THE ASTROPHYSICAL JOURNAL, Vol. 158, December 1969

GROUP VELOCITY OF SPIRAL WAVES IN GALACTIC DISKS

ALAR TOOMRE
Department of Mathematics, Massachusetts Institute of Technology
Received 1969 March 13; revised 1969 June 2

ABSTRACT

Studied here are density waves of the kind proposed especially by Lin to explain the spiral structures of disk galaxies. It is shown that any packet of such waves propagates radially (and toward increasingly short wavelengths) with a group velocity that is sufficient to obliterate it within a few galactic revolutions This does not necessarily mean that the density-wave hypothesis is wrong. But it does imply that any existing spiral waves in the disk of a galaxy must somehow be replenished if their pattern as a whole is to persist. Three conceivable sources of such replenishment are also discussed in this paper.

I. INTRODUCTION

It has now been suggested by a number of authors starting with B. Lindblad (1963, and several earlier papers) and Lin and Shu (1964, 1966) that at least the grander spiral patterns in disk galaxies represent density waves governed primarily by gravity. Such a suggestion seems quite attractive, and indeed Lin, Yuan, and Shu (1969) have found it helpful in reconciling certain observations concerning this Galaxy. But the theory remains incomplete. Among other things, it has not yet been established that any complete disk would permanently admit spiral waves as self-consistent *modes* of oscillation. On the contrary, indications such as the "anti-spiral theorem" of Lynden-Bell and Ostriker (1967) for differentially rotating gas disks have raised doubts whether any purely oscillatory (that is, neither growing nor decaying) modes of spiral planform can be found at all.

And yet, though it may be premature to speak of spiral waves as true modes of oscillation, it seems entirely appropriate to ask how some postulated spiral wave pattern in a galactic disk would *evolve* with time. It is the latter type of question which this paper tries to answer. It does so by calling attention to a group velocity that should be applicable to any reasonably tightly wound spiral wave. This group velocity describes at least qualitatively how various information from the given disturbance propagates radially, and that in turn has important implications to any permanent (or quasi-permanent) maintenance of spiral wave patterns.

Like the WKBJ analyses of Lin and Shu, the present remarks will be confined to a single azimuthal Fourier component of disturbance to an originally axisymmetric thin disk. Such a component of, say, the perturbation surface density can always be written as

$$\mu'_m(r,\theta,t) = S(r,t)\cos[\Phi(r,t) - m\theta]\,, \tag{1}$$

where r is the distance from the disk center, θ denotes longitude increasing in the direction of rotation, and t is time.

Lin and Shu considered that (i) the total disturbance is infinitesimal, (ii) both the amplitude $S(r,t)$ and the radial wavenumber

$$k(r,t) = -\partial\Phi/\partial r \tag{2}$$

vary only slowly with r (in order to represent a smooth spiral), (iii) the angular wavenumber m is a small integer (usually 2) but $|k|r \gg 1$ for all radii of interest (to make that spiral tightly wound), and (iv) $S(r,t)$ and $k(r,t)$ are both independent of t, and the frequency

$$\omega(r,t) = \partial\Phi/\partial t \tag{3}$$

is the same for all r and t (to postulate a single mode).

The first three of those assumptions also apply here. But instead of dealing explicitly with modes, this discussion admits from the start that $S(r,t)$, $k(r,t)$, and $\omega(r,t)$ may all depend weakly on both of their arguments. We simply postulate at first that any approximate dispersion relation of the kind

$$F(k,\omega,r;m) = 0 \tag{4}$$

estimated for modes is also adequate for describing the local behavior of those slightly more general disturbances. This additional assumption is often made (cf. Whitham 1960; Lighthill 1965) in semiquantitative discussions of wave trains or packets with gradually varying properties. As usual, it seems very reasonable, but we will here not even attempt to defend it rigorously.

II. PROPAGATION OF INFORMATION

a) Group Velocity

Imagine that equation (4) has been solved for ω as

$$\omega = f(k,r;m) . \tag{5}$$

Then the above is equivalent to supposing that

$$\partial\Phi/\partial t = f(- \partial\Phi/\partial r,r;m) \tag{6}$$

regardless of whether those partial derivatives themselves vary with time.

Equation (6) is ripe for a "kinematic" derivation of group velocity, the merits of which have been reviewed by Whitham and Lighthill among others. Just by differentiating it with respect to time, and using equations (2) and (3), one obtains

$$\frac{\partial\omega}{\partial t} + \left(\frac{\partial f}{\partial k}\right)_r \frac{\partial\omega}{\partial r} = 0 . \tag{7}$$

Likewise, a differentiation with respect to r gives

$$\frac{\partial k}{\partial t} + \left(\frac{\partial f}{\partial k}\right)_r \frac{\partial k}{\partial r} = -\left(\frac{\partial f}{\partial r}\right)_k . \tag{8}$$

These convective derivatives indicate that at least the frequency and wavenumber information in a slowly evolving spiral disturbance propagates radially with the speed

$$dr/dt = (\partial f/\partial k)_r \equiv c_g(k,r;m) , \tag{9}$$

which clearly plays the role of a group velocity.

b) Characteristic Curves

Viewed geometrically, equations (7) and (8) define a variety of so-called characteristic curves or "rays" in the (r,t)-plane. Along every such curve the frequency ω remains constant and the wavenumber k changes at a known rate. The local "slope" dr/dt, moreover, is given by equation (9). Generally, this group velocity varies even along any given characteristic curve. But it can everywhere be determined from the known k and r, or—implicitly via equation (4) or (5)—from r and the conserved value of ω.

c) Dispersion Relation

Let us now apply these concepts to that important version of Lin and Shu's (1966, eq. [4.1]) dispersion relation which refers to an infinitesimally thin disk composed only of stars with a Schwarzschild distribution of velocities.

