^{-1}M^{-5/2}} \approx 0.05,$$

(2.4)
a result which was later borne out by much larger numerical simulations (Barnes and Efstathiou 1987, Zurek, Quinn and Salmon 1988). However, such a low value of $\lambda$ leads to problems with the theory as envisaged by Hoyle and Peebles, for if a self-gravitating gas cloud collapses from an initial radius $R_i$ to form a centrifugally supported exponential disc of scale-length $\alpha$, it must collapse by a huge factor of

$$\alpha R_i \approx 0.7/\lambda^2 \approx 300,$$

(2.5)
if $\lambda \approx 0.05$. The collapse time for a typical spiral disc would therefore be longer than the age of the Universe, hence the theory is untenable. A solution to this problem was
proposed by Efstathiou and Jones (1980) and worked out in detail by Fall and Efstathiou (1980). If spiral discs form from the collapse of baryonic material within a dissipationless dark halo (as envisaged in the two component theory of White and Rees 1978), and provided that the gas conserves its angular momentum and is only marginally self-gravitating when it reaches centrifugal equilibrium, then it only needs to collapse by a factor of

$$\alpha R_i \approx \sqrt{2}/\lambda \approx 30,$$

if $\lambda \approx 0.05$.

Does this lead to a viable theory of galaxy formation? Many cosmologists would probably say yes, but there are some thorny issues that have not yet been resolved. The first gas dynamics simulations of galaxy formation showed that much of the gas collapsed into dense sub-units at early times (as expected from the over-cooling problem identified by White and Rees 1978 and discussed in more detail in the next Section). These sub-units lose their orbital angular momentum by dynamical friction as they merge to form larger sub-units. The end result is a blob of gas or stars with specific angular momentum one or two orders of magnitude lower than observed in real disc galaxies† (see, e.g. Navarro and White 1994, Navarro and Steinmetz 1997). One possible solution to this problem is to invoke feedback from supernovae to prevent the gas from collapsing at early times. (There are differing views, for example, Governato et al. (2002) argue that the catastrophic angular momentum loss found in early simulations is caused by their limited numerical resolution.) Recently, numerical simulations that include simplified models of stellar feedback have shown that it is possible to make disc systems with specific angular momenta and scale-lengths comparable to those of $L^*$ galaxies (Weil, Eke and Efstathiou 1998, Sommer-Larsen, Götz and Portinari 2002, Abadi et al. 2002). An example is shown in Figure 2 (Wright, Efstathiou and Eke 2003). This simulation begins with scale-invariant adiabatic initial conditions in a $\Lambda$-dominated cold dark matter (CDM) cosmology ($\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$). The gas is (artificially) prevented from cooling before a redshift of unity, but once cooling sets in, the gas collapses to form a disc within the dark halo preserving a large fraction of its angular momentum. As one can see from the figure, the disc displays (transient) spiral arms at early times and eventually forms a strong bar. This richness of structure is a direct result of the angular momentum acquired by tidal torques during the early stages of gravitational instability, just as Hoyle predicted more than fifty years ago.

3. Explaining Galaxy Masses

Shortly after the discovery of the expansion of the Universe, it was realised that galaxies had a well defined upper luminosity (Hubble and Humason 1931). The distribution of galaxy luminosities (the galaxy luminosity function) was subsequently studied by many authors (see e.g. Binggeli, Sandage and Tammann 1988) and has recently been estimated with extremely high precision using the 2dF and SDSS galaxy redshift surveys (Blanton et al. 2001, Norberg et al. 2002). These studies show that the galaxy luminosity function is well described by a Schechter (1976) function and that 90% of the mean stellar luminosity density is contained in galaxies spanning the luminosity range $0.02 \lesssim (L/L^*) \lesssim 2.5$, where $L^* \approx 2.6 \times 10^{10}L_\odot$ in the $b$-band. Most of the stellar luminosity in the Universe is therefore confined to galaxies covering a range of about a hundred or so in luminosity. Furthermore, there is compelling evidence for a critical stellar mass of about $3 \times 10^{10}M_\odot$ (Kauffmann et al. 2002); galaxies of higher mass have roughly similar central surface

† This possibility was first pointed out to me by Peter Goldreich in 1981.
Within region C, bounded by the solid curve, a gas cloud of uniform density \( n \), temperature \( T \), and primordial composition can cool within a free-fall time. Clouds within region B can collapse quasi-statically within a Hubble time. Clouds in region A cannot cool within a Hubble time. (From Rees and Ostriker, 1977).

brightnesses, independent of mass or luminosity, while galaxies with lower stellar masses have lower surface brightnesses with \(<\mu> \propto M^{0.54}\).

