THE ASTROPHYSICAL JOURNAL, 457:125–144, 1996 January 20

SECULAR EVOLUTION OF SPIRAL GALAXIES. I. A COLLECTIVE DISSIPATION PROCESS

XIAOLEI ZHANG
Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Mail Stop 78, Cambridge, MA 02138
Received 1994 March 14; accepted 1995 July 26

ABSTRACT

A collective dissipation mechanism responsible for the secular evolution of the disks of spiral galaxies is proposed and analyzed. The key element in this process is the outward transport of angular momentum. Although it has been previously shown by Lynden-Bell & Kalnajs (1972) that a trailing spiral pattern transports the angular momentum outward, it has also been claimed by them that the exchange of angular momentum between the disk stars and the spiral density wave happens only at the wave-particle resonances. This implies that for the majority of the disk stars there is no secular orbital decay or increase, and, as a result, there is little redistribution of disk surface density over the lifetime of a spiral galaxy. In this paper, we demonstrate that such a conclusion results from the fact that Lynden-Bell & Kalnajs had solved the problem locally and considered only the orbital *response* of stars to an *applied* spiral potential. They did not incorporate the constraint for a self-sustained global spiral solution. It is shown that this constraint is in the form of a phase shift, which exists between a self-consistent, open spiral potential and density pair. A phase shift between the potential and density spirals indicates that there is a torque applied by the spiral potential on the spiral density, and a secular transfer of energy and angular momentum between the disk matter and the spiral density wave. For the actual density distribution of a spiral wave mode, it is shown that the sense of this phase shift is such that for a trailing spiral, the disk matter inside the corotation radius should lose energy and angular momentum to the density wave and accrete inward, and the matter outside corotation should gain energy and angular momentum from the wave and excrete. As a result, the disk surface density should become more and more centrally concentrated, together with the buildup of an extended outer envelope. This trend is consistent with the direction of entropy evolution in self-gravitating systems (Antonov 1962; Lynden-Bell & Wood 1968) and is also consistent with the trend found in the recent *N*-body simulations of stellar disks (Donner & Thomasson 1994; see also the simulation results in this paper). It is further demonstrated that a local physical mechanism can be found to account for the secular dissipation as is revealed and required by the phase shift. This mechanism takes the form of a temporary local gravitational instability of the streaming disk material at the spiral arms. The presence of this instability, coupled with the fact that a phase shift appears to cause a finite amplitude, open spiral wave to steepen until there is sufficient dissipation in the spiral instability to offset the steepening tendency, indicate that the nature of the large-scale spiral density waves are large-scale spiral gravitational shocks. The typical width of the spiral gravitational shock is on the order of 1 kpc, the same as the effective mean free path of stars in the spiral-arm local gravitational instability. As a result of the instability condition at the spiral arms, a single disk star, when crossing the spiral arms, experiences many small-angle scatterings produced by the combined potential of its neighboring stars, besides experiencing the smooth axisymmetric potential and the smooth spiral potential. The is leads to a secular decay in the mean orbital radius for those stars inside corotation, as well as a secular increase in the mean orbital radius for stars outside corotation.

Subject headings: galaxies: evolution — galaxies: kinematics and dynamics — galaxies: spiral — galaxies: structure — Galaxy: structure — waves

1. INTRODUCTION

It is well known that for systems governed by long-range forces such as gravity or electromagnetic forces, the secular evolution of these systems is often determined chiefly by collective processes. In particular, in nonequilibrium systems where the microscopic means of evolution become too slow, these systems often first build a tool, i.e., create a global structure by going through a nonequilibrium phase transition, and then make use of the collective properties of such structures to accelerate the speed of reaching equilibrium. In this paper, as well as in the two subsequent papers in this series (Zhang 1995b, c, hereafter Papers II and III), we seek to establish that the spiral structure in disk galaxies is, in fact, another example of such a tool built by nature employing the long-range property of gravitational interactions. It is shown that the emergence of spiral structure greatly accelerates the speed of the evolution of a disk galaxy toward higher entropy configurations, i.e., those with a centrally concentrated core, together with an extended and diffused envelope.

The evidence that there is a secular dissipative process operating in the stellar disks of spiral galaxies has already emerged in the early and recent *N*-body simulations (Sellwood & Carlberg 1984; Carlberg 1986; Donner & Thomasson 1994). In these simulations, it is found that there is significant redistribution of disk matter, which amounts to halving the disk exponential length scale in a Hubble time (Carlberg 1986). For the quasi-stationary spiral mode obtained by Donner & Thomasson (1994), it is found that on average the disk matter inside corotation accretes to the center, and the matter outside the corotation excretes, so a more centrally condensed configuration gradually develops, together with a diffused outer envelope.

The phenomenon of radial mass accretion and excretion observed in these N-body simulations, which include no gas component and its associated dissipation effect, has no explanation within the framework of the existing density wave theory. For the case of a quasi-stationary spiral, a single star's orbit conserves the Jacobi integral in the corotating frame. The stellar trajectory oscillates between two limiting radii, and no secular orbital increase or decrease is possible. Even in the transient spiral case, as we will show later in this paper, for a single star moving in an applied spiral potential, only heating in the form of increased epicycle amplitude is possible; again, no secular mean-radius change is ever observed. These facts lead us to speculate that whatever mechanism causes the secular redistribution of disk matter observed in the N-body simulations has to be related to the collective nature of the self-sustained global spiral instability and has to involve the graininess of many neighboring stars, since a single orbit in an applied spiral potential never displays secular radial migration.

There is also a second hint given by the result of the N-body experiments. The spiral patterns which spontaneously emerge from the amplification of noise in these N-body simulations are predominantly of the trailing type, if a realistic basic state which has decreasing angular velocity with increasing galactic radii is used. This is true even for the quasi-stationary spiral mode obtained by Donner & Thomasson (1994), which lasted for more than five revolutions. From the "antispiral" theorem of Lynden-Bell & Ostriker (1967), we know that, unless there is dissipation in the underlying basic state, a quasi-stationary trailing spiral mode cannot be obtained from the set of dynamical equations which is manifestly reversible. So it is only logical to infer that there is some form of dissipation present in these N-body experiments. "But stars are nondissipative!" This is an objection the author has often heard. Indeed, one of the goals of the current paper is to discredit the myth that dissipation in a galactic disk can occur only through the mediation of interstellar gas. It is true that stars cannot convert a significant amount of their orbital energy into radiation. However, in terms of their ability for the irreversible conversion of regular orbital motion energy into random motion (epicycle) energy, which is the sense in which we use the word "dissipation," stars can indeed be considered dissipative. The stellar dissipation is mediated by the spiral density wave and is achieved through a series of small-angle scatterings when a star crosses the spiral arm. The fact that stars can scatter off their neighboring stars, despite the large mean free path they have, is due to the presence of a temporary local gravitational instability at the location of spiral arms, as we will demonstrate later. The dissipated orbital energy is partly used to heat the disk locally and partly carried to the outer disk by the trailing spiral density wave to be absorbed there. So the outer disk becomes a partial sink to the dissipated orbital energy.

The organization of the current paper is as follows. In § 2, as well as in the associated appendices, we introduce the concept of the phase shift between a self-consistent spiral potential and density pair, as well as the secular dissipation effect it indicates. In § 3 we discuss how the dissipation effect indicated by the phase shift is achieved, through analyzing the local stability condition at the spiral arms, and through establishing that a self-sustained global spiral pattern is a propagating front of local gravitational instability and gravitational shocks. Section 4 compares the results of the current paper with previous results in the literature and also outlines the astrophysical consequences of the spiral collective dissipation process, the details of which will be further analyzed in Papers II and III.

2. THE PRESENCE OF A PHASE SHIFT BETWEEN A SELF-CONSISTENT SPIRAL POTENTIAL AND DENSITY PAIR

The secular morphological evolution of a disk galaxy is ultimately gauged by the redistribution of angular momentum. From a comparison with other types of collective phenomena, we expect that the secular evolution of spiral galaxies is mediated chiefly by the spiral structure. Lynden-Bell & Kalnajs (1972, hereafter LBK) have already shown that, in general, a trailing spiral structure transports angular momentum outward. However, the proper means for loading the angular momentum of the disk stars onto the wave is yet to be found. In order for the disk stars to exchange angular momentum with the wave at galactic radii other than the resonances, we ask what kind of conditions must a spiral structure satisfy? In the following we will show that, for a spiral structure that is quasi-stationary on the dynamical timescale, the *only* possible way for the stars and the wave to exchange angular momentum secularly is to have a spiral density distribution which is phase-shifted in azimuth with respect to the spiral potential distribution.

Consider a disk galaxy with a total potential distribution of $\mathscr{V}$ and a total density distribution of Σ, each of which contains an axisymmetric part and a perturbation of the spiral form. For an annular ring located at radius r with width dr, the (z-component) torque applied by the total potential field on the material in this annular ring is

$$T(r) = r\,dr \int_0^{2\pi} -\Sigma(\boldsymbol{r} \times \nabla\mathscr{V})_z\, d\phi$$

$$= r\,dr \int_0^{2\pi} -\Sigma\left(\frac{\partial\mathscr{V}}{\partial\phi}\right) d\phi\ , \qquad (1)$$

where we have used $\boldsymbol{r} = r\hat{\boldsymbol{r}} + z\hat{\boldsymbol{z}}$.

Equation (1) can also be written in the form

$$T(r) = r\,dr \int_0^{2\pi} -\Sigma(r, \phi)\,\frac{\partial\mathscr{V}_1(r, \phi)}{\partial\phi}\, d\phi\ , \qquad (2)$$

where the subscript 1 on the potential denotes the spiral perturbation component, since the axisymmetric component of the potential gives a zero ϕ derivative. Manifestly, equation (2) describes the torque applied by the *spiral* part of the potential on the *total* disk surface density. Therefore it also gives the amount of the angular momentum transport from the spiral wave to the disk matter in this annular ring per unit time.

Equation (2) can be further written in the form of

$$T(r) = r\,dr \int_0^{2\pi} -\Sigma_1(r, \phi)\,\frac{\partial\mathscr{V}_1(r, \phi)}{\partial\phi}\, d\phi\ , \qquad (3)$$

with the subscript 1 on both the potential and density variables. This is because the axisymmetric component of the density integrates to a null value in equation (2) as long as the perturbation potential is periodic in ϕ. Equation (3), however, should still be regarded as the torque applied by the spiral wave on the *total* disk matter in the annular ring.

Dividing the above expression by $2\pi r\,dr$, the area of the annular ring, we obtain that the azimuth-averaged torque density $\bar{\mathscr{T}}$, which is equal to the averaged rate of angular momentum flow per unit area from the density wave to the disk material, at a particular radius r is

$$\bar{\mathscr{T}}(r) = \overline{\frac{dL}{dt}}(r) = -\frac{1}{2\pi}\int_0^{2\pi} \Sigma_1(r, \phi)\,\frac{\partial \mathscr{V}_1(r, \phi)}{\partial \phi}\,d\phi\ . \qquad (4)$$

A similar expression was first written down by Kalnajs (1972) in analyzing the angular momentum exchange between the stellar and gaseous density waves.[1]

The torque integral in equation (4) vanishes for potential and density profiles which have identical waveforms, even if they are nonsinusoidal. This can be demonstrated by expanding these waveforms into their Fourier components, which would have the property that the Fourier coefficients of the same harmonics for the potential and density are proportional to each other. Using trigonometric identities, we can easily show that the torque integral in equation (4) for such a pair of waveforms vanishes. Therefore the noncoincidence of the potential and density spirals in the form of an equivalent phase shift[2] is the *only* means for secular angular momentum transfer between the disk matieral and a quasi-stationary spiral density wave. This is what gives importance to the study of the phase shift.

In the Appendices, we show that a phase shift can be found for the potential and density spirals related through the Poisson equation, in the solution of the linearized Eulerian equations of motion, as well as in the orientation of the linear periodic orbit. The Poisson equation gives a constant phase shift with radius for a spiral pattern which is infinite in radial extent. This constant phase shift is positive (which means that the density leads the potential) for spirals with radial density falloff slower than $r^{-3/2}$ and is negative for faster falloff. The sign of the phase shift obtained from the Eulerian equations of motion and from the linear periodic orbit solution is such that the spiral potential lags the spiral density inside corotation, and vice versa outside corotation. This shows that, in principle, a self-consistent spiral wave solution can be constructed with a phase shift between the potential and density spirals. our N-body simulation in the next section confirms this point.

Before closing this section, we want to comment briefly on the meaning of the phrase "energy and angular momentum exchange between the disk material and the density wave." Another question the author has frequently been asked is: "How could stars give energy and angular momentum to the wave (or vice versa)? Isn't the wave itself made of stars?" Yes, a self-sustained spiral wave is supported by the motion of many individual stars. However, a wave indeed has a separate existence other than the straightforward superposition of the individual stellar orbits. In the potential field of a relatively large amplitude spiral wave, it can be shown (Zhang 1995d) that the orbits themselves become chaotic. After each crossing of the spiral arm, a typical star loses all information (memory) of its previous orbital phase. Thus the information about the wave is not stored merely in the individual star's orbital orientation, as appears to be the case in Kalnajs's kinematic spiral representation (Kalnajs 1973), which is accurate only for infinitesimal wave amplitude (Kalnajs himself had, in fact, already stressed this point in his 1973 paper). Rather, the information about the wave is stored in the collective force field (or potential field) of the wave, which "collapses the chaos" inherited in the individual star's orbital motion. The local field in the spiral pattern is effectively contributed by all the stars in the disk, due to the long-range nature of gravitational interaction. Thus, when we say that a star exchanges energy and angular momentum with the wave, this exchange occurs simultaneously with all the rest of the stars in the disk which contribute to the wave motion. This exchange is not random but is organized by the wave. So, for practical purposes, it is infinitely more convenient to talk about the energy and angular momentum exchange between the wave and the stars than to talk about the energy and angular momentum exchange of one star with the rest of the stars in the disk in the manner constrained by the wave field.

In the next section, we present a detailed analysis of the collective dissipation process induced by a spiral density wave. This collective dissipation process, on the one hand, is made possible by the presence of the phase shift and, on the other hand, is also responsible for the continued maintenance of the phase shift between a self-consistent spiral potential and density pair.

3. A COLLECTIVE DISSIPATION PROCESS INDUCED BY A SELF-SUSTAINED SPIRAL WAVE

It is well known that the relaxation and evolution of systems governed by long-range forces are determined predominantly by collective effects. The physical behavior of such systems is not completely reflected in the result of a single orbit calculation (see, e.g., Pfenniger 1986). For example, although a single orbit (circular or epicyclic) is known to be stable in the potential field of an axisymmetric galaxy, global axisymmetric and bisymmetric instabilities can still grow spontaneously from such disks (Toomre 1964; Lin & Shu 1964). On the other hand, certain systems consisting of irregular orbits, such as a common gas ensemble, are known to be structurally stable.

It is also well known that collective effects can enhance the speed of relaxation in many plasma systems (Kulsrud 1972). Collective effects, which invariably involve local or global instabilities, operate by forcing a particle to "collide with" or scatter off a bunch of other particles collected together by the wave. The effective impact parameter in such a "collision" process is of the order of the size of the inhomogeneity formed (Kulsrud 1972). In effect, besides experiencing the smooth part of the potential, a particle which participates in the collective process also bounces off the local (short-to-intermediate range) scattering potential, with the "graininess" of the neighboring particles coming into full play.