With its terms slightly rearranged, this reads

$$[\omega - m\Omega(r)]^2 = \kappa^2(r) - 2\pi G\mu(r)\,|k|\,\mathfrak{F}_\nu(\chi)\ , \tag{10}$$

where $\Omega(r)$ is the equilibrium angular speed of rotation, $\kappa(r) = r^{-3/2}\,[d(r^4\Omega^2)/dr]^{1/2}$ is the epicyclic frequency, G is the gravitational constant, $\mu(r)$ is the unperturbed projected mass density, and

$$\begin{aligned}\mathfrak{F}_\nu(\chi) &\equiv \frac{1-\nu^2}{\chi}\left[1 - \frac{\nu\pi}{\sin \nu\pi}\,\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-\chi(1+\cos s)}\cos \nu s ds\right] \\ &= (2/\chi)(1-\nu^2)\exp(-\chi)\sum_{n=1}^{\infty}[1-(\nu/n)^2]^{-1}I_n(\chi)\ .\end{aligned} \tag{11}$$

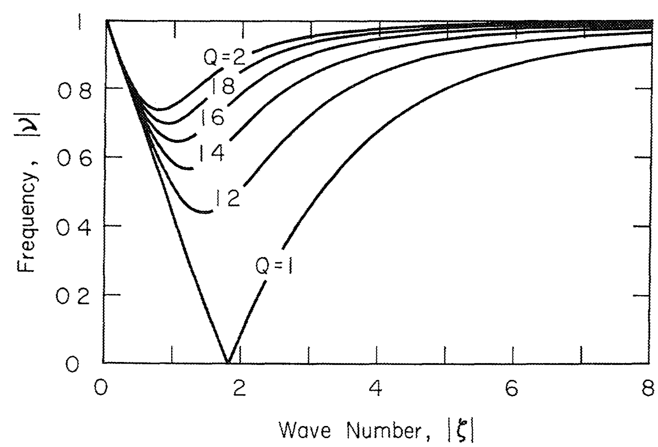

FIG. 1.—The Lin-Shu-Kalnajs dispersion relation

Here $I_n(\chi)$ is a modified Bessel function, $\nu = (\omega - m\Omega)/\kappa$ is a dimensionless frequency, and $\chi = k^2\sigma_u^2/\kappa^2$ is an auxiliary variable involving σ_u, or the unperturbed rms radial speed of the stars.

The frequency ν implied by equation (10) has been plotted in Figure 1. The abscissa is a dimensionless radial wavenumber,

$$\zeta = k/\alpha_{\rm crit}\ , \tag{12}$$

where

$$\alpha_{\rm crit}(r) = \kappa^2/2\pi G\mu \tag{13}$$

is that wavenumber beyond which a cold, thin disk becomes gravitationally unstable in the axisymmetric sense (cf. Toomre 1964). (These curves are reflection-symmetric about the ζ- and ν-axes. Hence only the absolute values of those variables appear.) The labels

$$Q = \sigma_u/\sigma_{u,\rm min} \tag{14}$$

on the different curves compare the assumed velocity dispersion with the minimum

$$\sigma_{u,\rm min} \cong (0.2857)^{1/2}\kappa/\alpha_{\rm crit} \tag{15}$$

that is needed to avoid all local axisymmetric instabilities in this disk of stars (Toomre 1964; Kalnajs 1965).

Owing to the fact that the force estimates implicit in Lin and Shu's dispersion relation neglect the slight inclinations of the spiral waves, the curves shown in Figure 1 (as well as the latter form of eq. [11]) are identical with those implied by Kalnajs's analysis of strictly axisymmetric oscillations. The latter have a simple physical interpretation: The self-gravitation of the disk material generally reduces the frequency of *collective* oscillation below the epicyclic frequency $\kappa(r)$ which alone would be feasible in the absence of interactions[1] and which always remains the basic frequency of oscillation of individual stars. As indicated by Figure 1, the amount of that reduction depends both on the wavelength and on the stability parameter Q.

With this understanding, equation (10) is best viewed simply as a consistency relation equating the rates at which various parts of a disk are *able* to vibrate in an essentially axisymmetric manner and those at which they are *required* to oscillate in order to constitute a single mode. These conceptually distinct frequencies are compared in equation (10) as they would appear to an observer (or to a group of stars) orbiting with the approximate mean speed $\Omega(r)$: The left-hand side is the square of the required frequency; the right-hand side is the square of the frequency that is actually feasible at the given radius and wavelength.

Except for modes, the frequencies $\omega = f(k,r;m)$ implied by equation (10) will hardly be the same at all radii, nor will ω at any given radius generally remain constant with time. The latter expectation stems from the different angular rates ω/m of wave travel at various radii. Owing to this, the wave pattern must shear gradually, the radial wavenumber will change, and that in turn will affect the intrinsic frequency of oscillation via the self-gravitation.

To follow such changes in detail, let the curves in Figure 1 be abbreviated by the function

$$|\nu| = N(|\zeta|;Q) , \tag{16}$$

so that equation (10) may be rewritten

$$\omega = f(k,r;m) = m\Omega(r) + \operatorname{sgn}(\nu)\kappa(r)N(|\zeta|;Q) , \tag{17}$$

where $\operatorname{sgn}(\nu) = \pm 1$ depending on whether $\nu \gtrless 0$. Then the group velocity from equation (9) becomes

$$dr/dt = c_g(k,r;m) = \operatorname{sgn}(\zeta\nu)[\kappa(r)/\alpha_{\text{crit}}(r)]\partial N/\partial|\zeta| , \tag{18}$$

and it also follows that

$$\frac{d\zeta}{dt} = \frac{\kappa(r)}{\alpha_{\text{crit}}(r)}\left(\frac{d\nu}{dr} - \operatorname{sgn}(\nu)\frac{\partial N}{\partial Q}\frac{dQ}{dr}\right) \tag{19}$$

along any given characteristic curve. Although numerical values of $\partial N/\partial|\zeta|$ can easily be estimated from Figure 1, let it be recorded for § IV*a* that

$$\frac{\partial N}{\partial|\zeta|} = \left(1 + 2\frac{\partial \ln \mathfrak{F}_\nu(\chi)}{\partial \ln \chi}\right)\Bigg/\frac{\partial}{\partial|\nu|}\left(\frac{1-\nu^2}{\mathfrak{F}_\nu}\right). \tag{20}$$

d) Behavior in the Large

To examine one fairly typical set of such characteristic curves in the (r,t)-plane, consider that infinite model galaxy in which

$$\Omega(r) = V/r \quad \text{and} \quad \mu(r) = V^2/2\pi Gr , \tag{21}$$

[1] In his thesis, Kalnajs also mentions other axisymmetric oscillations whose $\mu \to 0$ frequencies are 2κ, 3κ, ... But such overtone oscillations are confined mainly to velocity space, and they were justifiably ignored by Lin and Shu even though contained in equations (10) and (11).

with speed V = constant. Assume that its local stability parameter Q is likewise independent of the radius. For this model,

$$\kappa(r)/a_{\text{crit}}(r) = 2\pi G\mu(r)/\kappa(r) = V/\sqrt{2}\ . \tag{22}$$

Therefore, the relative frequency $\nu(r;\omega,m)$ implied by a given ω works out as

$$\nu = (\omega - m\Omega)/\kappa = (m/\sqrt{2})[(r/r_p) - 1]\ , \tag{23}$$

where $r_p = mV/\omega$ is the radius at which the assumed angular wave speed ω/m matches that of the average material.