Clearly, some physical process is required to explain the restricted range of galaxy luminosities and masses. In Hoyle's own words, one of the aims of his 1953 paper was to explain 'Why are the masses of galaxies mainly confined in the range \(3 \times 10^9 M_\odot\) to \(3 \times 10^{11} M_\odot\), with possibly a tendency to fall into two groups at the end of this range?'. Hoyle's explanation was based on simple gas dynamics – the requirement that a proto-cloud of hydrogen, at a few times the mean cosmic density, be able to cool radiatively on a timescale shorter than the timescale for gravitational collapse.

Hoyle's argument can be illustrated in a way more easily accessible to a modern readers using the famous cooling diagram, reproduced in Figure 3, from Rees and Ostriker (1977). A gas cloud with a particle density of \( n \sim 10^{-3} \text{cm}^{-3} \) (\( \rho \sim 10^{-27} \text{g/cm}^3 \), as considered by Hoyle) intersects the solid curve delineating region C in the diagram at two points, as indicated by the heavy vertical line in the Figure. Lines of constant Jean's mass (with slope 1/3 in this diagram) are shown by the dashed lines. Evidently a density of \( n \sim 10^{-3} \text{cm}^{-3} \) defines a lower Jean's mass of \( \sim 5 \times 10^8 M_\odot \) and an upper Jean's mass of \( \sim 10^{12} M_\odot \). The interpretation of this result is as follows. Region C delineates the region of the \( T-n \) plane in which gas clouds can cool radiatively within a free-fall time. Thus, clouds with mean density \( n \sim 10^{-3} \text{cm}^{-3} \) heated to their virial temperatures by gravitational collapse can cool and achieve the high overdensities (\( \gg 10^5 \)) of normal galaxies only if their masses lie within the range \( 5 \times 10^8 M_\odot \lesssim M \lesssim 10^{12} M_\odot \). These numbers differ from those in Hoyle's paper because he neglected cooling by line emission, and he did not phrase the argument in terms of the virial temperatures of protoclouds in quite the way discussed above. The key point, however, that cooling times delineate upper and lower bounds for galaxy masses, is contained in Hoyle's paper.
Figure 4. Semi-analytic models of the galaxy luminosity function from Cole et al. (2000) (solid lines) compared with observations of the $b$-band (left-hand panel) and $K$-band (right-hand panel) luminosity functions. (Figure from Baugh et al. 2002).

Does this cooling argument really explain galaxy masses? Since we no longer believe in the Steady State theory, we can ask what happens to clouds with densities much higher than $n \sim 10^{-3} \text{cm}^{-3}$. If the virial temperature of such a cloud exceeds $10^4 \text{K}$, then it can cool efficiently on a timescale much shorter than a free-fall timescale. The lower mass limit of Hoyle’s argument is ‘soft’ because it is sensitive to the redshift at which protogalaxies collapse.

In the cold dark matter model, this leads to a cooling catastrophe, as mentioned in the previous Section. In the absence of additional heating sources, the baryonic material would be expected to collapse efficiently in small non-linear systems with virial temperatures $\gtrsim 10^4 \text{K}$ at high redshifts (White and Rees 1978). According to the Press-Schechter (1974) theory, in a hierarchical model with a power-law spectrum of fluctuations $P(k) \propto k^n$, the mass function of dark haloes at low masses varies as

$$
\frac{dN(m)}{dm} \propto m^{-\left(9-n/6\right)}.
$$

(3.1)