Following is an incomplete list of references on collective effects in the context of galactic dynamics: Gilbert (1968), Kulsrud (1972), Gurzadyan & Savvidy (1986), Pfenniger (1986), Kandrup (1988), Romeo (1990), and Weinberg (1993).

3.1. *Local Gravitational Instability at the Spiral Arms*

For collective effects to operate in a spiral galaxy, individual stars have to be aware of their "neighbors" directly, besides

[1] Note that dL/dt in eq. (4) refers to the rate of angular momentum *flow* per unit area from the density wave to the disk matter, due to the operation of spiral gravitational torque. To calculate the net angular momentum gain for the material in a particular annulus we need also to consider the flux through the boundaries of the annulus. This, as well as other issues related to the balancing of the angular momentum budget for a quasi-stationary spiral mode, will be further analyzed in paper II. We only comment briefly here that since the process responsible for the angular momentum transfer given by eq. (4) is irreversible in nature, as will be shown in § 3, the amount of angular momentum transfer from the wave to the disk matter (or vice versa) due to gravitational torques is independent of the amount of the flux through the boundaries of the annulus.

[2] The definition of the equivalent phase shift for nonsinusoidal waveforms will be given in Paper II.

experiencing the smoothed axisymmetric as well as the smoothed spiral potential. However, as is well known, binary encounters in disk galaxies are extremely rare compared to the age of a galaxy (Binney & Tremaine 1987, p. 4). Under this circumstance, the scattering of a star off its neighboring stars can only happen when the disk is locally gravitationally unstable. So the first step in establishing that a spiral structure can induce collective dissipation is to show that a spiral structure can lead to local gravitational instability (albeit a temporary one) in an originally marginally stable disk.

3.1.1. *Local Stability Condition at the Spiral Arm and Interarm Region*

The local stability condition for a flattened stellar disk against axisymmetric perturbations is given by (Toomre 1964)

$$Q = \frac{\sigma_r \kappa}{3.36 G\Sigma} > 1 \,, \tag{5}$$

where σ_r is the radial velocity dispersion, κ is the epicycle frequency, Σ is the surface density of the disk, and Q is Toomre's stability parameter. For a fluid disk, the factor 3.36 in the denominator is changed to π. An order-of-magnitude estimate (Binney & Tremaine 1987, p. 313) shows that equation (5) can also serve as the approximate stability criterion for the formation of more localized instability features in a rotating disk, if the compression and expansion are mainly in the radial direction.

At the different azimuthal locations, the streaming motion of the disk material, under the influence of a spiral perturbation potential, changes the values of the radial velocity dispersion σ_r, the epicycle frequency κ, and the surface density Σ from their original values appropriate for an axisymmetric disk. In the following, we will derive the variations of these parameters with the phase of the spiral, and calculate how these variations influence the value of Q at the spiral arm and interarm region. We will first consider a linear and WKBJ (i.e., tightly wrapped) spiral wave, and then discuss what modifications we need to introduce when considering a more open type of wave in the nonlinear regime. Part of the results for the WKBJ waves (the variation of κ with the spiral phase) has been previously derived by Balbus & Cowie (1985) using a different approach.

For an m-armed spiral, the gravitational potential can be written as (Rohlfs 1977)

$$\mathcal{V}(r, \phi) = \mathcal{V}_0(r) + A(r) \exp\{i[m\Omega_p t - m\phi + \Phi(r)]\} \,, \tag{6}$$

where $|A| \ll |\mathcal{V}_0|$, and where $\Phi(r)$ is related to the pitch angle i and the wavenumber k of the spiral through

$$\frac{d\Phi}{dr} = \frac{m}{r \tan i} = k \,, \tag{7}$$

with $k < 0$ corresponding to a trailing spiral. The WKBJ approximation further demands that

$$|kr| \gg 1 \,. \tag{8}$$

The solution for the azimuthal velocity is

$$v(r, \phi) = v_c(r) + i \frac{kA}{2\Omega_0} \frac{1}{1 - v^2 + x} \times \exp\{i[m\Omega_p t - m\phi + \Phi(r)]\} \,, \tag{9}$$

where v is the normalized interaction frequency for mode m, $v = m(\Omega_p - \Omega)/\kappa$, $x = k^2\sigma_{r0}^2/\kappa_0^2$, and where v_c, σ_{r0}, Ω_0, and κ_0 are the unperturbed circular velocity, radial velocity dispersion, angular frequency, and epicyclic frequency, respectively. For convenience, in the following discussions we assume a flat rotation–curved galaxy, i.e., $v_c(r) = v_c$ is a constant.

The corresponding density variation is

$$\Sigma(r, \phi) = \Sigma_0(r) - \Sigma_0(r) \frac{k^2 A}{\kappa_0^2} \frac{1}{1 - v^2 + x} \times \exp\{i[m\Omega_p t - m\phi + \Phi(r)]\} \,. \tag{10}$$

Since

$$\kappa^2 = 2r\Omega \frac{d\Omega}{dr} + 4\Omega^2 \,, \tag{11}$$

in the following we first calculate the change in Ω and $d\Omega/dr$ due to the presence of a spiral.

The angular frequency at a location (r, ϕ) in the presence of spiral perturbation becomes

$$\begin{aligned}\Omega(r, \phi) &= \frac{v(r, \phi)}{r} \\ &= \Omega_0(r) + i \frac{kA}{2\Omega_0 r} \frac{1}{1 - v^2 + x} \\ &\quad \times \exp\{i[m\Omega_p t - m\phi + \Phi(r)]\} \,.\end{aligned} \tag{12}$$

Therefore,

$$\frac{d\Omega}{dr}(r, \phi) = -\frac{v_c^2}{r^2} - \frac{k^2 A}{2\Omega_0 r} \frac{1}{1 - v^2 + x} \times \exp\{i[m\Omega_p t - m\phi + \Phi(r)]\} \,. \tag{13}$$

The effective κ^2 in the presence of the spiral potential can thus be calculated to be

$$\kappa^2 = \kappa_0^2 \left[\!\!\left[1 - \frac{k^2 A}{\kappa_0^2} \frac{1}{1 - v^2 + x} \exp\{i[m\Omega_p t - m\phi + \Phi(r)]\} \right]\!\!\right] \,, \tag{14}$$

where $\kappa_0^2 = 2\Omega_0^2$ for a flat rotation–curved galaxy, and where we have dropped a few terms of order $1/kr$ or higher compared to the dominant terms in deriving equation (14).

On the other hand, from equation (10) we know that

$$\frac{\Sigma}{\Sigma_0} = 1 - \frac{k^2 A}{\kappa_0^2} \frac{1}{1 - v^2 + x} \exp\{i[m\Omega_p t - m\phi + \Phi(r)]\} \,. \tag{15}$$

We have thus demonstrated that

$$\frac{\kappa}{\kappa_0} = \left(\frac{\Sigma}{\Sigma_0}\right)^{0.5} \,. \tag{16}$$

The change in the velocity dispersion σ_r depends on the energy conversion process assumed. If, as in the linear density wave theory, we assume an adiabatic process for stars entering and leaving the spiral arms, so that the self-gravitating potential energy of the streaming stars is temporarily converted into the random velocities of these stars, we expect that

$$\frac{\sigma_r^2}{\sigma_{r0}^2} \approx \frac{\Sigma}{\Sigma_0} \,, \tag{17}$$

where we have assumed that the compression is one-dimensional, as is appropriate for WKBJ waves.

From equations (5), (17), and (16), we see that at the location of the spiral arm, which corresponds to the location of the density enhancement, the epicycle frequency κ increases so that the stabilizing effect of the Coriolis force is increased. The velocity dispersion increases too by an amount determined by the degree of density enhancement, and also by the overall energy conversion process at the spiral arms. The final stability state at the spiral arm region will be determined by the competition of these different factors.

In the case when only orbit crowding but no energy loss occurs, the effective Q will not change significantly from its unperturbed value for the basic state of the disk, i.e., we have $Q_{\mathrm{arm}} \approx Q_{\mathrm{interarm}} \approx Q_0$ for a linear WKBJ wave, due to the fact that both κ and σ_r in equation (5) scale as $\Sigma^{0.5}$.

However, for a more open spiral pattern with finite amplitude, the potential field at the spiral arm is generated not only by the local streaming mass but also by the matter in the rest of the spiral pattern. The phase shift between the potential and density spirals of an open spiral pattern (it can be easily shown that the phase shift for the tightly wound WKBJ wave is zero) means that a patch of disk material, when entering the spiral arm, experiences an extra compression contributed by the rest of the disk matter, besides that contributed by its own self-gravity. Thus the originally marginal stable disk material becomes temporarily unstable when crossing the spiral arms. structure on the originally marginal stable axisymmetric disk. The local instability at the spiral arms is the constituent of the global spiral instability.

In § 3.2, we will use the result of an N-body simulation to show that a significant variation of Q from the spiral arm to the interarm region can indeed be observed.

3.1.2. *Derivation of the Length Scale of the Local Gravitational Instability at the Solar Neighborhood*

At the solar neighborhood ~10%–20% of the local surface density is contributed by the gas component, which has random velocities between 4 and 8 km s^{-1} (Spitzer 1978, p. 231; Stark 1979; Liszt & Burton 1981). The presence of the low velocity dispersion gaseous component generally makes the galactic disk less stable. The quantitative effect of the gas on the stability condition of the star/gas combined disk can be analyzed through the two-fluid dispersion relation of Jog & Solomon (1984).[3] The distinctive features of the two-fluid results as compared to the one-fluid ones are, first, in the case where the stars and gas are separately stable, the combined two-fluid disk can be unstable; second, the length scale of the most unstable instability feature is significantly reduced from that of the pure stellar case.

Including the finite–disk-thickness correction, the two-fluid dispersion relation is in the form (Jog & Solomon 1984)

$$\omega^2 = \tfrac{1}{2}\{(\alpha_s' + \alpha_g') - [(\alpha_s' + \alpha_g')^2 - 4(\alpha_s'\alpha_g' - \beta_s'\beta_g')]^{1/2}\} , \quad (18)$$

where

$$\alpha_s' = \kappa^2 + k^2 v_s^2 - 2\pi G k \Sigma_s\{[1 - \exp(-kh_s)]/kh_s\} , \quad (19)$$

$$\alpha_g' = \kappa^2 + k^2 v_g^2 - 2\pi G k \Sigma_g\{[1 - \exp(-kh_g)]/kh_g\} , \quad (20)$$

$$\beta_s' = 2\pi G k \Sigma_s\{[1 - \exp(-kh_s)]/kh_s\} , \quad (21)$$

$$\beta_g' = 2\pi G k \Sigma_g\{[1 - \exp(-kh_g)]/kh_g\} , \quad (22)$$

where ω is the angular frequency of the two-fluid instability, h_s and h_g are the scale heights, Σ_s and Σ_g are the surface densities, and v_s and v_g are the velocity dispersions of the stellar and gaseous fluids, respectively, and $k = 2\pi/\lambda$ is the wavenumber of the axisymmetric two-fluid instability under consideration.

Making use of the two-fluid dispersion relation, we plot in Figure 1 the ω^2 versus $1/\lambda$ curve for parameters appropriate for the solar circle. Since there is expected to be extra compression on the streaming disk matter due to the fact that the potential and density are phase-shifted with respect to each other, we have adopted the highest known total surface density at the solar neighborhood, $\Sigma_{\mathrm{total}}(r = r_\odot) = 80\ M_\odot\ \mathrm{pc}^{-2}$ (Bahcall 1984) in the calculation of the instability length scale. From Figure 1, we see that the solar neighborhood is slightly below the stability threshold, and the most unstable mode (the one with the largest magnitude of negative ω^2) has wavelength

$$\lambda_{\text{most unstable}} \approx 3.6\ \mathrm{kpc}\ . \quad (23)$$

The instability structure formed usually has an extent of $\lambda/2$ for the region of density enhancement, thus a radius of $\lambda/4$. Therefore, the magnitude of the radus of the instability structure is

$$r_c \sim 0.9\ \mathrm{kpc}\ . \quad (24)$$

Compared to the galactic orbital circumference at the solar neighborhood, which is ~50 kpc, the size of the instability structure can fit comfortably in the circumferential direction. This length scale is also very close to that of the size of the giant H I and molecular cloud complexes observed near the spiral arm region of many external galaxies (Elmegreen & Elmegreen 1983). It is likely that the appearance that these giant molecular cloud (GMC) complexes are gravitationally bound is an indication that the underlying stellar disk is also unstable, since, on the scale of a kiloparsec, the gaseous material is not likely to be decoupled from the stars in its stability state.

3.2. *Spiral Gravitational Shocks*

The presence of a local gravitational instability at the spiral arms indicates that the streaming disk material experiences

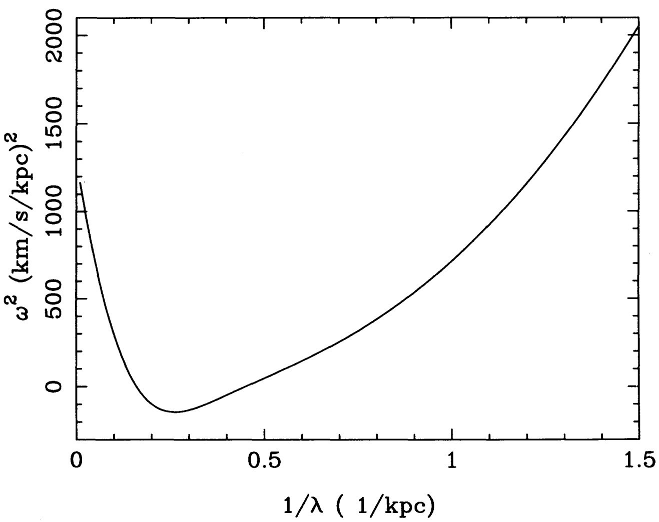

FIG. 1.—Growth rate of the two-fluid instability at the solar neighborhood. Parameters used: $\Sigma_{\mathrm{total}} = 80\ M_\odot$ (Bahcall 1984) (15% of the surface density is in the gas phase), $v_s = 30$ km s^{-1} (Binney & Tremaine 1987), $v_g \approx 6$ km s^{-1} (Liszt & Burton 1981; Stark 1979; van der Kruit & Shostak 1984), $h_s = 175$ pc (Gilmore et al. 1990), $h_g = 75$ pc (Jog & Solomon 1984), $\kappa = 36$ km s^{-1} kpc^{-1} (Binney & Tremaine 1987).