Now recall from § II*b* that it is ω that is conserved along each characteristic curve. Hence $d\nu/dr$ there follows directly from equation (23) and this, together with equations

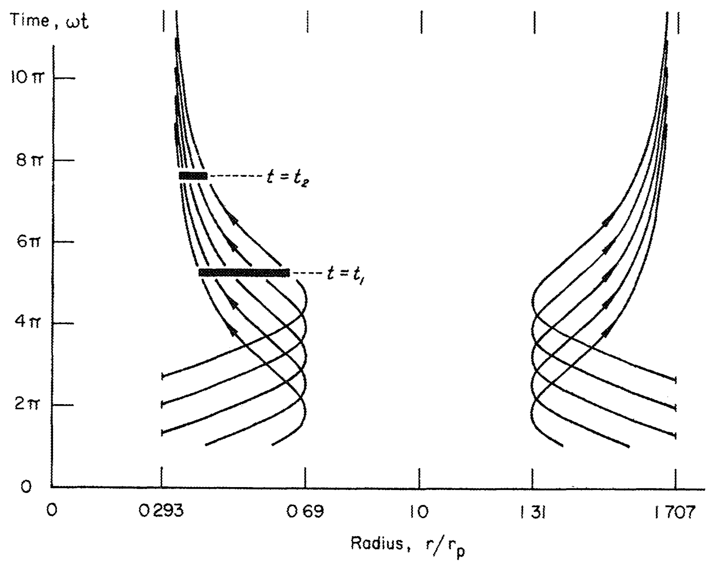

FIG. 2.—Some $m = 2$ characteristic curves for a disk in which $Q = 1.2$. (The early parts of these curves should be treated only as extrapolations, since the waves then are still very open.)

(19) and (22) and the fact that $dQ/dr = 0$ here, implies that the dimensionless wavenumber ζ increases at a constant rate

$$d\zeta/dt = \omega/2 \tag{24}$$

along every such curve.

These conveniently linear dependences of ν upon r, and of ζ upon t, reduce the present task of obtaining characteristic curves in the (r,t)-plane to little more than a relabeling of the (ν,ζ)-axes in Figure 1. In fact, Figure 2, which shows a set of such curves referring to a *trailing* (or $\zeta > 0$) spiral pattern of a *single* frequency ω and of angular wavenumber $m = 2$, was constructed with a template shaped exactly like the $Q = 1.2$ curve from the former diagram.

All radii in Figure 2 have been normalized with respect to the particle-resonance radius $r_p(\omega;m)$. Hence the so-called inner Lindblad resonance ($\nu = -1$) occurs there at $r/r_p = (1 - 2^{-1/2}) \simeq 0.293$, and the outer Lindblad resonance ($\nu = +1$) occurs at $r/r_p \simeq 1.707$. Between those two radii lies what Lin, Yuan, and Shu (1969) call the "principal part of the spiral pattern." Outside that range, $|\nu| > 1$ and no radially oscillatory WKBJ solutions are possible. However, as here indicated by the gap $0.69 < r/r_p < 1.31$, an additional annulus must be excluded whenever the stability parameter Q

exceeds unity; this is necessary because the local self-gravity is then also incapable of reducing $|\nu|$ below a certain minimum.

The fact that the characteristic curves in Figure 2 cross each other is of little consequence to the *forward* time development of an already tightly wound and trailing (i.e., $\zeta \gtrsim 1.5$) spiral wave whose $\omega \simeq$ const. at various radii. Consider, for instance, the wave-number and frequency information contained in the range $0.4 \lesssim r/r_p \lesssim 0.65$ at time t_1. According to these curves, such information is simply conveyed inward (e.g., $t = t_2$). It ends up eventually in the vicinity of the inner Lindblad resonance appropriate to the assumed value of ω. Likewise, any similar information beyond the particle resonance at $r = r_p$ tends to be carried outward to the outer Lindblad resonance. In addition, the tightness of winding of the spiral waves associated with either set of characteristic curves increases inexorably with time (cf. eq. [24]).

The interesting *past* history of the marked wave group in Figure 2 shows, however, that the radial transport along any given characteristic curve has not always been in the same direction. Obviously the details of the turnaround or refraction near $r/r_p = 0.69$ demand a more elaborate analysis. And especially for this disk, waves with $|\zeta| \lesssim 1$ also do not lend themselves to WKBJ analysis because they are as yet wound too loosely. But despite these shortcomings, it seems undeniable that the sense of propagation of that (proto)group must have been outward when its constituent waves had not yet become wrapped as tightly as $\zeta \simeq 1.5$.

e) Implications for the Galaxy

All the qualitative behavior discussed in the last three paragraphs carries over to the trailing and two-armed ($m = 2$) spiral wave which, according to Lin and Shu (1967) and Lin, Yuan, and Shu (1969), is present in the Galaxy. The only significant changes are: (i) For a pattern speed $\omega/2 \simeq 12.5$ km sec^{-1} kpc^{-1} as judged in the latter paper, the particle resonance $\nu = 0$ occurs at $r \simeq 17$ kpc. This means that any $\nu > 0$ oscillations lie essentially outside the Galaxy and may be ignored. (ii) Unlike our equation (23), the relationship $\nu = \nu(r)$ is now nonlinear, as reflected by the values $\nu = -1, -\frac{3}{4}, -\frac{1}{2}$, and $-\frac{1}{4}$ occurring at $r \simeq 3.5$, 10.5, 13, and 15 kpc, respectively.

To estimate the group velocity near the Sun, recall that Lin *et al.* considered this star disk with the observed $\sigma_u \simeq 35$ km sec^{-1} to be barely stable, or $Q \simeq 1.0$. Via equations (14) and (15), this implies immediately that $\kappa/\alpha_{\rm crit} \simeq (0.2857)^{-1/2}(\sigma_u/Q) \simeq 65$ km sec^{-1}, or that $\lambda_{\rm crit} \equiv 2\pi/\alpha_{\rm crit} \simeq 13$ kpc.[2] As a check, note also that Lin *et al.* judged a radial wavelength $\lambda \equiv 2\pi/k = 3$–4 kpc in this vicinity to be consistent with the cited wave speed. This means that $\zeta \equiv k/\alpha_{\rm crit} \simeq 4$. Indeed, the line $|\nu| = \frac{3}{4}$ in Figure 1 intersects the $Q = 1.0$ curve near that value of the abscissa. But, more important, the slope of that curve thereabouts is $\partial N/\partial|\zeta| \simeq 0.15$.