For a CDM model, $n \approx -2$ on the scales relevant for galaxy formation and so the ‘cooling catastrophe’ would lead to a much steeper mass spectrum than inferred from the faint-end slope of the galaxy luminosity function ($dN(L)/dL \propto L^{-1}$). A photoionising background can prevent the gas from cooling in haloes with circular speeds less than about 30 km s$^{-1}$ (Efstathiou 1992, Benson et al. 2002) and this can help prevent the formation of dwarf galaxies. However, most authors are agreed that substantial feedback from supernovae driven winds is needed to reproduce the shape of the galaxy luminosity function in CDM models (see e.g. White and Frenk 1991, Cole et al. 2000). As with the angular momentum problem discussed in the previous section, it seems that stellar feedback is required to explain the observed properties of galaxies in a CDM universe.

This is illustrated in Figure 4 from Baugh et al. (2002). The left hand panel shows estimates of the $b$-band luminosity function from Blanton et al. (2001) and Norberg et al. (2002), while the right hand panel shows estimates of the $K$-band luminosity function from Cole et al. (2001). The solid lines in the figure show the predictions of the
Figure 5. Slices through a $\Lambda$-dominated dark matter simulation at various redshifts. Simple prescriptions from semi-analytic models have been applied to compute the properties of visible galaxies (masses, stellar ages, etc) shown by the open circles. (See Kauffmann et al. 1999).

semi-analytic models of Cole et al. (2000) which include substantial stellar feedback. These models provide an acceptable match to the observations, but the model predictions are sensitive to the details of the feedback model which are poorly know (see e.g. Efstathiou, 2000, for a discussion of stellar feedback) and to other cosmological parameters. For example, the models of Cole et al. (2000) assume a baryon density of $\Omega_b = 0.02$ which is lower than the value $\Omega_b \approx 0.045$ inferred from recent observations of the cosmic microwave background radiation (Spergel et al. 2003). A higher baryon density introduces problems with the bright end of the luminosity function as well as the faint end (Benson et al. 2003). A larger baryon density increases the efficiency of radiative cooling leading to an overproduction of massive galaxies. (This problem is closely related to the long-standing ‘cooling flow’ problem in cluster cores, see e.g. Fabian 2002). The resolution of this problem is still unclear and probably requires additional sources of feedback, perhaps associated with a massive central black hole (see e.g. Blandford 2001).

In summary, Hoyle correctly identified radiative cooling as an important process in
determining the baryonic masses of galaxies. However, fifty years after Hoyle’s paper, we still do not understand many of the physical processes that produce galaxies from an almost featureless spectrum of density fluctuations. Cooling clearly plays a role, but so does energy injection from stars and possibly AGN. The physics of these feedback mechanisms is poorly understood and yet is crucial for understanding galaxy formation. As Figure 5 shows, this physics is vital, for without it we cannot relate the dark matter distribution to the visible mass distribution at the present day and at higher redshifts.

4. Conclusions

Hoyle’s explanation of the rotation of galaxies and his recognition of the importance of radiative cooling in fixing the baryonic masses of galaxies are not usually listed amongst his most important contributions to Astronomy. Yet these processes are central to modern theories of galaxy formation. It is a measure of Hoyle’s breadth and inventiveness that he recognised the importance of these processes at such an early stage in the subject. Nevertheless, more than fifty years after Hoyle’s papers there are still major gaps in the theory of galaxy formation. This is because complex physical effects, in particular energy injection from massive stars, are important in determining the baryonic content and angular momentum properties of galaxies.

I will end with a quotation from Fred Hoyle’s book *Galaxies, Nuclei and Quasars*, published in 1966:

‘It is not too much to say that the understanding of why there are these different kinds of galaxy, of how galaxies originate, constitutes the biggest problem in present-day astronomy. The properties of the individual stars that make up the galaxies form the classical study of astrophysics, while the phenomena of galaxy formation touches on cosmology. In fact, the study of galaxies forms a bridge between conventional astronomy and astrophysics on the one hand, and cosmology on the other.’

This remains as true today as it was nearly fourty years ago.

4.1. Acknowledgments

I think Lisa Wright and Vince Eke for allowing me to reproduce Figure 2 and Carlton Baugh for providing a copy of Figure 4.

REFERENCES