[3] An earlier two-component treatment was given in Lin & Shu (1970), which also incorporated the finite–disk-thickness effect.

something akin to true collisions[4] when it crosses the spiral arm. Therefore, the analogy of the stellar random velocity to the sound speed of a gas is much better established at the location of spiral arms. As is well known, over most of the galactic radii the entry speed of the disk material into the spiral arms is supersonic; and for spiral forcing strength greater than $\sim 3\%$, the periodic orbits of stars at neighboring galactic radii are found to intersect (Wielen 1975). Furthermore, the presence of the phase shift between the potential and density distributions means that the local state of the disk matter and the wave has to be in one of the following situations. First, if the wave pattern has reached a quasi-steady state, then the phase shift indicates a secular dissipation process as given by the torque integral in equation (4). Second, if there is insufficient dissipation to acheive a quasi-steady state, then the wave has to change in shape. A natural way for the wave to change shape is for it to steepen into a shocklike profile, so that shock dissipation can relieve some of the "stresses" applied or required by the phase shift. These factors, together with the analogy to the steepening of nonlinear acoustic waves into shock waves (Appendix C), point to the formation of galactic scale spiral shocks in the stellar medium.

The calculation of galactic spiral shocks, originally thought to exist only in the gaseous component, began with the work of Fujimoto (1968). The possibility of quasi-stationary spiral shocks of galactic scale was first demonstrated by Roberts (1969) within the framework of WKBJ theory, and with the self-gravity of the gas ignored. Shu, Milione, & Roberts (1973) demonstrated that, at least in the non–self-consistent case (i.e., with the self-gravity of the gas ignored from the forcing spiral potential), the equilibrium flow solution always contains shocks as long as the forcing is more than a few percent. A finite-amplitude spiral potential modifies the velocity field of the galactic flow from that of entirely supersonic (or entirely subsonic) to that which contains sonic transitions, and a shock forms near the location where the supersonic flow velocity changes to subsonic velocity in the form of a sudden jump. The time-dependent calculation of the formation and steepening of spiral shocks was first carried out by Woodward (1973, 1975), again ignoring the self-gravity of the gas and again using the WKBJ approximation. Recent work on the nonlinear development of spiral structures, although still employing the WKBJ approximation, has incorporated the self-gravity of the stars (Shu, Yuan, & Lissauer 1985) and gas (Lubow, Balbus, & Cowie 1986; Lubow 1988). It is found that the self-gravity of the disk material generally makes the density peaks formed near the spiral potential minimum more symmetric than that of the forced hydrodynamic shock (Shu et al. 1985; Lubow et al. 1986; Lubow 1988). It is also found that for fully self-gravitating nonlinear WKBJ waves, the streamlines of the fluid never cross one another (Shu et al. 1985), so presumably no dissipative shocks can form. This result is also hinted in the earlier nonlinear analysis of the self-consistent WKBJ waves by Vandervoort (1971).[5]

From the above mentioned results, and also from the arguments given in Appendix C, we expect that shock formation could be a general characteristic of the nonlinear development of the galactic spiral waves even in stellar disks, as long as the wave is somewhat open in morphology. The presence of a phase shift between the potential and density of an open spiral pattern indicates that the streaming matter is, locally, only partially self-gravitating before entering the spiral arm, and this provides the possibility for the material to "unexpectedly" shock onto the potential wave as it crosses the spiral potential minimum.

The most appropriate or satisfying way to demonstrate the steepening of nonlinear spiral wave modes into spiral shocks is by the iterative solution of the set of nonlinear Eulerian fluid equations, as well as the equation of continuity and the Poisson equation, starting from a known linear spiral modal distribution for a given basic state, together with the proper inner and outer boundary treatment. This approach is similar in spirit to what Lubow et al. (1986) had used for a nonlinear WKBJ wave. The numerical implementation of this two-dimensional initial–boundary-value problem seems to be a formidable task to the author at the moment. In the following, however, we will adopt another route by employing the well-developed technique of N-body calculations. There is an added advantage for adopting the N-body approach, in that the viscosity due to the graininess of the particles is naturally incorporated into the simulation, closely resembling the situation in real stellar disks.

The N-body code used for the simulation of spiral disks is a two-dimensional polar code, written by the author using the algorithms described in Thomasson (1989). The validity of the code is checked by running the first example described in Donner & Thomasson (1994), which simulates the spontaneous growth of a spiral mode in an unperturbed disk. Reasonably good agreements with the results of Donner & Thomasson (1994) were found, in terms of the morphology of the spiral mode formed, its growth rate, pattern speed, and amplitude evolution, when using the same set of simulation parameters, although small differences do exist. These small differences in the spontaneous spiral mode formed are to be expected, since these two versions of the N-body codes are implemented in slightly different ways (for example, Thomasson's code aver-

[4] In plasma physics, the term "collisionless shock" has been used to refer to the kind of shock which is associated with the particle-scattering process inside an instability front (Krall & Trivelpiece 1973), in contrast to the kind of hydrodynamic shock where true particle collisions are happening. The spiral gravitational shock we will discuss in this section is, in essence, the same as the plasma "collisionless shock." We will, however, not use the term "collisionless shock" extensively in the rest of the paper. Rather, we prefer the term "spiral gravitational shock," in order first to emphasize that the nature of the instability we are dealing with here is gravitational and is associated with the spiral structure itself, and, second, to avoid the misconception that there is actually any fundamental difference between a true collision and a scattering process—all collisions are scatterings of varying strength if we look close enough!

[5] It is reasonable outcome that a self-consistent WKBJ wave does not contain shocks. This is true for the following reasons: (1) A spiral shock invariably introduces viscous dissipation, therefore the potential is, on the average, phase-shifted with respect to the density, whereas for a self-consistent (lowest order) WKBJ wave, the potential is everywhere proportional to the density (see, e.g., Shu 1992, eq. [11.41]), so no phase shift is allowed. (2) If a self-consistent WKBJ spiral shock does exist, then its potential will also have a finite jump at the location of the shock, as has the density (Shu 1992, eq. [11.41]). This, however, indicates an infinite force, which is the derivative of the potential. This infinite force cannot be consistently accommodated by the equation of motion perpendicular to the spiral arm direction (cf. Roberts 1969, eq. [10]), since some of the terms there are nonlinear in the perturbed quantities. We note, however, that Shu et al. (1985) did not employ the conventional type of lowest order WKBJ approximation; rather, they had obtained the potential through an *integration* of the density distribution of the tightly wrapped waves. Therefore, the perturbation density they obtained allowed the cusp-shaped solution which has discontinuous derivatives. True shock solution is nonetheless absent from their self-consistent nonlinear results even with such relaxed WKBJ approximations, presumably because their treatment still did not introduce a phase shift.

ages the grid mass and grid force around the centers of the mesh bin, whereas the code used for the current paper averages these quantities around the boundary nodes of the mesh bin, even though the same mesh configuration is used in both codes). The difference in the details of the spiral mode formed could also be due to the slight differences in the random number generation part of the initial position and velocity assignments for the disk stars.

Although a basic spiral shock structure already emerged in our simulation using the same set of mesh parameters and number of particles as used by Donner & Thomasson (1994), we have decided to use a computation mesh (110 radial divisions and 128 azimuthal divisions) about twice as fine and a total number of particles (200,000) about 4 times as many as theirs in the example presented below, in order to better resolve the shock structure. A gravity softening length of 1.5 times the mesh length unit is used. The cloud-in-cell (CIC) method is employed for mass and force interpolation. Further details about the numerical algorithms used can be found in Thomasson (1989).

For the spiral mode calculated in this paper, we have chosen to use the same basic-state parameters as those used in Donner & Thomasson (1994), mainly for the purpose of not having to recalculate and represent many of the modal characteristics which are already discussed in their paper, since these take up a lot of space and are not the central issues of the current paper. These modal characteristics are nonetheless important in their own right, so interested readers are encouraged to look into their paper.

The disk surface density in this case is in the form of a modified exponential

$$\Sigma(r) = \Sigma_0(e^{-r/R} - e^{-2r/R}) , \qquad (25)$$

where R is a constant scale length and Σ_0 is a constant. An inactive bulge and a rigid halo are also being used, which are assumed to be of the regular exponential shape. The (normalized) disk mass is 0.5, halo mass is 0.4, and bulge mass is 0.1. The scale length used is 10, 5, and 1 for the disk, halo, and bulge, respectively.

In Figure 2 we plot the calculated disk morphology at six selected time steps. Note that, since our choice of time step is twice as fine as that used in Donner & Thomasson (1994), our time step 3200 corresponds to their time step of 1600, which is seen to be the time step when the spiral morphology is best organized for both sets of simulations. This indicates that the two simulations gave about the same spiral modal growth rate. The large-scale morphologies of the spiral mode in these two simulations are also very similar, although the current simula-

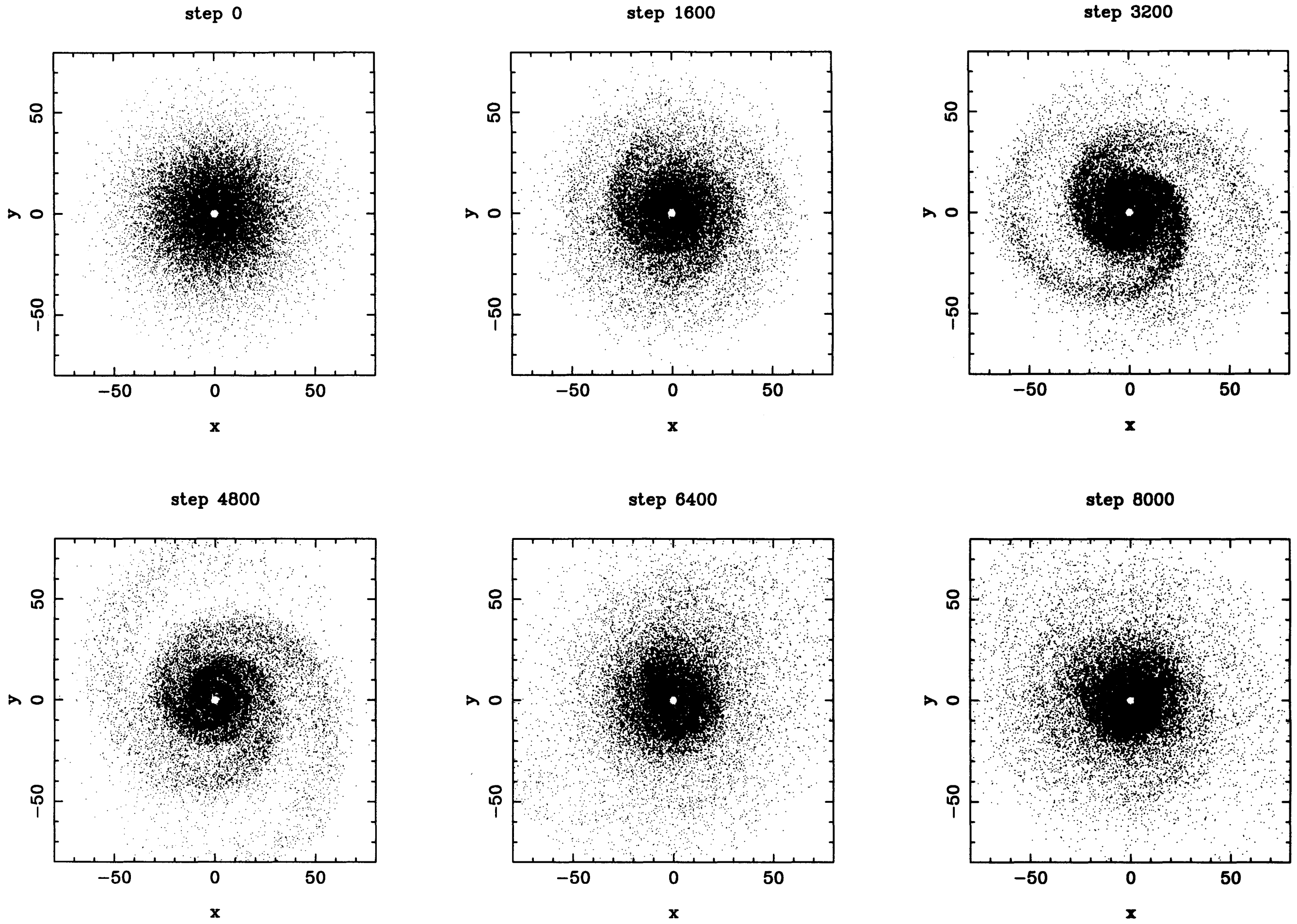

FIG. 2.—Time development of an unstable spiral mode in an unperturbed disk. Parameters for the simulation: number of grid cells in the radial direction = 110, number of grid cells in the azimuthal direction = 128, total number of particles used = 200,000. Every tenth particle used in the simulation is plotted. The time step used corresponds to 628 steps per rotation period at a radius of 20. The basic-state parameters of this disk are given in Donner & Thomasson (1994).

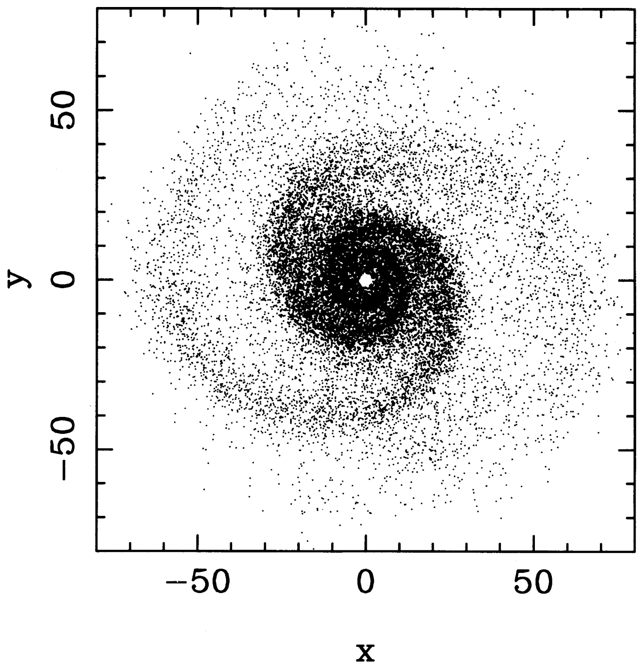

FIG. 3.—Enlarged view of the spiral mode of Fig. 2 at time step 3200

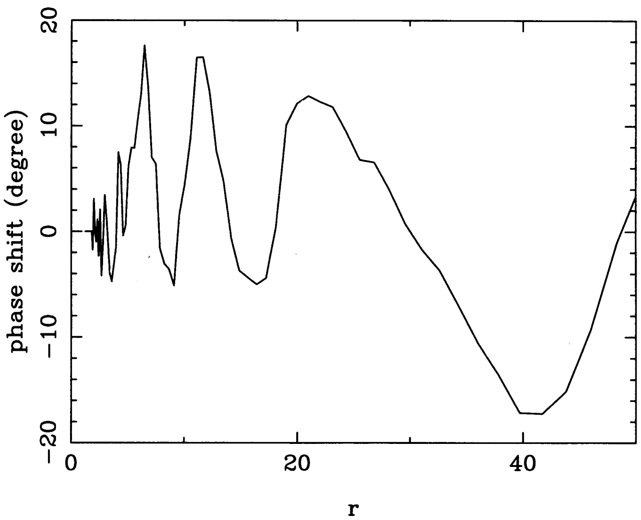

FIG. 4.—Phase-shift vs. radius plot for the spiral mode of Fig. 3 at time step 3200. For this pearticular mode, the inner Lindblad resonance is near $r = 10$, the corotation radius is near $r = 30$, and the outer Lindblad resonance is near $r = 42$ (Donner & Thomasson 1994).

tion has much better resolution to define the fine details of the mode.