From these data and from equation (18) it follows that the nearby group velocity is $dr/dt \simeq -10$ km sec^{-1}. The minus sign merely reaffirms that the proposed spiral waves are already wound so tightly that their information propagates radially *inward*, or away from the particle-resonance radius. The actual speed may not seem remarkable. But even at a rate of only 10 km sec^{-1} one would cover 10 kpc in about 10^9 years, or in roughly 4 galactic years as figured at the Sun. Of course, this is not to say that such a group velocity would be sustained along any given characteristic curve, or that the "message" would ever reach the galactic center. However, it can be verified by simple integration that the time of propagation from $r = 12$ kpc, say, to $r = 5$ kpc is at most 10^9 years, regardless of whether $Q = 1.0$, 1.2, or 1.4.

[2] This shortcut avoids any explicit corrections for the finite disk thickness. Taken literally, our deduced wavelength $\lambda_{\rm crit}$ implies a local surface density $\mu = 77\ M_\odot$ pc^{-2}. But that is only an *equivalent* density for a comparable disk of infinitesimal thickness, and is not to be confused with the total $\mu \simeq 90\ M_\odot$ pc^{-2} adopted by Lin *et al.* for a modified version of Schmidt's (1965) model of the Galaxy.

The above estimates obviously raise the question of how any existing spiral wave could have persisted longer than a few galactic years. We shall return to that in § IV.

III. DENSITY WAVES IN A LOCAL MODEL

The following computer experiment corroborates much of the preceding theory. It also shows emphatically that wave energy is transported in an analogous manner.

a) Review of the Model

This experiment rests entirely on the thin, *local* model of a shearing disk of stars that was introduced by Julian and Toomre (1966, hereinafter JT) to discuss certain forced responses of disk galaxies. Like the well-known Hill problem (cf. Szebehely 1967, p. 608) from restricted three-body theory, or the model of a shearing gas disk devised by Goldreich and Lynden-Bell (1965), that model assumes that the total *radial* extent of any region of interest is much less than its distance from the center of rotation. It also supposes that any relative motion is only a small fraction of the full rotation speed.

With this rationale, the model pretends that even the perturbed (r,θ) motion

$$x = r - r_0\,, \qquad y = r_0(\theta - \Omega_0 t) \tag{25}$$

of any given star is described exactly by the equations

$$\frac{d^2x}{dt^2} - 4\Omega_0 A_0 x - 2\Omega_0\frac{dy}{dt} = F_x'\,, \qquad \frac{d^2y}{dt^2} + 2\Omega_0\frac{dx}{dt} = F_y'\,. \tag{26}$$

Here r_0 is some suitable reference radius, Ω_0 stands for $\Omega(r_0)$, A_0 is the Oort constant $-\frac{1}{2}r_0(d\Omega/dr)_0$, and F_x' and F_z' are the radial and circumferential components of any disturbance self-gravitation and/or imposed force. Less important, the model also supposes that F_x' and F_z' may be deduced from densities in the plane exactly as if the coordinates x, y were truly Cartesian and the disk exceedingly thin, that the unperturbed stellar velocities are everywhere distributed about the mean $\langle dx/dt\rangle = 0$, $\langle dy/dt\rangle = -2A_0 x$ according to a given Schwarzschild distribution, and that all disturbances to that uniform state are infinitesimal.

The small-scale assumption is obviously very drastic. But aside from that, this idealized thin stellar system has several advantages as an example: (i) Its analysis requires no further asymptotic approximations of any sort, nor is it limited to tightly wrapped waves. (ii) It embodies the basic effects both of Coriolis forces and of the differential rotation. (iii) Its axisymmetric (or y-independent) vibrations obey the Lin-Shu-Kalnajs $m = 0$ dispersion relation (10). (iv) All its nonaxisymmetric disturbances may be decomposed into shearing wavelets. The separate time development of each of those can be followed through a Volterra integral equation (eq. [21] of JT), and the results can subsequently be reassembled. (v) And finally, for any pattern of y-dependence like sin or cos $(2\pi y/\lambda_0)$, presumed to remain stationary to an observer orbiting with angular speed Ω_0, the model also exhibits Lindblad resonances. Those occur here at $x = \pm x_L$, where

$$x_L = (\kappa_0/2A_0)(\lambda_0/2\pi) \tag{27}$$

and where $\kappa_0 = [4\Omega_0(\Omega_0 - A_0)]^{1/2}$ is the relevant epicyclic frequency.

b) Results

The present computations are in fact confined to a single y-harmonic. They presume the entire stellar disturbance to have been provoked by the gravity forces of a hypothetical mass distribution of infinitesimal surface density

$$\mu_{\rm imp}(x,y,t) = \epsilon\mu_0 \exp\left[-(x - x_0)^2/x_L^2\right] \cos(2\pi y/\lambda_0) \exp(-t^2/t_0^2) \tag{28}$$

introduced in the plane of the model. For the moment, let us not ask what this forcing could conceivably correspond to in a real galaxy. Just note that these perturbing forces *could* in principle have been realized with a suitable distribution of extraneous masses, that they are periodic in y but lack any spiral bias, and that they are applied only for a limited time of $O(t_0)$ and predominantly at a distance x_0 from the radius r_0 at which they would appear most nearly stationary.

Typical of the results of these computations is Figure 3 showing the evolution of the

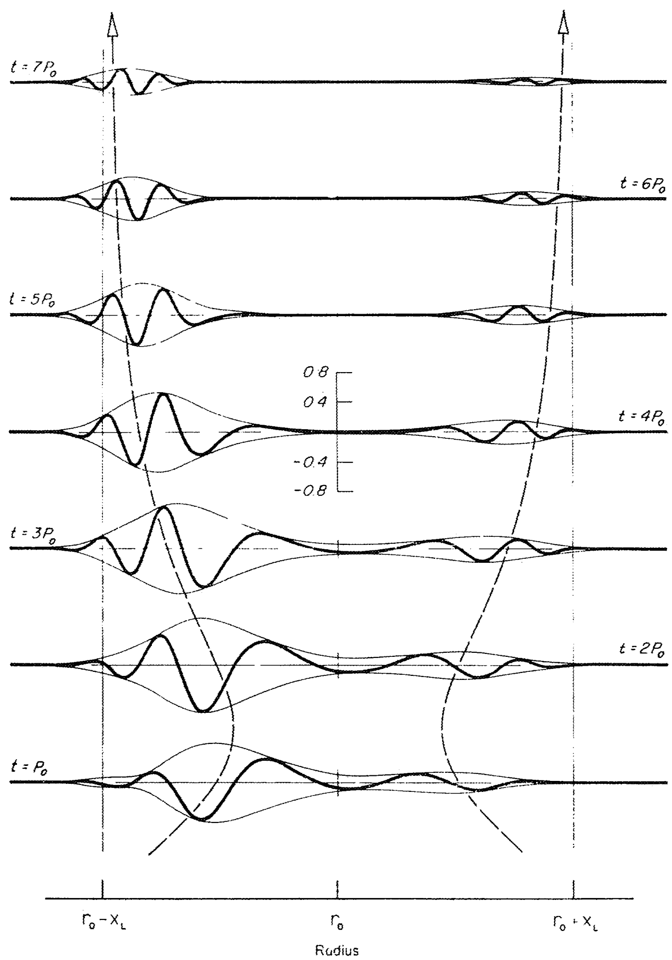

FIG. 3.—An evolving density wave in the $Q = 1.2$ local model. Heavy curves denote density amplitudes $S_c(x,t)$ (cf. eq. [29]), and the envelopes represent $\pm[S_c^2 + S_s^2]^{1/2}$, at seven equally spaced instants of time. The broken curves are two of the (r,t) characteristic curves defined in § II.