In Figure 3 we plot an enlarged view of the spiral morphology at time step 3200. Figure 4 plots the calculated phase shift versus radius at this time step.[6] Here we observe that the phase shift is mostly positive in the inner disk, with oscillations which correspond to the winding of the spiral arms in Figure 3,[7] and is negative in the outer disk. It is reassuring to see that the transition between the positive and negative phase shifts happens near the corotation radius of $r = 30$. To observe this kind of coherent phase-shift distribution, the spiral model used would need to have achieved a high degree of organization throughout the disk. At time steps other than 3200, as the spiral mode loses its coherence, the phase-shift distribution also becomes less regular.

Although the global coherence of the nonlinear spiral mode degrades after time step 3200 (or 1600 in Donner & Thomasson [1994]), the underlying $m = 2$ component of this spiral mode has been shown in Donner & Thomasson (1994) to survive with almost the same amplitude and pattern speed until the end of their simulations, which is about 13 rotation periods at radius 20. This longevity of the $m = 2$ mode is also confirmed in our simulation. The dissolution of the (nonlinear) spiral modal coherence in the simulations is thought to be mainly the result of secular heating,[8] the heating effect is more pronounced for a simulation which has 10^6 times fewer particles than a real galactic disk. We will address this issue further in § 4.1.

We now focus on the main issue of our concern here, which is the formation and steepening of spiral gravitational shocks. In Figure 3, we could already discern some evidence for the presence of spiral shocks. Sharp density maxima are seen to be present at the leading edge of the spiral pattern, reminiscent of the narrow dust lanes found at the leading edges of the spiral arms of real galaxies, which are thought to represent the location of large-scale gaseous shocks. In Figures 5*a*–5*f*, we further plot the azimuthal distributions of surface density and negative potential, radial velocity dispersion, epicycle frequency κ, Toomre's Q-parameter, and the velocity components parallel and perpendicular to the spiral arms. Here the parallel and perpendicular velocity components are related to our grid velocities in the radial and azimuthal direction v_r and v_ϕ, as well as to the spiral pattern speed Ω_p and pitch angle i through (Roberts 1969)

$$v_\perp = v_r \cos i + v_\phi \sin i - \Omega_p r \sin i \qquad (26)$$

and

$$v_\parallel = -v_r \sin i + v_\phi \cos i - \Omega_p r \cos i \, . \qquad (27)$$

The pattern speed for this spiral mode is ~0.006 radians per time step (again the numerical value appears to be one-half of that obtained in Donner & Thomasson 1994 because our time step is half as fine as theirs), and the pitch angle is ~16°.8. The distributions in Figures 5*a*–5*f* are obtained by averaging the relevant characteristic for each individual bin in an annulus centered around $r = 14.5$, at time step 3800.

From Figure 5*a*, we see that the density profile is of the nonlinear shape similar to that found in the N-body simulations of gaseous spiral shocks of Levinson & Roberts (1981), with a maximum arm-interarm density contrast of 3 to 1 (for comparison, Levinson & Roberts obtained a density contrast of 2 to 1 in their gaseous simulations). Due to the self-gravity of the disk material, the spiral shock here acquires a more symmetric shape, as was also found in the previous gaseous shock simulations. The potential profile is seen to be phase-shifted

[6] The definition of equivalent phase shift for nonsinusoidal waveforms, which is used to calculate the curve in fig. 4, will be given in Paper II.

[7] These oscillations can be shown to be due to the truncation of the spiral at the outer disk.

[8] Certain basic-state and initial condition specifications are also found to lead to spiral modes which are more robust to dissolution than others. For example, the global shock pattern of another spiral mode, which we will present in Paper II, is much more long lasting than the current one.

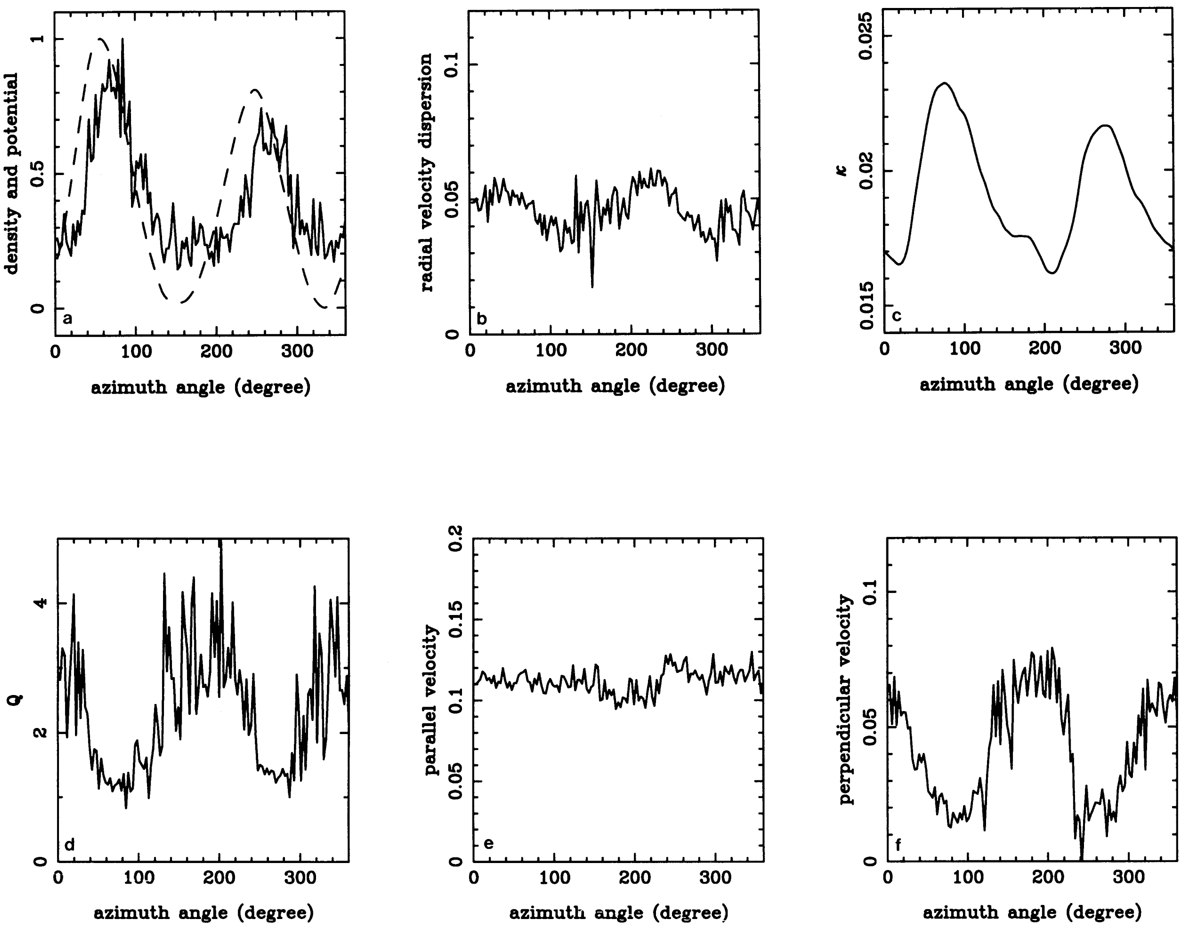

FIG. 5.—Spiral gravitational shock. Different frames show the azimuthal distributions of the following parameters. (*a*) Surface density (*solid line*) and negative potential (*dashed line*). The density is normalized to have a maximum of 1. The potential has an arbitrary scale and is shifted in the vertical direction to be displayed on the same frame as the density. (*b*) Radial velocity dispersion. (*c*) Epicyclic frequency κ. (*d*) Toomre's Q-parameter. (*e*) Velocity component parallel to the spiral arm. (*f*) Velocity component perpendicular to the spiral arm. The centers of the bins where these properties are averaged have a radius of 14.5. the time step is 3800.

from the density profile in the correct sense for a radial location inside corotation.

Figures 5*a*–5*d* give the same type of variations of Σ, σ_r, κ, and Q versus the spiral phase as we have predicted in § 3.1, except that the radial velocity dispersion σ_r starts to decrease not long after we enter the negative potential region (or positive half-cycle of the potential curve on this plot). The spiral phase where σ_r starts to decrease coincides with the phase where $v_\perp$ suffers a sharp downward jump (Fig. 5*f*), with the value of $v_\perp$ going from supersonic (note that the equivalent sound speed in this case is around 0.04, as is indicated by Fig. 5*b*) before the jump to subsonic after the jump. This clearly indicates the presence of a shock.[9] It has been checked (not plotted) that the shock actually caused the radial velocity component v_r itself to change sign near the location of the shock, which is partially responsible for the minimum of σ_r observed there.[10] Thus we see that because of shock dissipation, the velocity dispersion σ_r at the spiral arm region is smaller than that given by the linear WKBJ theory. This causes Toomre's Q-parameter to suffer a drastic decrease in the spiral arm region, to a value very close to (and sometimes smaller than) 1. This confirms our prediction in § 3.1 that there should be a temporary local gravitational instability at the spiral arms. Since the presence of this instability is a result of the nonlocal (long-range) nature of the gravitational interaction and is brought about by the relative phase shift of potential and density in an open spiral pattern, we expect the strength of the gravitational instability to correlate with the value of the phase

[9] In any realistic physical systems which contain dissipation, the transition from supersonic to subsonic flow can *only* be accomplished by a shock (see, e.g., Shu 1992, p. 77; Woodward 1975).

[10] This decrease in σ_r at a spiral phase after the shock does *not* mean that the stars, on the average, are cooled by the shock. First of all, an average star is not following a circular trajectory such as our azimuthal plot is displaying; rather its trajectory is turned sharply inward at the location of the shock, similar to the gas streamlines shown in Fig. 4 of Roberts (1969). So the downstream location for the mean orbit of a star would be at a slightly smaller radius after the shock. Second, as we have shown in § 3.1, the spatial distribution of the velocity dispersion is partially produced by the mean flow field in the spiral potential, and this is not the same as the temporal behavior of stellar velocity dispersion at a fixed location. There is in fact a secular increase in the stellar velocity dispersion, as well as a secular increase in Q, which is observed both in the Donner & Thomasson (1994) simulation and in the current simulation.

shift in a quasi-steady state. If the spiral pattern changes significantly on the local dynamical timescale, however, even at galactic radii where the phase shift is small, the disk material could still "unexpectedly" run into the spiral potential when crossing the spiral arm and be driven beyond the instability threshold. This has been found to be the case in our N-body results.

Similar behavior of Σ, σ_r, κ, and Q are also found for other radial locations, except for the fact that for a location outside corotation, the maximum of σ_r occurs at a spiral phase *after* the potential minimum, instead of before the potential minimum as is shown in Figure 5*b*. This last point is consistent with the fact that outside corotation, the sense of the material entering the spiral shock is reversed, so the spiral phase after the shock now corresponds to the upstream location for the streaming stars.

From Figures 5*e* and 5*f*, we see that the velocity components parallel and perpendicular to the spiral arm follow a trend of variation similar to that found by Levinson & Roberts (1981) in their two-dimensional hydrodynamical simulations of gaseous shocks (compare especially with their Figs. 8 and 9). Note that due to the inclusion of the self-gravity of gas, in their case, and the self-gravity of stars, in our case, the velocity jumps at the spiral arms both in their calculation and in our current calculation are not as sharp as that for the non–self-gravitating gaseous shock of Roberts (1969).

In Figure 6, we plot the disk morphology at three different time steps, as well as the azimuthal profiles of the grid density and potential at radius $r = 15$ (the location of the circle in the first frame) for the corresponding time steps, to illustrate the shock steepening process and its relation to the local phase-shift value. The density profile is seen to steepen with time from a more sinusoidal profile to a shocklike nonlinear profile. It is also seen from Figure 6 that the shock strength increases as the local phase shift increases, although there seems to be a time lag between the moment when the phase shift is largest (*bottom middle frame*) and the moment when the phase shift succeeded in compressing the density distribution into the narrowest nonlinear profile (*bottom right frame*). The shape of the density distribution in the middle frame, however, has the best resemblance to the classical hydrodynamic shock, where there is a very steep rising edge followed by a much more gradual falloff.

The time lag observed in Figure 6 between the moment of the large phase shift and the moment of narrowest density

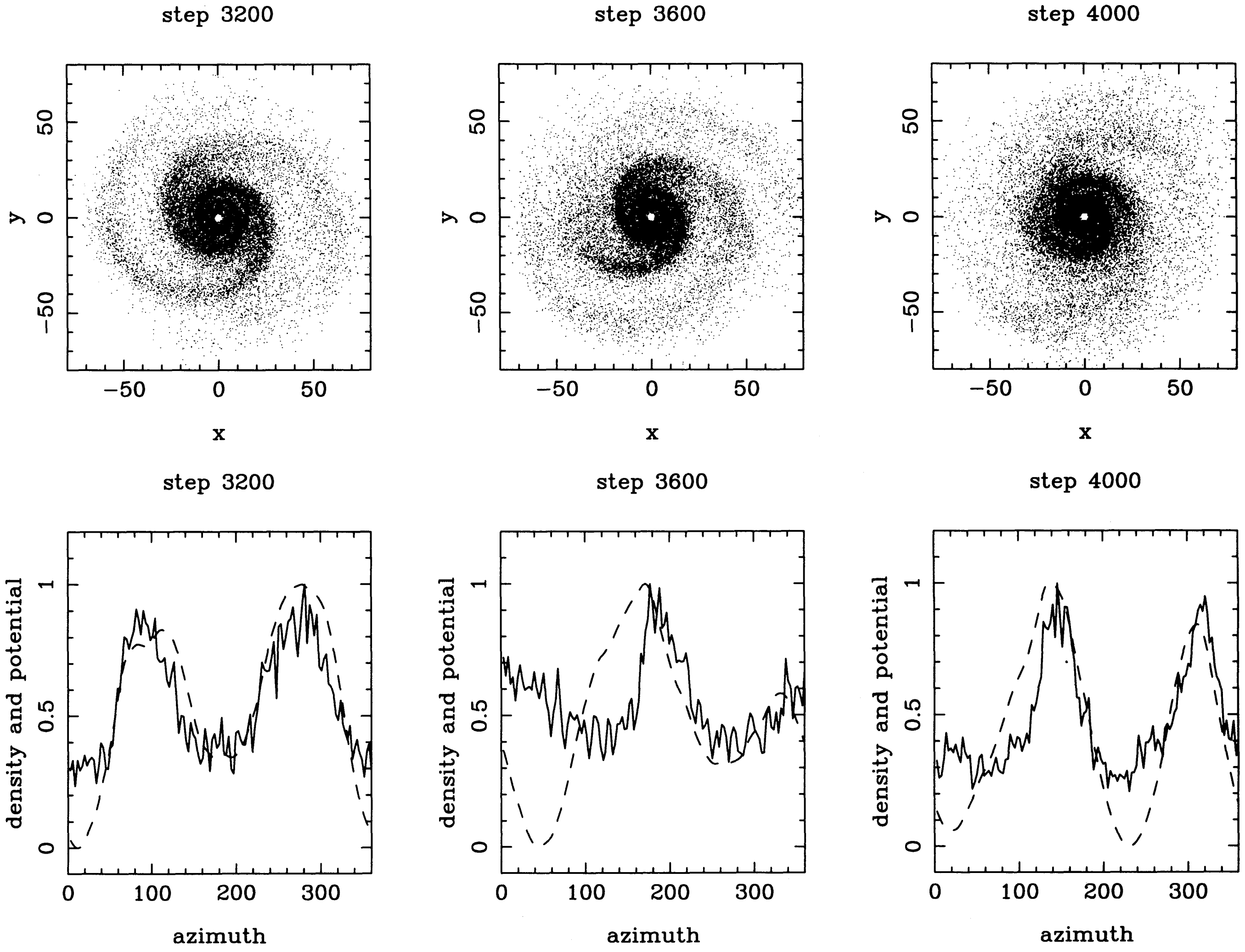

FIG. 6.—Steepening of spiral shock with time. *Top three frames:* Surface morphology evolution. *Bottom three frames:* Azimuthal density (*solid line*) and negative potential (*dashed line*) profiles at radius $r = 15$. The circle in the first frame indicates the radial location where the bottom azimuthal profiles are plotted. For the bottom three frames, the density is normalized to have a maximum of 1. The potential has arbitrary scale and is shifted in the vertical direction.

distribution is not unexpected. Only when a (nonlinear) spiral mode reaches a quasi-stationary stage can we expect an exact correspondence between the local phase-shift value and the local shock strength, whereas the state of the art of N-body simulations is still such that the nonlinear profile of a spiral mode cannot yet be made quasi-stationary, even though the underlying $m = 2$ mode is indeed found to have achieved a quasi-stationary amplitude in both the Donner & Thomasson (1994) simulation and in our current simulation. Despite the incompetency of the current N-body simulations in producing long lasting large-scale spiral shock distribution, we do continue to believe that real galaxies can do so for much longer than the local dynamical timescale, just from the number of observed spirals which have well-defined grand design spiral patterns. Although we still wait for the further improvement of the computational capabilities[11] and the art of N-body building to fully confirm our prediction that the dissipation effect induced by the collective instability at the spiral wave crest is responsible for sustaining the phase-shift distribution in a quasi-stationary spiral structure, even with the present N-body results, we have at least verified that whenever there is a shock steepening process in a piece of spiral arm, the present and immediately previous potential and density profiles are phase-shifted in the correct sense. So the phase shift appears to be driving the shock to steepen toward a level at which dissipation in the shock matches the local phase-shift value, with the phase shift itself given by the global Poisson integral.

Due to the presence of the local gravitational instability at the spiral arms, the width of the spiral shocks is effectively the size of the instability structure formed, i.e., on the order of 1 kpc for a galaxy like our own, instead of the free particle mean free path.[12] In other words, whenever there is a local gravitational instability, the range of influence of the gravitational scattering potential is no longer restricted to that around the individual particle itself but rather around the collection of particles that form the instability. Such views have already been expressed in Kulsrud (1972). The observed narrow width of the dust lanes in the leading edge of many spirals, on the other hand, may very well reflect the size of the gas cloud mean free path in the spiral instability, which could be smaller than the stellar mean free path there if the two fluids are not completely coupled.

The signature of spiral shocks in the stellar component has also been found in observations (Elmegreen & Elmegreen 1989), where it is noticed that for many spiral galaxies, the leading edge of the spiral arm is generally sharper than the trailing edge, and the shape of the wave in both the blue and the near-infrared band has the appearance of a water wave on the verge of breaking up. Looking back to the famous photographs of M51 referenced in Binney & Tremaine (1987, p. 342), which were originally obtained by Elmegreen (1981), we are much more aware now that the sharp density contrast, especially in the red frame of the picture, is strong evidence for the presence of spiral shocks in the stellar medium.

The term "gravitational shock" has been defined and used by Spitzer (1987, p. 110; see also Spitzer & Chevalier 1973) to refer specifically to the transient external gravitational perturbations experienced by a globular cluster when crossing the galactic plane or passing near the galactic center. We comment briefly here on the relation of Spitzer's gravitational shock to the spiral gravitational shock discussed in the current paper. Both types of gravitational shock are capable of inducing a temporary local gravitational instability on the originally marginally stable, self-gravitating mass ensemble, therefore accelerating the relaxation and evolution of the local cluster of matter. However, due to the unidirectional entry speed of the orbiting matter into the spiral shock, there is a unidirectional transfer of angular momentum caused by the spiral gravitational shock, besides the relaxation effect it induces, whereas in the case of the globular cluster the transient external gravitational perturbation causes mainly core contraction and the escape of stars from the cluster (Spitzer & Chevalier 1973).[13]

4. DISCUSSION

4.1. *Further Comments on the Results of N-Body Simulations*

In virtually all of the N-body experiments which formed spiral patterns, it is observed that there is a tendency for stars to accrete inward in the inner disk. In cases where a spiral mode is formed, the particles are found to accrete inward inside corotation and excrete outside corotation (Donner & Thomasson 1994). This same trend is also observed in our own N-body results.

In Figures 7 and 8, we have plotted the different segments of a typical orbit inside corotation and a typical orbit outside corotation, respectively, both obtained from the N-body simulations described in § 3.2. Note that the force interpolation scheme used in the N-body simulation would smear out any small-scale "kinks" in the orbit that could be produced by the local gravitational instability at the spiral arms. These kinks are expected to be very gentle in any case for real galaxies, since they are produced by small-angle scatterings, with the particle mean free path for scattering on the order of 1 kpc. The cumulative effect of these small-angle scatterings nevertheless survives, as is reflected in the often sharp change of the orbital orientation, observable especially in Figure 7, which is reminiscent of the sharp-pointed oval-shaped streamlines in a gaseous spiral shock (Roberts 1969, Fig. 4).[14] The dissipation effect of the spiral shock is also revealed through the secular decease (or increase) in the mean orbital radius. In Figure 9, the corresponding frames for the disk surface density are plotted.

[11] It has often been claimed in the literature that the number of particles used in the N-body simulations of spiral structures does not seem to matter, as long as the number is large enough so that binary relaxation is negligible. We have found that this conclusion seems to be valid only in cases where a spiral is formed as the result of the tidal interaction with a companion galaxy, or, in the case of a spontaneously formed spiral, if one is interested in the growth rate and pattern speed of $m = 2$ component only, but not in the highly nonlinear spiral shock structure.

[12] Note that for very small wave amplitude, the degree of steepening of the spiral wave can be very mild, as is the case during the initial growth phase of the wave. For such a wave, the mean free path of the particles in the spiral-arm gravitational instability is of the same order as the azimuthal wavelength of the $m = 2$ mode. Whether to call such a wave a spiral shock or not becomes purely a matter of semantics. In essence, there is not a sharp transition from a mild spiral instability to a strong spiral shock. There is only the gradual reduction of the mean free path of the particles in the spiral arm instability as the wave amplitude increases. The dissipation effect due to the phase shift is expected to be present throughout the wave growth process, although not always at its full capacity if the wave has not reached quasi-steady state.

[13] Though there could indeed be a small amount of momentum transfer of the globular cluster to the galactic plane due to the impact, if the crossing speed is high.

[14] Note, however, that Fig. 4 of Roberts (1969) plots the streamlines in a corotating frame, whereas our orbits are plotted with respect to the lab frame. In the corotating frame of the spiral pattern, the orbits obtained in our N-body simulation are found to be essentially chaotic. A coherent spiral pattern, however, is nonetheless seen to be supported by such chaotic-looking orbits, presumably because at each spiral-arm crossing, the action of the spiral shock enhances the phase correlation of the chaotic orbits with the spiral potential.

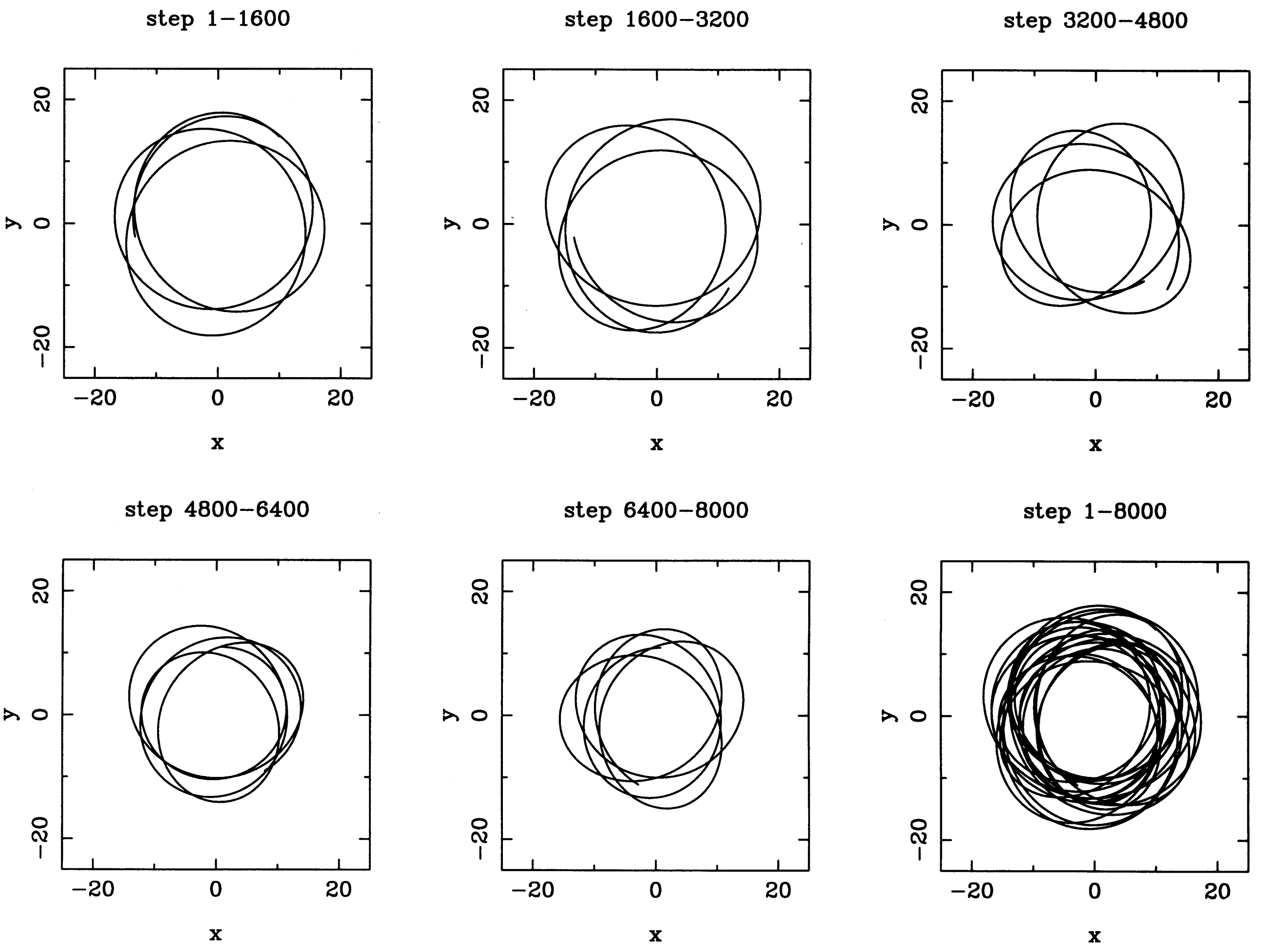

FIG. 7.—Evolution of orbit trajectory for a typical star inside corotation

The figure clearly demonstrates that there is a secular increase in the disk surface density in the inner disk region, together with a slight density increase in the outer disk, consistent with the trend shown in Figures 7 and 8 for the mean orbit evolution. Note that the second density peak on some of the frames in Figure 9 near $r = 10$ is due to the temporary accumulation of disk matter near the inner Lindblad resonance (ILR).

It has been suggested (Carlberg 1986) that the secular orbital change observed in the N-body simulations, at least in the case of transient spirals, is due to the effective broadening of resonances as a result of the transient forcing. However, a simple integration of the stellar orbit shows that this could not possibly be the case. In Figures 10*a* and 10*b* we have plotted a single star's orbital response to two transient spiral forcings of differing durations. The parameters used for the simulation are similar to those for the solar neighborhood and are given in the figure caption. The first case is a slow turning on and off of a spiral potential, with a timescale of variation of 10^{10} yr. The second case is a rapid spiral perturbation, with the timescale of the impulse being 10^8 yr. It is seen that in the impulse forcing case there is a residual epicycle motion after the forcing is past, whereas in the slow forcing case there is no residual heating (which is strictly true only for spiral forcing amplitude that is not too large). Both of these results are consistent with the order-of-magnitude estimate of Binney & Tremaine (1987, p. 482). However, what is also clear is that there is no secular change in the mean orbital radius, whether the transient is fast or slow. The view that the secular orbital change observed in the N-body simulations cannot be due to the resonance-broadening effect is also supported by the fact that these secular changes in the N-body orbits always consist of the simultaneous decrease of the inner *and* outer limits of orbital migration (for a star inside corotation), or the simultaneous increase of the two limits (for a star outside corotation), whereas, under the influence of resonant potential, the inner radius of the migration of a star will decrease, and the outer radius of the migration will increase, and the resonant orbit covers a larger and larger region of the space as time goes on, which is not the kind of behavior we observe for the orbits in Figures 7 and 8. Therefore, the secular orbital change observed in the N-body simulations cannot be explained by the behavior of a single orbit in an applied spiral potential. It has to be a result of the collective dissipation effect induced by a self-sustained global spiral structure.