induced stellar disturbance density

$$\mu'(x,y,t) = \epsilon\mu_0[S_c(x,t)\cos(2\pi y/\lambda_0) + S_s(x,t)\sin(2\pi y/\lambda_0)] . \tag{29}$$

That figure refers to a situation where

$$Q = 1.2 , \quad A_0 = \tfrac{1}{2}\Omega_0 , \tag{30}$$

and

$$x_L = \lambda_{crit} , \quad x_0 = -1.5x_L \quad t_0 = 2P_0 , \tag{31}$$

with $P_0 = 2\pi/\kappa_0$ denoting the epicyclic period. The two parameters in equation (30) merely describe the unperturbed state. The three in equation (31) refer to the excitation; their choice was influenced by the following, more subjective considerations: (i) As verified with a similar calculation for $x_L = 3\lambda_{crit}$, the WKBJ expectations would indeed have been met more accurately with a larger ratio x_L/λ_{crit}. However, judging from the rough equality of the characteristic length $\lambda_{crit} \equiv 4\pi^2 G\mu_0/\kappa_0^2$ and of the separation between the inner Lindblad and the particle resonance radii in this Galaxy according to Lin *et al.*, the choice $x_L = \lambda_{crit}$ seems more realistic. (ii) The offset $x_0 = -1.5\ x_L$ was designed to insure that the "inner" ($-x_L < x < 0$) strip of the model was excited considerably more than the "outer" ($0 < x < x_L$). (iii) The chosen duration was found to yield about the most compact wave packet.

In other respects, Figure 3 almost speaks for itself. For instance, it clearly exhibits the expected shortening with time of the typical radial wavelengths within the propagating packet. Also, the almost negligible motion of individual wave crests and nodes in that diagram confirms that the wave frequency is approximately conserved everywhere—that is, it *remains* small as viewed from our rotating coordinates. What is more, the smallness of that frequency itself is nicely explained by the fact that the exciting force field did not move with respect to these coordinates.

But the most striking thing about Figure 3 is the advance of the wave envelopes in the same qualitative manner as predicted for the wavenumber and frequency information. (That expected behavior is here represented by the pair of characteristic curves emanating from $r_0 \pm x_L$ at $t = 0$. Note that the relationship $\nu = \nu(r)$ is again linear in this example.) Although the initial propagation of wave energy toward r_0 is not shown clearly in this diagram, it is evident that most of that energy remains on the side where it was introduced. The amount of "tunneling" or "barrier jump" from one side of r_0 to the other may be estimated from the fact that the same imposed forces (not entirely negligible on the side $x > 0$) led, in the case $x_L = 3\lambda_{crit}$, to an "outer" disturbance of roughly half the relative amplitude shown here. Thus, easily 90 percent of the wave energy in the present example seems to be reflected from the forbidden annulus surrounding r_0.

The subsequent decay of either wave packet in Figure 3 stems from a slight phase mixing—or, broadly speaking, a Landau damping—of the perturbed oscillations of various stars even in the presence of collective forces. As discussed by JT, such a decay is the eventual fate of all nonaxisymmetric disturbances to our idealized stellar system when the forcing does not persist indefinitely. Only the rate varies, increasing fairly rapidly with decreasing circumferential wavelength: This dependence may be gauged either from equation (32) of JT or from Figure 6 there. It is also reflected by the fact that that damping was barely evident at $t = 15\ P_0$ in the $x_L = 3\lambda_{crit}$ calculations, whereas in the case $x_L = \frac{1}{2}\lambda_{crit}$ it proved virtually complete by $t = 2.5\ P_0$. (Those specific instants are cited because the typical *radial* wavelengths in the respective cases are then roughly the same as those found at $t = 5\ P_0$ in Fig. 3).

Figure 4 refers to the same transient forcing as Figure 3. It differs only in its vertical scale and the fact that the stability parameter Q now equals unity instead of the previous $Q = 1.2$. Such an equilibrium state, of course, is only marginally stable with respect to certain axisymmetric disturbances. Hence it seems likely (e.g., Julian 1967) that real star disks are by now somewhat "hotter" than this.

Nevertheless, the case $Q = 1.0$ is of theoretical interest because Figures 1 and 2 left open the possibility that a wave packet then originating on one side of r_0 might simply travel to the other side with hardly any reflection. However, Figure 4 shows the truth to be more complicated: It appears instead that when the packet arrives in the general vicinity of r_0—which also means that its radial wavelengths have shrunk to $O(\frac{1}{2}\lambda_{crit})$—the density wave amplifies roughly threefold owing to the gravitational near-instability. The ensuing pattern at $t = 3\ P_0$ looks remarkably uniform, symmetric, and extensive.

But soon thereafter, both the group transport and the phase mixing reassert themselves, and the final outcome resembles that of Figure 3.

IV. DISCUSSION

The present estimates were meant as no substitutes for a comprehensive study of disturbances to a complete galactic disk. But they leave little doubt that the fate of any *already tightly wound* packet of spiral density waves in a disk of *stars* may be summarized