The presence of a collective dissipation process at the spiral arms could also explain why, in most N-body simulations of spiral structures, the heating rate of the stars is found to be much higher than the angular momentum transport rate (see,

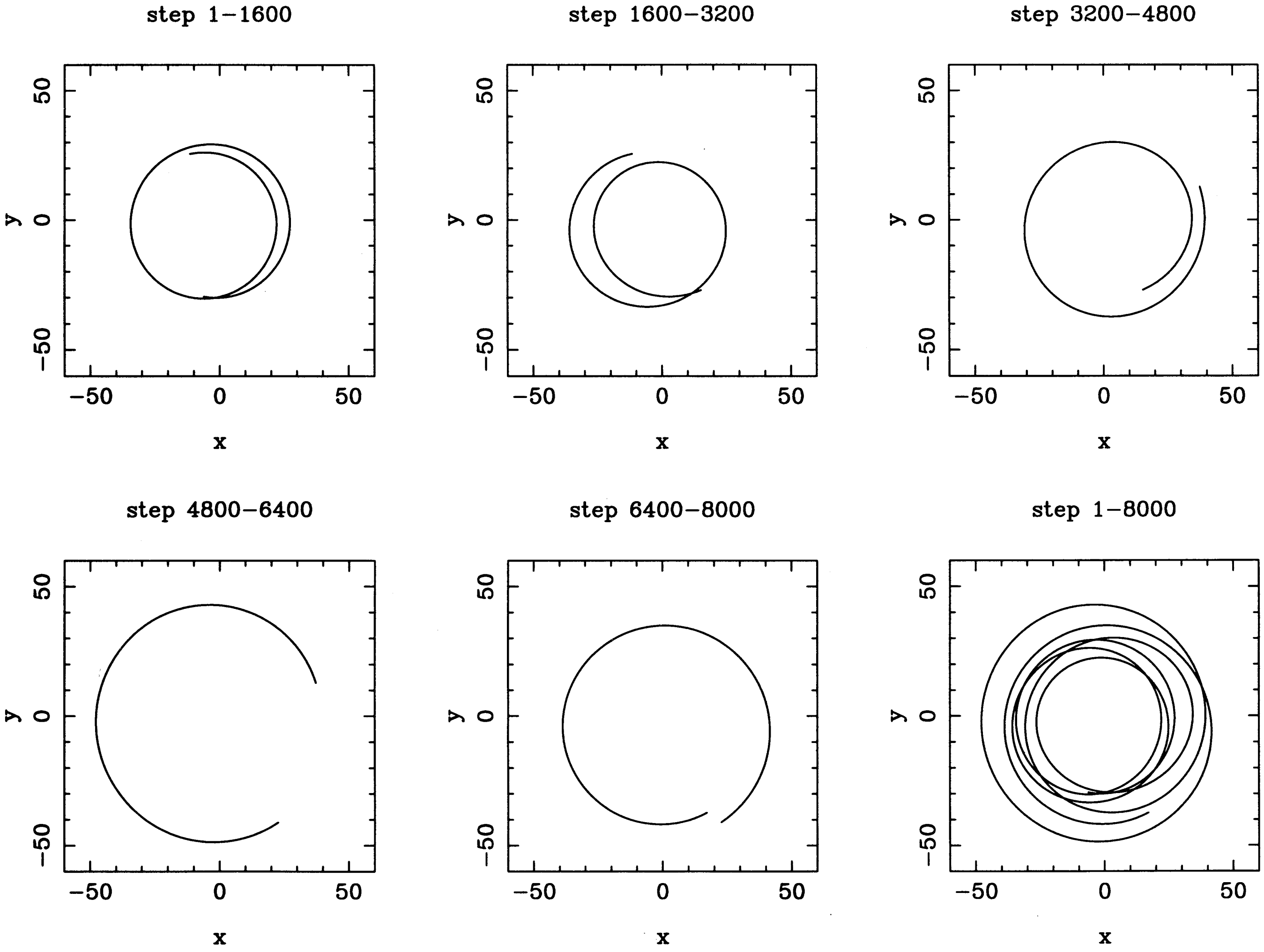

FIG. 8.—Evolution of orbit trajectory for a typical star outside corotation

e.g., Carlberg 1986), whereas, in reality, we expect the two to be of the same order for quasi-steady evolution, as is required by the virial theorem. From the discussion in the current paper, we see that collective dissipation effectively causes "collisions" (scatterings) among the stars which are crossing the spiral arms. This greatly accelerates the rate of local relaxation. From such a collisional process some stars can carry away much of the orbital angular momentum but leave energy behind in the form of heat. This effect is expected to be more pronounced in the N-body simulations than in real spiral galaxies, since Gaussian random noise has an $N^{-1/2}$ dependence, where the relevant N here is the total number of particles in the local–gravitational-instability clump at the spiral arms. Thus it is possible for the N-body disk to evolve to a Q of 2–3 in just a few rotation periods, whereas in the outer disk of observed spiral galaxies the Q is found to be close to unity.

Because of the nature of the spiral pattern as a global instability, which always has its associated collective dissipation and relaxation effect, the N-body simulations which model the spontaneous formation of spiral patterns cannot be made truly collisionless, even if such a code has no binary relaxation at all before the emergence of a spiral structure. A similar view has also been expressed in Weinberg (1993).

4.2. *Comparison with the Results of Lynden-Bell & Kalnajs (1972)*

There are two major conclusions in the seminal paper of LBK. The first is that a trailing spiral structure transports angular momentum outward; the second is that a disk star does not exchange angular momentum with a quasi-stationary spiral wave except at the resonances. We now compare these two conclusions of LBK with the results of the current paper.

First of all, what LBK showed about the angular momentum transport was in fact a weaker result than that stated above. They have indeed shown that the sign of gravitational torque coupling C_z is such that a trailing spiral wave transports angular momentum outward. However, as is also shown in the same paper of LBK, there is a second torque coupling C^* due to advection (lorry transport), which also contributes to the total torque coupling between the inner and outer disk in a spiral galaxy. When these two contributions are summed together, LBK found that, at least for waves which are not too long, the total torque coupling is in the form of angular momentum density multiplied by the group velocity of the wave (this result had previously been obtained by Toomre 1969). Therefore, for a wave inside corotation, which has nega-

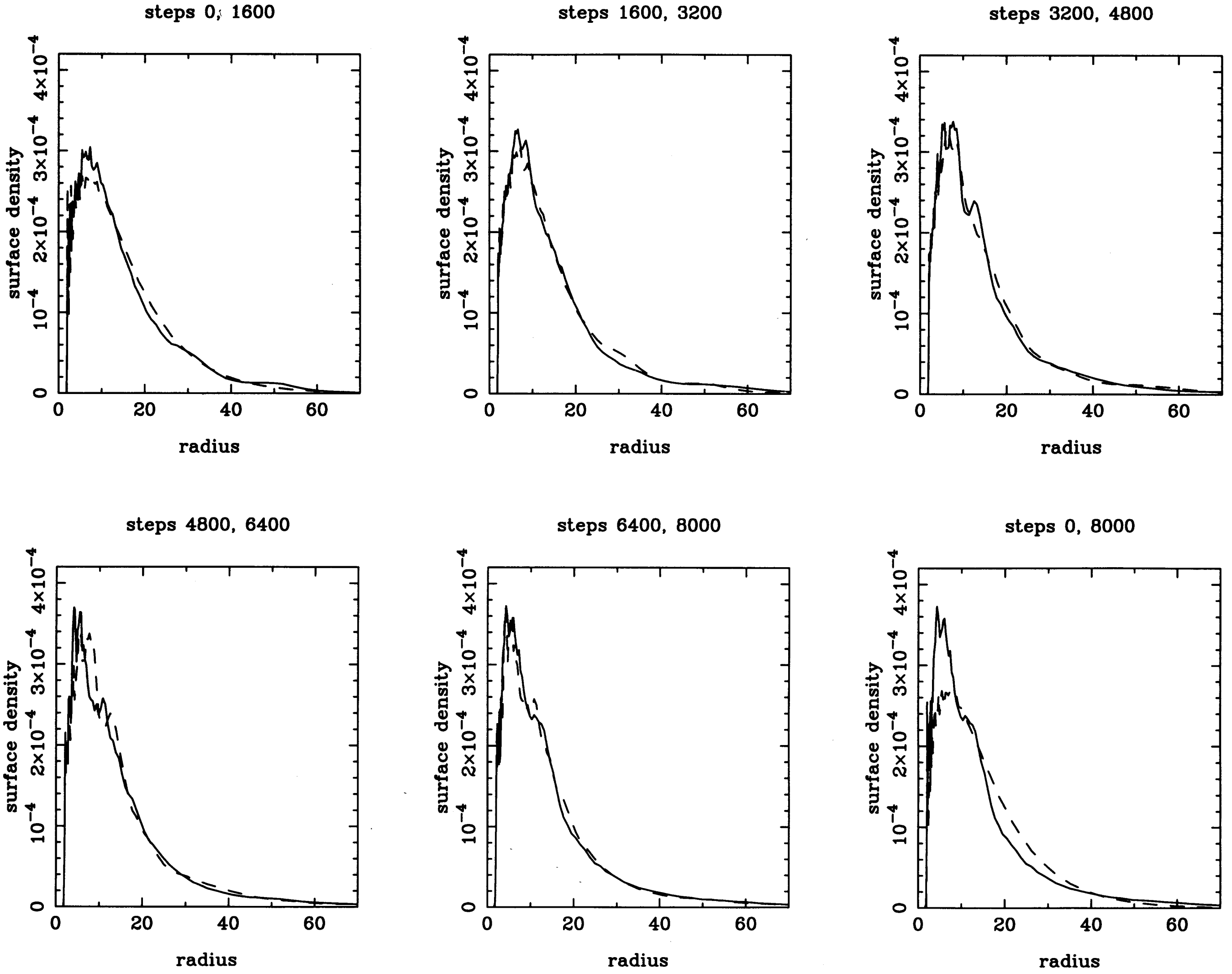

FIG. 9.—Evolution of disk surface density. The dashed line in each frame indicates the surface density at the earlier time step, and the solid line the later time step.

tive angular momentum density, the net group velocity of the wave must be directed inward in order for the trailing wave to transport angular momentum outward. Fortunately, this condition is satisfied for most of the trailing spiral structures we observe, thanks to the overreflection mechanism at corotation (Mark 1976; Toomre 1981), which makes the inward-propagation trailing wave train always more powerful than the outward-propagating leading or trailing wave trains.

Therefore, the essence of LBK's first conclusion can be rephrased as follows: a spiral wave carries angular momentum with it as it propagates. However, due to the second conclusion of LBK, that of no angular momentum exchange between a nonresonant disk star and a wave, the source and sink of the outward angular momentum transport were thought to reside only at the Lindblad resonances for a transient wave train. Moreover, for galaxies which have a Q-barrier that shields the ILR, so that a spiral mode can grow, the outward transport of energy and angular momentum merely leads to the growth of the wave mode both inside and outside of corotation, but no evolution in the basic-state morphology. In the latter case, the waves inside and outside corotation become the source and sink of each other, with essentially no wave interaction with the basic state except for the fact that the growing number of stars which participate in the growing wave ultimately come from the basic state.[15]

The torque coupling integrals obtained by LBK concerns a different physical process from the torque integral in equation (4) of the current paper. The detailed discussion of the relation between the different torque integrals will be given in Paper II. In essence, LBK torque coupling integrals describe an angular momentum transport process by a trailing spiral wave, whereas the torque integral in equation (4) describes the angular momentum exchange between the disk stars and a

[15] In fact, the facilitation of wave growth through the outward transport of angular momentum was the chief function that LBK had attributed to a trailing spiral structure. This is because a spiral wave has negative angular momentum density inside the corotation, so a process which removes angular momentum from the inner disk is expected to lead to wave growth. This was also the reason that LBK had named their paper, "On the Generating Mechanism of Spiral Structure."

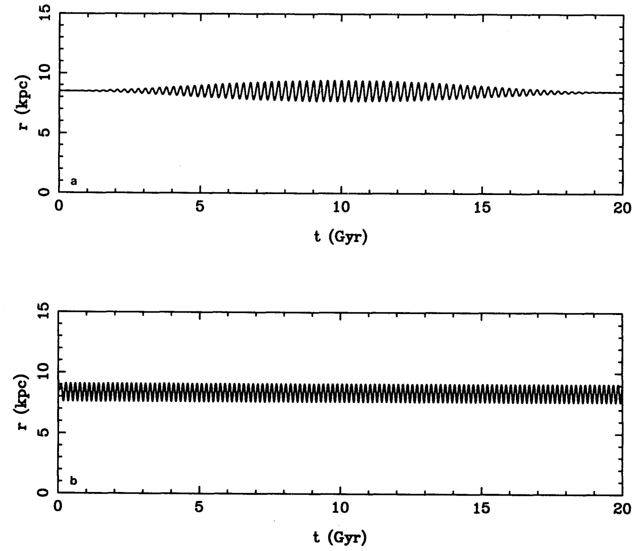

FIG. 10.—Orbit response to transient spiral excitations. Modulation function is $-\cos(t/t_1\pi) + 1$ for $t \leq 2t_1$, and zero otherwise. (*a*) Slow transient with $t_1 = 10^{10}$ yr. (*b*) Fast transient with $t_1 = 10^8$ yr. Parameters common to these two cases are as follows: the strength of the spiral is 5%, the pitch angle of the spiral $i = 20°$, $\Omega_p = 13.5$ km s^{-1} kpc^{-1}, $r(t = 0) = 8.5$ kpc, $v_c = 220$ km s^{-1}, and the total integration time is 20 Gyr. The test particle was initially on a circular orbit.

quasi-stationary density wave (i.e., the loading and unloading of the angular momentum onto and out of the wave). A steady-amplitude density wave mode acts simply as an intermediary for the absorption, transport, and dissipation of the disk energy and angular momentum, and the wave itself can remain quasi-stationary on the timescale of a Hubble time (Paper II). In this sense, the entire wave is being used as a lorry for the outward transport of angular momentum. Therefore, we now see that over the larger portion of the lifetime of a spiral galaxy, the main function of the outward angular momentum transport by a trailing spiral structure is not to allow the wave to grow, as LBK had originally envisioned, but rather to allow the basic state to evolve.

We now take a closer look at the second conclusion of LBK, that of no angular momentum exchange between a spiral wave and a nonresonant star, which LBK had demonstrated to second order. From a direct numerical integration of nonlinear orbits in a spiral potential (Zhang 1995d), we found that the conclusion of no angular momentum exchange *is* the correct conclusion, in fact, to all perturbation orders in a nonlinear calculation, *if* the star only experiences a smooth axisymmetric plus spiral potential but no collective effect. This conclusion is also supported by the fact that the Jacobi integral is a constant for a stellar orbit in a smooth spiral potential. In order to conserve the Jacobi integral, the amount of energy loss and the amount of angular momentum loss of a star in its interaction with the wave have to have the ratio Ω_p, which is, in general, not equal to Ω. Therefore, the *secular* exchange of energy and angular momentum of a single star with a quasi-steady wave is prohibited by the constancy of the Jacobi integral at non corotation radii, if there is no collective dissipation mechanism which converts part of the stellar orbital energy into heat (i.e., epicycle motion), whereby the phase of the noncircular component of the stellar orbital velocity is decorrelated with the phase of the spiral wave.[16] This will be addressed further in Paper III. In fact, without collective dissipation, a finite-amplitude spiral wave could not even obtain its coherent organization and form a self-consistent pattern (Zhang 1995d). Only by dissipating part of their orbital energy at each crossing of the spiral arm potential do stars participate in wave motion. This is a view of the maintenance of the spiral pattern which is quite different from that offered by the "kinematic spiral" mechanism, where the orbits are the true "building blocks" of a global pattern. In essence, the single-orbit response in an applied spiral potential does not tell the whole story of a self-consistent spiral wave, since it does not incorporate collective effects.

4.3. *Other Astrophysical Consequences*

The spiral-induced collective dissipation process leads to the secular transfer of energy and angular momentum from the disk material to the spiral density wave in the inner disk. Since the wave inside corotation has negative energy and angular momentum, upon receiving energy and angular momentum the wave will be damped in amplitude. Similarly, the wave outside corotation is also damped, owing to its giving away energy and angular momentum to the basic state. This is, in general, a nonlinear as well as a dissipative process, as is reflected in the gradual deformation of the wave from sinusoidal profile to the more sharply peaked nonlinear profile as the wave amplitude increases. It is found that the phase shift calculated through the Poisson equation is not sensitively dependent on the azimuthal profile of the wave, as long as the global distribution of the spiral density (i.e., pitch angle and radial falloff) remains the same. Consequently, the dissipation rate is largely independent of the evolution in the azimuthal profile of the wave. The growth rate of the wave, on the other hand, decreases as the azimuthal profile of the wave is deformed due to energy conversion from the $m = 2$ mode into the higher harmonics in the azimuth. As a result, the growth rate and the dissipation rate of the wave eventually reach an equilibrium, so a nonlinear quasi-stationary spiral mode can be obtained (Paper II).