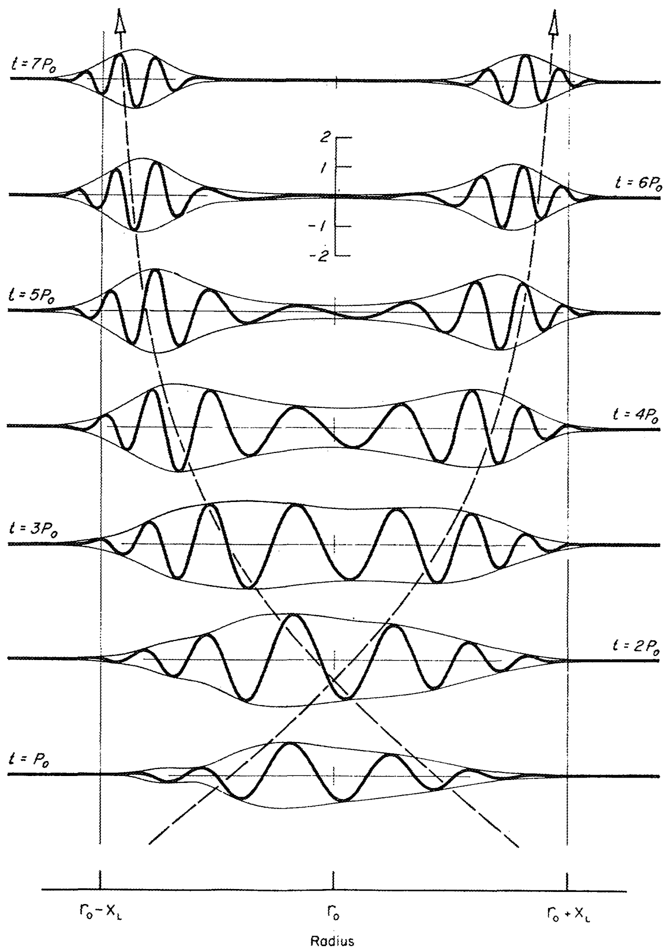

FIG. 4.—An evolving density wave in the $Q = 1.0$ local model

as follows:

i) In effect, even such *waves* are subject to differential rotation. The latter is not simply the material shear $d\Omega/dr$ (which is, typically, several times faster). Yet the effect is qualitatively the same, as least when $d\nu/dr > 0$ and $dQ/dr \simeq 0$ (cf. eq. [19]): As viewed from their packets, all "trailing" waves tend to wrap yet more tightly, and those wound in a "leading" sense tend to loosen.

ii) Simultaneously, such packets of waves also drift in radius. The relevant group velocity was derived in § II*a*. Strictly speaking, that simple derivation refers only to frequency and wavenumber data. But the examples of § III suggest that wave energy propagates roughly likewise.

iii) The time scale of that propagation is of the order of a few galactic years.

iv) The destination of any trailing wave packet of approximate frequency ω is the vicinity of the corresponding Lindblad resonance, *provided this* (and also any needed turnaround radius) *lies within the disk.*[3]

v) Especially with the last proviso, it is clear that the wave packet will eventually decay through phase mixing: Simply because the wavelengths then will decrease indefinitely, it must sooner or later become meaningless to regard any remaining oscillations as collective. However, this does not mean that most of the wave energy (or, more precisely, wave action—cf. § IV*a* below) will necessarily be deposited into random motions *near* a Lindblad radius. On the contrary, as Figures 3 and 4 suggest, much of that may be lost already in transit.

Needless to say, these conclusions about the group velocity constitute a serious criticism of the existing theoretical case for spiral waves in galaxies. However, note also that neither the damping nor the radial propagation of individual wave packets actually excludes really long-lived spiral wave patterns in a galaxy. If such patterns are to persist, the above simply means that fresh waves (and wave energy) must somehow be created to take the place of older waves that drift away and disappear.

Where could such fresh and relatively open spiral waves conceivably originate? The only three logical sources seem to be: (*a*) Such waves might result from some relatively *local* instability of the disk itself. (*b*) They may be excited by tidal forces from *outside*, such as from a companion or satellite galaxy. (*c*) Or they might be a by-product of some truly *large-scale* (but not necessarily spiral) distortion or instability involving an entire galaxy.

The rest of this paper is simply a review and reappraisal of each of these three possibilities. Although rather lengthy, these remaining comments are not to be regarded as exhaustive!

a) *Local Amplification*

No local mechanism seems yet to have been established whereby spiral density waves in a disk that is stable with respect to axisymmetric disturbances can grow spontaneously at wavelengths comparable to the observed radial spacings of spiral arms. To be sure, the pronounced growth of certain shearing disturbances to differentially rotating gas disks that was discovered by Goldreich and Lynden-Bell (1965) may appear to be one such candidate. However, though the same transient amplification was later found also in star disks by JT, the present writer (cf. JT, and also § IV*c* below) regards it, not as any true instability, but only as evidence that shearing disks are very willing to *respond* in a spiral manner.

Different "indications" of a local overstability of a disk of stars were reported by Lin and Shu (1966). Shu (1968) has since elaborated on them in his thesis. But it now turns out that those signs, too, were slightly misinterpreted.

Shu's otherwise comprehensive second-order WKBJ analysis of the various gradient effects contains two small errors. Upon their removal, that analysis predicts in fact that the density amplitude $S(r)$—cf. eq. (1)—of any steady wave of frequency ω should obey

$$\frac{d}{dr}\left[\frac{rS^2(r)}{k^2(r)}\left(1 + 2\,\frac{\partial \ln \mathfrak{F}_\nu(\chi)}{\partial \ln \chi}\right)\right] = 0\,. \tag{32}$$

In combination with the dispersion relation (10), with the group velocity c_g as given by equations (18) and (20), and with the fact that the wave energy density works out as a nonnegative

$$E = \frac{\pi G}{2\alpha_{\rm crit}(r)}\,S^2\nu\,\frac{\partial}{\partial\nu}\left(\frac{\mathfrak{F}_\nu(\chi)}{1-\nu^2}\right) \tag{33}$$

[3] This condition is not always met. For instance, as noted before, the $\nu = +1$ resonance corresponding to the spiral wave deduced by Lin, Yuan, and Shu occurs well outside the Galaxy. More important, even the inner Lindblad resonance ($\nu = -1$) may be absent if the true frequency ω of an assumed $m = 2$ disturbance is too large. In such cases, the given wave packet must in some sense be reflected either from the outer edge of the disk or from its center; in the process, its character will presumably change from trailing to leading, and the sign of the group velocity should also reverse.

in this nominal case of the Schwarzschild distribution,[4] equation (32) states simply that

$$\frac{d}{dr}\left(\frac{rc_gE}{\omega - m\Omega}\right) = 0 . \tag{34}$$

The above refers, of course, to only a single time harmonic. But by superposing a continuum of such "modes" and applying the usual stationary-phase arguments to integrals like

$$\int S(r;\omega) \cos [\phi(r;\omega) - \omega t] d\omega , \tag{35}$$

it is easily deduced from equation (34) that the corresponding statement for a slowly evolving disturbance must be

$$\frac{\partial}{\partial t}\left(\frac{E}{\omega - m\Omega}\right) + \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{rc_gE}{\omega - m\Omega}\right) = 0 . \tag{36}$$

This last result implies that, just as for a wide class of shearing fluid motions discussed by Whitham (1965) and by Bretherton and Garrett (1968), it is not exactly the wave energy density but rather the wave "action density," here $E/(\omega - m\Omega)$, which propagates with the group velocity c_g. *Note added in proof:* This particular action density has recently been identified by Kalnajs as just the excess density of angular momentum associated with the wave.