Because of the secular energy and angular momentum exchange between the disk matter and the wave, the disk material (including both stars and gas) inside corotation spirals inward, and the material outside corotation drifts outward, and this leads to the secular evolution of the disk surface density. The timescale of the radial mass accretion process for a galaxy like our own is estimated to be on the order of a Hubble time (Zhang 1992; Paper III), which is of the same order as given by the early *N*-body simulations (Carlberg 1986), and which is found to be able to account for the formation of the quasi-exponential disk surface density profile and the formation of bulges in galaxies (Zhang 1995a; Paper III).

For stars moving on nearly circular orbits, the ratio of their energy loss to the angular momentum loss is proportional to Ω, the circular speed of the disk. On the other hand, the ratio of energy to angular momentum that can be absorbed by the density wave is proportional to Ω_p, the pattern speed of the wave. Since inside corotation $\Omega > \Omega_p$, the complete transfer of

[16] This is expected to be the case even at the Lindblad resonances. In other words, the energy and angular momentum exchange between the basic state and a quasi-stationary wave at the Lindblad resonances, found in many previous linear-order calculations, has implicitly incorporated collective effects.

angular momentum from the disk material to the wave (since the angular momentum has nowhere else to go) means that the energy released by the disk material cannot be completely absorbed by the density wave. The surplus of this released energy contributes to the secular heating of the disk stars. This is found to quantitatively explain the observed age-velocity dispersion relation of the solar neighborhood stars (Zhang 1995a; Paper III).

The trend of evolution of the basic state of a spiral galaxy due to the spiral-induced collective dissipation process coincides with the trend found in the Hubble sequence from the late to early spiral types, whereby a spiral galaxy gradually acquires a thicker inner disk and a larger bulge. This change in the basic-state property, in turn, results in the change of the kind of spiral modes present, from the more open type to the more tightly wound type (Bertin et al. 1989a, b), again consistent with the correlation observed in the Hubble classification. So it is likely that the Hubble sequence, when viewed in the reverse direction, indicates a temporal evolution sequence. We will address this further in Paper III.

5. CONCLUSIONS

In this paper we have proposed and analyzed a collective dissipation process induced by a galactic spiral structure. This process reveals itself as a phase shift between a self-sustained spiral potential and density pair, and the dissipation effect indicated by the phase shift is achieved though a mild local gravitational instability at the spiral arms. Owing to the instability condition and the dissipation process at the spiral wave crest, a large-scale spiral structure is, in essence, a large-scale spiral gravitational shock. We argue that the spiral-induced collective dissipation is the only means to account for the secular orbital decay or increase observed in the N-body simulations of spiral structure. The dissipation must be spiral-induced because there is a clear distinction between the behavior of an average stellar orbit that is inside or outside corotation, which indicates that the secular orbital change must be due to a mechanism which is related to the presence of the spiral structure, instead of due to a nondiscriminating viscous force such as dynamical friction. The dissipation must also be achieved through a collective process because the non–self-consistent calculation of orbital response under an applied spiral potential shows that the mean radius of a single orbit never changes secularly, whether the applied spiral potential is quasi-steady or transient. Collective dissipation can operate only at the spiral arms, because only there can neighboring stars interact with one another directly, as a result of the instability condition produced there by the potential-density phase shift. We expect that a similar phase-shift–related dissipation mechanism is also operating in disks which contain other nonaxisymmetric (as well as nonbilaterally symmetric, etc.) types of instability structures, such as in barred galaxies where the two sides of the bar are offset, and in stellar accretion disks which contain $m = 1$ spiral instabilities.

The author would like to express her sincere gratitude to Colin Masson for his support and encouragement, as well as many helpful comments during the completion of this work. Thanks also to Paul Jaminet, Colin Masson, and Scott Paine for helpful criticism of the manuscript. Thanks to Alessandro Romeo for sending me his Ph.D. thesis, which contains many helpful references and insights. Thanks to Magnus Thomasson for his research report 162, which made the task of writing the N-body code used for the current work a much more manageable one. During the various stages of this work the author has benefited from correspondence and conversations with Bruce Elmegreen, Agris Kalnajs, Ivan King, C. C. Lin, Ramesh Narayan, Eve Ostriker, Alice Quillen, Paul Schechter, Frank Shu, Antony Stark, Alar Toomre, and Scott Tremaine. The helpful comments and suggestions of the referees, especially the second referee, Frank Shu, are gratefully acknowledged. The early part of this work was conducted when the author was working on her Ph.D. thesis at the University of California, Berkeley, with the Radio Astronomy Laboratory. The latter part of this research was supported by a Smithsonian postdoctoral fellowship.

APPENDIX A

A PHASE SHIFT GIVEN BY THE POISSON EQUATION

Due to the long-range nature of the gravitational interaction, the potential field generally has a different distribution from the mass density which generates it. In the case of the spiral wave, this difference appears in the form of a phase shift between the potential and density spirals related through the Poisson equation.

Although the author discovered the existence of phase shift in a Poisson transform pair independently through numerical integration (Zhang 1994), motivated by the conviction that spiral galaxies are dissipative structures, and a phase shift must arise if this dissipation is to be achieved, she later learned that mathematicians have long since known about this fact in the study of potential theory, especially in the branch which deals with the spiral transformation (Snow 1952), which has been applied successfully by Kalnajs (1965, 1971) to the study of galactic spiral structures. However, before the current work, this phase shift in the Poisson transform pair was mostly considered just a nuisance in the analytical calculations of the spiral modes, and no one had suspected the relevance of it to the collective dissipation and evolution of spiral galaxies.

Since the presence of phase shift in the Poisson equation has never been discussed before in any of the astrophysical literature, we present here a brief derivation due to A. J. Kalnajs (1994, private communication), by employing the spiral transformation.

Define $u = \ln r$; it follows that a power of α of r, r^α, can be written as $e^{\alpha u}$.

From Snow (1952) or Kalnajs (1971), we know that a reduced density

$$r^{3/2}\Sigma(r, \phi) = e^{i(\alpha u + m\phi)} \tag{28}$$

will produce a reduced potential

$$r^{1/2}\mathscr{V}(r, \phi) = -2\pi G K(\alpha, m) e^{i(\alpha u + m\phi)} , \tag{29}$$

where K is the ratio of several gamma functions (Kalnajs 1971). Since $K(\alpha, m)$ is an analytic function of α, the above relation between the reduced potential and reduced density is still true for complex values of α, provided that the imaginary part is sufficiently small. Write

$$\alpha = \alpha_r + i\alpha_i \ , \tag{30}$$

and we have

$$e^{i(\alpha_r + i\alpha_i)u} = e^{-\alpha_i u} e^{i\alpha_r u} \ . \tag{31}$$

The above expression, when substituted back in equations (28) and (29), indicates that, for an infinitely long density spiral with radial density falloff differing from $r^{-3/2}$, its corresponding potential spiral, although it still has the property that its radial modulation function is r times the radial modulation function of the density spiral, is phase-shifted with respect to the density spiral because $K(\alpha, m)$ now becomes slightly complex. By expanding $K(\alpha, m)$ in a Taylor series around α_r, we have

$$K(\alpha, m) = K(\alpha_r + 0i) + \left.\frac{\partial K}{\partial \alpha}\right|_{\alpha = \alpha_r} (i\alpha_i) + \text{higher order terms} \ . \tag{32}$$

For $\alpha_r > 0$ (which corresponds to a trailing spiral in our convention), it can be shown that the first derivative of $K(\alpha, m)$ is negative (Kalnajs 1971, Table 1) for small α. Therefore, if $\alpha_i > 0$, which means that the density (potential) falloff is faster than $r^{-3/2}$ ($r^{-1/2}$), the potential spiral will lead the density spiral (i.e., the potential is in the form of $Ce^{i[\alpha_r \ln r + m(\phi - \phi_0)]}$, with $\phi_0 > 0$), and vice versa for $\alpha_i < 0$.

The presence of a phase shift in the sense we described above can also be seen from the higher order asymptotic expansion of the differential form of the Poisson equation (Shu 1970, eq. [11]).

APPENDIX B

PHASE SHIFTS IN THE EULERIAN EQUATIONS OF MOTION AND IN THE LINEAR PERIODIC ORBIT SOLUTION

In order to obtain a self-consistent spiral wave solution which admits a phase shift, there has to be a corresponding phase shift which exists in the potential and density relation given by the equations of motion. As is shown in equation (D12) of Lin & Lau (1979), the relation between the spiral potential and density, obtained from the higher order asymptotic solution of the linearized Eulerian equations of motion and the equation of continuity, indeed contains a so-called out-of-phase term; and careful analysis shows that the sign of this term is such that for a trailing spiral, the density leads in phase to the potential inside corotation, and vice versa outside corotation.

Since the surface density in the case of a stellar disk is ultimately made of the superposition of stellar orbits, we expect that the relative phase shift of the potential and density spirals is also reflected in the orbital response of a star in a more open type of spiral potential. Especially in the linear regime, where there is an exact correspondence between the Eulerian and Lagrangian approaches, at least for the pressureless case (Lau & Bertin 1978), a phase shift has to exist in the orbit solution if it exists in the Eulerian fluid solution. In what follows we demonstrate that there is indeed a phase offset in the orientation of the linear periodic orbit, which is obtained in the corotating frame of a spiral potential.

The linearized orbit equations in a frame that corotates with the pattern at an angular speed Ω_p are (Binney & Tremaine 1987, eqs. [3-114a] and [3-114b])

$$\ddot{r}_1 + \left(\frac{d^2\Phi_0}{dr^2} - \Omega^2\right)_{r0} r_1 - 2r_0 \Omega_0 \dot{\phi}_1 = -\left(\frac{\partial \Phi_1}{\partial r}\right)_{r0} , \tag{33}$$

$$\ddot{\phi}_1 + 2\Omega_0 \frac{\dot{r}_1}{r_0} = -\frac{1}{r_0^2}\left(\frac{\partial \Phi_1}{\partial \phi}\right)_{r0} , \tag{34}$$

where r_1 and ϕ_1 are the perturbed orbital coordinates, Φ_0 is the axisymmetric potential, $\Omega = [(1/r)(d\Phi_0/dr)]^{1/2}$ is the angular speed, $\Omega_0 \equiv \Omega(r_0)$, with r_0 the zeroth-order radius where the potential and the angular speed are evaluated, and, finally, Φ_1 is the perturbation potential, which we choose to be of the spiral form

$$\Phi_1(r, \phi) = F(r) \cos [f(r) + m\phi] , \tag{35}$$

where $\phi = \phi(t) = \phi_1(t) + (\Omega_0 - \Omega_p)t$, which for nonresonant stars can be approximated by $\phi \approx \phi_0(t) \equiv (\Omega_0 - \Omega_p)t$.

Integrating equation (34), we obtain

$$\dot{\phi}_1 + 2\Omega_0 \frac{r_1}{r_0} = -\frac{(\Phi_1)_{(r_0, \phi_0)}}{r_0^2} \frac{1}{\Omega_0 - \Omega_p} + \text{constant} \ . \tag{36}$$

Substituting equation (36) in equation (33) to eliminate ϕ_1, we obtain

$$\ddot{r}_1 + \kappa_0^2 r_1 = -\left(\frac{\partial \Phi_1}{\partial r}\right)_{(r_0,\phi_0)} - 2\Omega_0 \frac{\Phi_1(r_0)}{r_0} \frac{1}{\Omega_0 - \Omega_p}, \tag{37}$$

where we have ignored the constant term for the same reason as given in Binney & Tremaine (1987, p. 480), and where

$$\kappa_0 = \left(r \frac{d\Omega^2}{dr} + 4\Omega^2\right)_{r_0} = \left(\frac{d^2\Phi_0}{dr^2}\right)_{r_0} + 3\Omega_0^2 \tag{38}$$

is the epicycle frequency.

Using the expression of the perturbation potential of equation (35), the forcing terms on the right-hand side of equation (37) can be found to be

$$\text{rhs} = \left[-F'(r_0) - 2\Omega_0 \frac{F(r_0)}{r_0} \frac{1}{\Omega_0 - \Omega_p}\right] \cos\left[f(r_0) + m\phi_0\right] + F(r_0)k(r_0) \sin\left[f(r_0) + m\phi_0\right], \tag{39}$$

where $\phi_0 = \phi_0(t) = (\Omega_0 - \Omega_p)t$, and $k(r_0) \equiv (df/dr)_{r_0}$.

We now see clearly that the forcing consists of two terms which are 90° out of phase with each other. We thus expect that the forced orbital response, or the particular solution of equation (37), will also contain two similar terms. In fact, the particular solution of equation (37) can be written as

$$r_1(t) = \frac{1}{\kappa_0^2 - m^2(\Omega_0 - \Omega_p)^2} \left\{\left[-F'(r_0) - 2\Omega_0 \frac{F(r_0)}{r_0} \frac{1}{\Omega_0 - \Omega_p}\right] \cos\left[f(r_0) + m\phi_0\right] + F(r_0)k(r_0) \sin\left[f(r_0) + m\phi_0\right]\right\}. \tag{40}$$

This solution for r_1 can be further written as

$$r_1(t) = C \sin(m\phi' + m\delta), \tag{41}$$

where $\phi' = f(r_0)/m + \phi_0$, and

$$C = \sqrt{A^2 + B^2}, \tag{42}$$

$$\delta = \frac{1}{m} \tan^{-1}\left(\frac{A}{B}\right), \tag{43}$$

with

$$A = \frac{1}{\kappa_0^2 - m^2(\Omega_0 - \Omega_p)^2} \left[-F'(r_0) - 2\Omega_0 \frac{F(r_0)}{r_0} \frac{1}{\Omega_0 - \Omega_p}\right], \tag{44}$$

and

$$B = \frac{1}{\kappa_0^2 - m^2(\Omega_0 - \Omega_p)^2} F(r_0)k(r_0). \tag{45}$$

Note that it is the sine form of the orbital response that should be compared with the negative cosine form of the forcing spiral potential in order to derive the relative phase shift. We have also assumed $F(r) < 0$ here.

Therefore, we have derived that the phase shift δ of the orbit with respect to the forcing potential can be expressed as

$$\delta = \frac{1}{m} \tan^{-1}\left\{\frac{-F'(r_0) - [2\Omega_0 F(r_0)/r_0]/(\Omega_0 - \Omega_p)}{F(r_0)k(r_0)}\right\}. \tag{46}$$

Since the rate of amplitude variation of the density wave $F'(r_0)$ is expected to be small, equation (46) can be further simplified to

$$\delta \approx \frac{1}{m} \tan^{-1}\left[-\frac{2\Omega_0}{\Omega_0 - \Omega_p} \frac{1}{k(r_0)r_0}\right], \tag{47}$$

which tells us that the phase shift δ is negative for a trailing wave ($k > 0$ in our current convention) inside corotation. A negative δ here means that the orbit leads in space (lags in time) with respect to the spiral potential; the opposite is true for orbits outside the corotation. This agrees with that obtained from the fluid equations of motion.