With that answer, Shu's work unwittingly closes the main logical gap of the present paper. However, it also means that no true "gradient instability"—nor indeed any damping[5]—exists to this asymptotic order in a disk of stars.

b) Tidal Forcing

One alternative that deserves very serious consideration is that much of any spiral density wave in *this* Galaxy may have evolved from vibrations set up during a close passage of at least the Large Magellanic Cloud some 5×10^8 years ago.

Such an approach of the LMC to a perigalactic distance of 20 or 25 kpc was judged necessary by Hunter and Toomre (1969) to explain the well-known warp of the outer plane of this Galaxy. Subsequent theoretical efforts (Toomre 1970) to match the amplitude and phase of that observed vertical distortion indicate that the orbit of the LMC must be of relatively low inclination. Assuming the standard distance $R_0 = 10$ kpc to the galactic center, the eccentricity of that orbit appears to be roughly 0.5, and the minimum mass required of the LMC comes close to $3 \times 10^{10}\, M_\odot$. The sense of revolution is less certain, but both the bending and some circumstantial evidence suggest a direct—as opposed to a retrograde—orbit.

If one accepts the close approach, it is only this last point that leaves much uncertainty about the "tidal wave" excited within the plane of this Galaxy: In either kind of orbit, the LMC would have spent only some 1.5–2.5 $\times 10^8$ years traversing the nearest 90° of galactocentric longitude. In the retrograde case, the implied angular speed of

[4] This energy density was here calculated literally as the net work per unit area done between $t = -\infty$ and $t = 0$ in the course of "waving" an extraneous, axisymmetric, and very slowly ($s \rightarrow 0$) growing mass distribution

$$\mu_{\rm imp}(x,t) \propto se^{st} \cos (kx \pm \nu\kappa t)$$

in the plane of the initially undisturbed infinite stellar sheet defined in § III*a*.

[5] It seems quite likely that such "Landau damping" as exhibited in Figures 3 and 4 may be absent to all orders in an asymptotic expansion in powers of the (supposedly small) ratio of the radial to the circumferential wavelengths. This remark is based on the fact that the phase mixing associated with the differential rotation, at least, seems to decrease with exponential rapidity as the inclination of waves of any given length perpendicular to their crests is chosen smaller and smaller (cf. the term exp $(-\beta^2 S^2 \tau_d^2/2)$ in JT's eq. [32]).

6–10 km sec^{-1} kpc^{-1} would have had no special significance. But in the more favored direct orbit, this speed would have roughly matched the speed of advance, $\Omega - \kappa/2$, of the slow $m = 2$ "dispersion orbit" (cf. B. Lindblad 1959) of any ring of supposedly noninteracting material in the outer half of the disk. And that, coupled with the dominant $m = 2$ character of the tidal force in the plane, means that any *direct* close passage of the LMC should have been unusually effective in exciting $\nu \simeq -1$, $m = 2$ oscillations of the Galaxy. It also suggests that, even with self-gravitation taken into account, the resulting "pattern speed" should have been of the order of 10 km sec^{-1} kpc^{-1}.

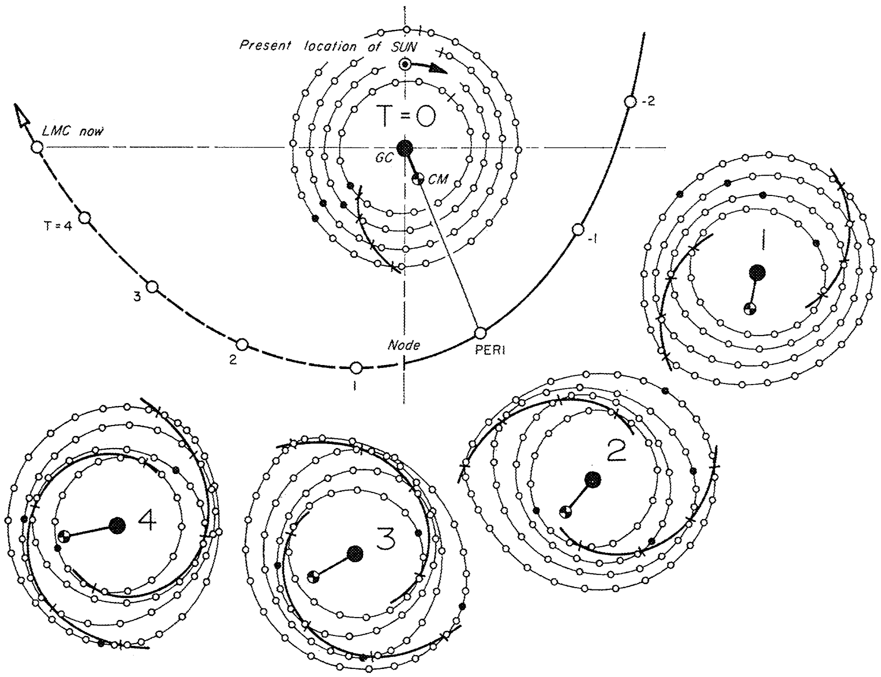

FIG. 5.—A time history of the displacements of four rings of noninteracting test particles provoked by a simulated direct passage of the LMC. All points have here been projected onto the present galactic plane, with the hypothetical mass point at the galactic center (= GC) treated as fixed. As explained in the text, the mass of the LMC was chosen differently for each of the four rings; the LMC orbit shown, and also the indicated positions of the center of mass (= CM) of the combined system, refer to $M_{\text{LMC}} = 2.0 \times 10^{10} M_\odot$ or the mass felt by the 10-kpc ring. Time in units of 10^8 years is here reckoned from the perigalactic point. The heavy spiral curves connect points on each ring which are at maximum distance from the GC.

But why should the simple tidal force of the LMC have caused a *spiral* wave? To understand that, consider Figure 5. Shown there are five successive "frames" from restricted three-body computations (patterned after those of P. O. Lindblad 1960; Pfleiderer and Siedentopf 1961; Pfleiderer 1963) in which the mass of the Galaxy was replaced by a single $1.2 \times 10^{11} M_\odot$ particle and in which the outer matter in this disk was represented by test particles orbiting initially in circles with radii of 8, 10, 12, and 14 kpc about that mass point. The LMC, represented likewise by a single mass point, was placed in a direct orbit of 35° inclination and a perigalactic distance of 25 kpc; its present distance was taken to be 55 kpc and its speed of recession 65 km sec^{-1}, both reckoned from the galactic center. Each calculation was begun with the LMC at the appropriate apogalactic point.

The violent commotion produced outward of about 12 kpc in this simulated passage of the LMC was foreshadowed by the results of Pfleiderer, and especially by the comments of Habing (1967) on even more closely related computations performed by himself and Visser. Indeed, with the LMC assigned a mass of $2.0 \times 10^{10} M_\odot$ here, roughly 15 percent of the test particles at 16 kpc are actually torn loose from their primary. Moreover, the transient spiral shapes of the very distorted 14- and 16-kpc rings found in such a calculation resemble those already reported by Pfleiderer, and thus are not reproduced here. Such outermost spiral structure seems largely a result of just an ordinary shearing of that far-flung material.

Not so evident in Pfleiderer's diagrams, however, is an additional wave shear that is probably more relevant to the bulk of the Galaxy. That is the reason why Figure 5 has been prepared as a *montage:* The particles initially at 14 kpc (together with the central mass point in that calculation) were exposed to a passing mass of only $0.5 \times 10^{10} M_\odot$, those at 12 kpc (and again the central particle) to a mass of $1.0 \times 10^{10} M_\odot$, and so forth in mass multiples of 2.

This montage has two advantages over an unadulterated diagram in which the outermost violence would tend to "steal the show." For one thing, it gives a fairly honest picture of the relative dynamical importance of the various rings, since their own realistic masses decrease outward roughly by factors of two. But, more important, it also emphasizes the relative phases of the various oval distortions, the latter being largely the aforementioned dispersion orbits by B. Lindblad. In each of those, the particle density is greatest near each end of the long axis. And, as we see here, the various ovals drift at different rates. It is just these last two facts together which explain qualitatively why even an initially unbiased vibration is sheared into a trailing spiral wave.

These test-particle calculations can, of course, be criticized for their total neglect of any interactions between the various particles. However, this is not to say that the self-gravity of these relatively low-density parts of the disk should immediately have been of major importance, nor does it contradict our qualitative picture about the evolution of the waves: For one thing, the relatively sudden passage of the LMC should have induced roughly the same initial velocities regardless of the subsequent *disturbance* gravity forces from within this system. And also, it seems that the principal effect of that later mutual attraction of the various disk particles should have been to enhance the shearing discussed above, since in effect it would have reduced the epicyclic frequency κ and thus caused the wave speeds $\Omega - \kappa/2$ at the various radii to become more disparate.

In addition, the self-gravitation should have had the following, more subtle effect: Once the shear had reduced the radial wavelengths below $\lambda_{crit}/2$ (cf. Fig. 3), it is now clear that the group transport would have begun to communicate the vigorous oscillations of the outer disk to the more massive but also more quiescent inner parts of the Galaxy. This last phenomenon is akin to a tail wagging the dog. As such, it runs contrary to much of our intuition. However, it has one corollary that is very welcome: The energy transport inward would itself have served to damp the outer oscillations from their embarrassingly large amplitudes as compared with present observations; by contrast, the same wave energy deposited into the denser parts is much less objectionable. The total elapsed time from the excitation to the eventual extinction of such a transient spiral wave has in effect been estimated already in § II*e:* It is roughly 10^9 years.

c) Forcing from Within

There are other galaxies in which a pronounced spiral structure might again be attributed to tidal forces from a presumed satellite. The most celebrated of those is, of course, M51. However, even granting the difficulty of observing faint companion galaxies, it seems far-fetched to suggest that any large-scale density waves in the majority of the disklike spirals could owe their existence to external influences.

In such galaxies, it appears from the present discussion that the cause of any reasonably tightly wound spiral waves not only must be internal but also must not be primordial, nor can it very plausibly be sought in any local overstabilities or amplification of the waves in transit. If this is indeed so, the only remaining alternative seems to be that such waves must themselves be the *consequences of some yet more basic density asymmetries*—e.g., disturbances such as mildly barlike waves or oval distortions which may be hard to detect but which cannot even remotely be approximated as tightly wrapped waves. And that is probably the most important hint contained in this paper.

Once again, I am indebted to Drs. C. C. Lin and A. J. Kalnajs, not only for their expert advice on this paper but also for many vigorous discussions of the spiral problem in general. I would also like to thank Drs. G. Contopoulos, C. Hunter, F. D. Kahn, N. R. Lebovitz, M. J. Lighthill, K. H. Prendergast, and F. H. Shu for various helpful and perceptive comments.

This work was supported in part by grants from the National Science Foundation and from the National Aeronautics and Space Administration.

REFERENCES

Bretherton, F. P., and Garrett, C. J. R. 1968, *Proc. Roy. Soc. London, A*, **302**, 529.
Goldreich, P., and Lynden-Bell, D. 1965, *M.N.R.A.S.*, **130**, 125.
Habing, H. J. 1967, in *Proc. I.A.U. Symp. No. 31*, p. 159.
Hunter, C., and Toomre, A. 1969, *Ap. J.*, **155**, 747.
Julian, W. H. 1967, *Ap. J.*, **148**, 175.
Julian, W. H., and Toomre, A. 1966, *Ap. J.*, **146**, 810 ("JT").
Kalnajs, A. J. 1965, unpublished Ph.D. thesis, Harvard University.
Lighthill, M. J. 1965, *J. Inst. Math. Appl.*, **1**, 1.
Lin, C. C., and Shu, F. H. 1964, *Ap. J.*, **140**, 646.
———. 1966, *Proc. Nat. Acad. Sci.*, **55**, 229.
———. 1967, in *Proc. I.A.U. Symp. No. 31*, p. 313.
Lin, C. C., Yuan, C., and Shu, F. H. 1969, *Ap. J.*, **155**, 721.
Lindblad, B. 1959, *Hdb. d. Phys.*, **53**, 21.
———. 1963, *Stockholm Obs. Ann.*, Vol. **22**, No. 5.
Lindblad, P. O. 1960, *Stockholm Obs. Ann.*, Vol. **21**, No. 3.
Lynden-Bell, D., and Ostriker, J. 1967, *M.N.R.A.S.*, **136**, 293.
Pfleiderer, J. 1963, *Zs. f. Ap.*, **58**, 12.
Pfleiderer, J., and Siedentopf, H. 1961, *Zs. f. Ap.*, **51**, 201.
Schmidt, M. 1965, in *Galactic Structure*, ed A. Blaauw and M. Schmidt (Chicago: University of Chicago Press), p. 513.
Shu, F. H. 1968, unpublished Ph.D thesis, Harvard University.
Szebehely, V. 1967, *Theory of Orbits: The Restricted Problem of Three Bodies* (New York: Academic Press).
Toomre, A. 1964, *Ap. J*, **139**, 1217.
——— 1970 (in preparation).
Whitham, G. B. 1970, *J. Fluid Mech.*, **9**, 347.
———. 1965, *ibid.*, **22**, 273.