The orbit phase shift we have just derived is absent from the result of a similar derivation in Binney & Tremaine (1987, p. 480). This is due to the fact that Binney & Tremaine have used the short-wavelength WKBJ approximation ($kr \gg 1$) and have ignored a number of terms that account for the openness of the spiral. It is exactly the openness (i.e., the finite pitch angle) of the spiral structure which leads to the phase shift in the orbital response.

APPENDIX C

THE SECULAR DISSIPATION EFFECT INDICATED BY THE PHASE SHIFT

Since both the Poisson equation and the linearized Eulerian equations of motion give a spiral potential and density which are phase-shifted from each other, a self-consistent linear global spiral mode can be constructed with a phase shift between the potential and density, as long as the radial density modulation of the spiral mode is such that it causes the phase shift given by the Poisson equation to change sign at the corotation radius. The numerical calculation of linear and global spiral modes has been carried out successfully by C. C. Lin and his collaborators (Bertin et al. 1989a, b and references therein).

Within the context of the linear calculation, however, the phase shift we have just demonstrated does not have a dissipation effect. This is because linear calculation aims to obtain a self-consistent flow solution, which is periodic in the case of a spiral galaxy, with neighboring streamlines (in the fluid approach) never crossing one another. There are further signs that the dissipation effect indicated by the phase shift is not already incorporated in the linear theory:

1. The appearance of the torque integral in equation (4),

$$\mathscr{T} = -\frac{1}{2\pi}\int_0^{2\pi} \Sigma_1 \frac{\partial \mathscr{V}}{\partial \phi}\, d\phi \,,$$

is in the form of the product of two perturbation quantities, therefore a second-order interaction term with respect to the basic-state quantities. Such an interaction could not happen in the context of the linear theory, since in linear calculation the perturbation potential operates effectively on the zeroth-order basic-state solution and induces a perturbation density response; the perturbation quantities do not operate back again on perturbation quantities.

2. The dissipation effect indicated by the phase shift involves an energy and angular momentum exchange between the basic state and the density wave. However, in the stellar dynamical case it has been demonstrated that the spiral solutions obtained have the property of conservation of wave action during wave propagation (to second order in asymptotic expansion), and the amplification of the wave is through the energy and angular momentum exchange between the three waves inside and outside corotation (Mark 1974, 1976). In the fluid dynamical case, similar conclusions can be reached through analyzing the result of the growth rate of the global modes (Lin & Lau 1979). This further confirms that in linear theory there is no energy and angular momentum exchange between the density wave and the basic state. The result of the exact solution of linear equations for spiral modes is fairly close to the result of the asymptotic solution, in density contours as well as in the growth rate and the pattern speed obtained (see Lin & Lau 1979; Bertin et al. 1989b, especially Fig. 5). Thus we expect that, in the exact solution of the linear equations, there is also a separate conservation of energy and angular momentum for the basic state and the density wave.

3. It has been shown through numerical integration that the inclusion of the out-of-phase terms does not have a significant effect on the calculated modal growth rate (Lau & Bertin 1978; Li, Han, & Lin (1976). In fact, the growth rate even *increased* slightly when the out-of-phase terms were included, whereas the energy and angular momentum transfer process between the basic state and the density wave, as is indicated by the phase shift, should lead to wave damping (Paper II). This is why, in most of the asymptotic calculations of the spiral modal growth rate, the out-of-phase terms are usually neglected (Lau & Bertin 1978).

The dissipation effect indicated by the phase shift can only be revealed by going to the next order, through performing the torque integral in equation (4). This kind of dissipation calculation, using the product of two perturbation quantities which are obtained from solving the linearized dissipationless equations, is similar in nature to the kind of quasi-linear calculation which is standard practice in plasma physics (Krall & Trivelpiece 1973, chap. 10). The fact that the quasi-linear approach, by using the lower order results of the dissipationless equations, could "predict" the effect of a process which is dissipational in nature, is true because the quasi-linear approach essentially takes the linear solutions which are fixed in shape, and therefore artificially *avoided* the tangling of streamlines which would happen in the solution of nonlinear and dissipationless equations (see also the discussion at the end of this section). This disentangling of streamlines is exactly the function which is achieved by dissipation in real physical systems.

Incidentally, if we calculate the net angular momentum gain (or loss) of a star when completing a full cycle on the inclined linear periodic orbit, which we have obtained in Appendix B, through calculating the integral

$$-\int_0^{P_r} \frac{\partial \Phi}{\partial \phi}\, dt = -\int_0^{P_r} \left[\left(\frac{\partial^2 \Phi}{\partial r\, \partial \phi}\right)_{(r_0,\, \phi_0)} r_1 + \left(\frac{\partial^2 \Phi}{\partial \phi^2}\right)_{(r_0,\, \phi_0)} \phi_1\right] dt \,, \tag{48}$$

with $\phi_0 = \phi_0(t) = (\Omega_0 - \Omega_p)t$, and with the use of equation (47) which relates $\sin \delta$ and $\cos \delta$, we find that it is zero.[17] Although, as we have mentioned before, there is an exact correspondence between the Eulerian and Lagrangian approaches, there is in fact no contradiction between the two results that, on the one hand, the torque calculation in equation (4) gives a nonzero result, and, on the other hand, the calculated net angular momentum gain of a single star on a linear and periodic orbit is zero. This is because the linear periodic orbit corresponds to the orbital response under an infinitesimal spiral potential, and such linear orbits for the neighboring galactic radii do not intersect, and therefore there could be no collective dissipation effect. Or, looking at it in another way, in the limit of infinitesimal wave amplitude the torque integral in equation (4) would itself give a null result for the net amount of angular momentum exchange.

Even for finite wave amplitude, where there is already spiral-induced dissipation, the torque integral in equation (4) becomes exact only when the spiral structure is quasi-stationary on dynamical timescales. During the initial spontaneous growth stage of a spiral instability, when the wave has not reached a quasi-steady state, the phase shift could lead partially to a rapid change in the

[17] An earlier draft of the current paper contains an error in this calculation, which was corrected by S. Tremaine.

morphology of the spiral, rather than contributing entirely to the dissipation effect given by the torque integral in equation (4). This has actually been observed in our N-body simulation of the spontaneous growth of a spiral mode, as shown in § 3.2.

Finally, we comment that the dissipationless Eulerian equations are not capable of predicting all the physical effects in a spiral disk which admits collective dissipation processes. Therefore, the past proofs of no-wave/basic-state interaction using the nonlinear Eulerian equations could not invalidate our results in the current paper. Nevertheless, we are still curious about what new phenomena will emerge from the calculation of an open spiral pattern using the nonlinear Eulerian set of equations. A little contemplation shows that the situation here is very similar to the steepening of acoustic waves in a nonlinear and dissipationless medium (see, for example, Shu 1992, pp. 203 ff.). Because of the difference in effective propagation speed for waves of different amplitudes, when the nonlinear Eulerian equations are solved as an initial-value problem, the peak region of a finite-amplitude sound wave gradually catches up to the trough of the wave located originally at its front, and the solution first becomes singular and subsequently unphysical. The nonphysical solution in the nonlinear waves calculated using the dissipationless equations is usually remedied by the introduction of a hydrodynamic shock, and the associated shock viscosity and dissipation, to bring the solution back to the physically meaningful form.[18] In the case of a spiral wave, the dissipation effect introduced by the collective instabilities at the spiral arms also appears to be responsible for halting the wave-steepening process (§ 3.2).

In summary, an open spiral wave calculated by the linearized Eulerian equations can self-consistently sustain a phase shift without invoking dissipation. A self-consistent nonlinear and open spiral solution obtained with the Eulerian equations cannot be a time-steady solution but will always steepen with time.[19] A self-consistent nonlinear WKBJ wave, however, does not steepen, because it does not introduce a potential-density phase shift, which is what is responsible for driving the nonlinear wave-steepening process.

[18] Note that the formation of shocks by the steepening of free sonic waves is a different mechanism of shock formation from that of forced deceleration of supersonic flow when the flow encounters an intruding object. In the second case, the forced deceleration is determined mainly by the boundary conditions, instead of entirely by the nonlinearity in the wave equations as in the first case.

[19] This remains a conjecture at the present time and needs to be proved, possibly through an iterative solution of the nonlinear Eulerian equation set for the time development of an open and global spiral mode. Strongly supporting this conjecture, however, is the fact that the past non–self-consistent (Roberts 1969; Shu et al. 1973) or partially self-consistent (Levinson & Roberts 1981) WKBJ calculations have invariably found large-scale spiral shock solutions for finite amplitude spiral forcing. A phase shift between the potential and density spirals introduces exactly this element of local (i.e., within the same annulus) non–self-consistency between the forcing spiral potential and the response spiral density, in a globally self-consistent, open spiral wave solution.

REFERENCES

Antonov, V. A. 1962, Vestn. Leningr. Gos. Univ., 7, 135
Bahcall, J. N. 1984, ApJ, 287, 926
Balbus, S. A., & Cowie, L. L. 1985, ApJ, 297, 61
Bertin, G., Lin, C. C., Lowe, S. A., & Thurstans, R. P. 1989a, ApJ, 338, 78
———. 1989b, ApJ, 338, 104
Binney, J., & Tremaine, S. 1987, Galactic Dynamics (Princeton: Princeton Univ. Press)
Carlberg, R. G. 1986, in Nearly Normal Galaxies, ed. S. M. Faber (New York: Springer), 129
Donner, K. J., & Thomasson, M. 1994, A&A, 290, 785
Elmegreen, B. G., & Elmegreen, D. M. 1983, MNRAS, 203, 31
Elmegreen, B. G., & Elmegreen, D. M. 1989, in Evolutionary Phenomena in Galaxies, ed. J. E. Beckman & B. E. J. Pagel (Cambridge: Cambridge Univ. Press), 83
Elmegreen, D. M. 1981, ApJS, 47, 229
Fujimoto, M. 1968, in IAU Symp. 29, Nonstable Phenomena in Galaxies, ed. V. Ambartsumian (Yerevan: Armenian Acad. Sci.), 453
Gilbert, I. H. 1968, ApJ, 152, 1043
Gilmore, G., King, I. R., & van der Kruit, P. C. 1990, The Milky Way as a Galaxy, Saas-Fée Advanced Course Lecture Notes, No. 19 (Mill Valley: University Science Books)
Gurzadyan, V. G., & Savvidy, G. K. 1986, A&A, 160, 203
Jog, C. J., & Solomon, P. M. 1984, ApJ, 276, 114
Kalnajs, A. J. 1965, Ph.D. thesis, Harvard Univ.
———. 1971, ApJ, 166, 275
———. 1972, Astrophys. Lett., 11, 41
———. 1973, Proc. Astron. Soc. Australia, 2(4), 174
Kandrup, H. E. 1988, MNRAS, 235, 1157
Krall, N. A., & Trivelpiece, A. W. 1973, Principles of Plasma Physics (New York: McGraw-Hill)
Kulsrud, R. M. 1972, in IAU Colloq. 10, Gravitational N-Body Problem, ed. M. Lecar (Dordrecht: Reidel), 337
Lau, Y. Y., & Bertin, G. 1978, ApJ, 226, 508
Levinson, F. H., & Roberts, W. W., Jr. 1981, ApJ, 245, 465
Li, C. B., Han, N. K., & Lin, C. C. 1976, Sci. Sinica, 19, 665
Lin, C. C., & Lau, Y. Y. 1979, Stud. Appl. Math, 60, 97
Lin, C. C., & Shu, F. H. 1964, ApJ, 140, 646
Lin, C. C., & Shu, F. H. 1970, in Galactic Astronomy, Vol. 2, ed. H. Y. Chiu & A. Mureal (New York: Gordon & Breach), 80
Liszt, H. S., & Burton, W. B. 1981, ApJ, 778
Lubow, S. H. 1988, in Applied Mathematics, Fluid Mechanics, Astrophysics: A Symposium to Honor C. C. Lin, ed. D. J. Benney, F. H. Shu, & C. Yuan (Teaneck: World Scientific), 358
Lubow, S. H., Balbus, S. A., & Cowie, L. L. 1986, ApJ, 309, 496
Lynden-Bell, D., & Kalnajs, A. J. 1972, MNRAS, 157, 1
Lynden-Bell, D., & Ostriker, J. P. 1967, MNRAS, 136, 293
Lynden-Bell, D., & Wood, R. 1968, MNRAS, 138, 495
Mark, J. W.-K. 1974, ApJ, 193, 539
———. 1976, ApJ, 205, 363
Pfenniger, D. 1986, A&A, 165, 74
Roberts, W. W. 1969, ApJ, 158, 123
Rohlfs, K. 1977, Lectures on Density Wave Theory (New York: Springer)
Romeo, A. 1990, Ph.D. thesis, SISSA, Trieste, Italy
Sellwood, J. A., & Carlberg, R. G. 1984, ApJ, 282, 61
Shu, F. 1970, ApJ, 160, 99
Shu, F. 1992, The Physics of Astrophysics, Vol. 2 (Mill Valley: University Science Books)
Shu, F. H., Milione, V., & Roberts, W. W., Jr. 1973, ApJ, 183, 819
Shu, F. H., Yuan, C., & Lissauer, J. J. 1985, ApJ, 291, 356
Snow, C. 1952, Hypergeometric and Legendre Functions with Applications to Integral Equations of Potential Theory (NBS/AMS 19; Washington, DC: National Bureau of Standards)
Spitzer, L. 1978, Physical Processes in the Interstellar Medium (New York: Wiley)
———. 1987, Dynamical Evolution of Globular Clusters (Princeton: Princeton Univ. Press)
Spitzer, L., & Chevalier, R. A. 1973, ApJ, 183, 565
Stark, A. A. 1979, Ph.D. thesis, Princeton Univ.
Thomasson, M. 1989, Res. Rep. 162, Department of Radio and Space Science with Onsala Space Observatory (Chalmers Univ. Techn. Goteborg)
Toomre, A. 1964, ApJ, 139, 1217
———. 1969, ApJ, 158, 899
———. 1981, in Structure and Dynamics of Normal Galaxies, ed. S. M. Fall & D. Lynden-Bell (Cambridge: Cambridge Univ. Press), 111
van der Kruit, P. C., & Shostak, G. S. 1984, A&A, 134, 258
Vandervoort, P. O. 1971, ApJ, 166, 37
Weinberg, M. D. 1993, ApJ, 410, 543
Wielen, R. 1975, in Colloq. CNRS 241, La Dynamique des Galaxies Spirales, ed. L. Weliachew (Paris: CNRS), 357
Woodward, P. R. 1973, Ph.D. thesis, Univ. California, Berkeley
———. 1975, ApJ, 195, 61
Zhang, X. 1992, Ph.D. thesis, Univ. California, Berkeley
———. 1994, 24th DDA Meeting Abstracts, BAAS, 26(2), 1020
———. 1995a, IAU Symp. 169, Unsolved Problems of the Milky Way, ed. L. Blitz (Dordrecht: Kluwer), in press
———. 1995b, ApJ, submitted (Paper II)
———. 1995c, in preparation (Paper III)
———. 1995d, in Chaos in Gravitational N-Body Systems, ed. J. C. Muzzio (Dordrecht: Kluwer), in press