Xiaolei Zhang

Dynamical Evolution of Galaxies

Also of Interest

Origin of the Moon. New Concep. Geochemistry and Dynamics
Erik M. Galimov, Anton M. Krivtsov, 2012
ISBN 978-3-11-028628-1, e-ISBN (PDF) 978-3-11-028640-3

General Relativity: The most beautiful of theories. Applications and trends after 100 years
Carlor Rovelli (Ed.), 2015
ISBN 978-3-11-034042-6, e-ISBN (PDF) 978-3-11-034330-4,
e-ISBN (EPUB) 978-3-11-038364-5

Frontiers in Relativistic Celestial Mechanics. Theory
Sergei M. Kopeikin, 2014
ISBN 978-3-11-033747-1, e-ISBN (PDF) 978-3-11-033749-5,
e-ISBN (EPUB) 978-3-11-038938-8

Frontiers in Relativistic Celestial Mechanics. Applications and Experiments
Sergei M. Kopeikin, 2014
ISBN 978-3-11-034545-2, e-ISBN (PDF) 978-3-11-034566-7,
e-ISBN (EPUB) 978-3-11-037953-2

Xiaolei Zhang

Dynamical Evolution of Galaxies

DE GRUYTER

Mathematics Subject Classification 2010 and Physics and Astronomy Classification Scheme 2010
PACS: 98; MSC: 85-08, 85A05, 85A40

Author

Xiaolei Zhang, Ph.D.
George Mason University
Department of Physics and Astronomy
4400 University Drive
Fairfax, VA 22030
USA
xzhang5@gmu.edu

ISBN 978-3-11-052519-9
e-ISBN (PDF) 978-3-11-052742-1
e-ISBN (EPUB) 978-3-11-052544-1

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck
Cover image: Digital Vision/DigitalVision/thinkstock
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Preface

To explore the possibility of morphological transformation of galaxies over cosmic time had been a major motivation of Edwin Hubble, when he devised the galaxy classification scheme bearing his name (Hubble 1936). Galaxy observations in the recent decades have shown that Hubble's original conception of an evolution trend from the early to the late Hubble types, that is, from the more rounded (or bulgy) galaxies to the more flattened (or disky) galaxies, based on Sir James Jeans' theoretical framework (Jeans 1928), was untenable. The possibility of an evolution trend along the reverse direction, that is, from the late to the early Hubble types, on the other hand, has received growing observational support, and is currently an active area of research under the general heading of "secular morphological evolution of galaxies." The phrase "secular evolution" here refers to the slow and long-term transformation of galaxy morphology driven mostly by internal mechanisms.

In the title of the current book, we choose rather to use the phrase "dynamical evolution" because during the past few decades, the phrase "secular evolution" has often been associated with work on gas accretion in galaxies. The dynamical mechanism discussed in this work, on the other hand, is effective for the radial redistribution of both the stellar and the gaseous disk mass. When we use the phrase "secular evolution" in this book as a short-hand for "internal and dynamically driven morphological evolution of galaxies," it also has this broader meaning. We will show that long-term and, for all practical purposes, irreversible dynamical evolution in a nominally Hamiltonian system is in fact possible, when the system under consideration is of many degrees-of-freedom, is open, and is far-from-equilibrium. Such a system allows an *effective* singularity to develop when a self-organized global instability pattern emerges, which further generates an arrow of time.

Viewed in this light, the phrase "dynamical evolution" no longer appears to be an oxymoron[†] even without the assistance of resonant interactions: a complex Hamiltonian system can indeed exhibit dissipation-like, long-term evolution behavior, even though its local dynamics is nominally reversible. Here the paradox is resolved by the overwhelming volume of phase space occupied by the type of initial conditions for an

---

[†] The impression that a Hamiltonian system cannot exhibit true long-term evolution behavior probably results partly from the so-called Poincaré recurrence theorem (Poincaré 1890), which states that many dynamical systems will, after a sufficiently long time, return to a state very close to the initial state. One of the conditions under which the theorem was proved is the preservation of the volume of a phase space element along the particle orbit during the evolution of the dynamical system, which is indeed obeyed by all Hamiltonian systems (since they satisfy Liouville's theorem). However, for open and many degrees-of-freedom systems that are far from equilibrium (including disk galaxies), it can occur that the physical region a system explores expands indefinitely, even though the local phase-space density near the particle trajectory is conserved. This violates a second condition needed in the proof of the Poincaré recurrence theorem, that is, that of a finite upper bound on the total accessible phase space volume.

DOI 10.1515/9783110527421-202

open, complex, and nonequilibrium system, which, when evolved, would lead to the spontaneous emergence of global instabilities. The coarse-grained entropy is very low for almost any possible choice of initial condition for such a nonequilibrium many-body system, which leaves room for the eventual increase of coarse-grained entropy as a result of global-instability-induced secular evolution, even though the fine-grained entropy in such systems never changes, since all Hamiltonian systems obey Liouville's theorem.

Despite the growing observational support for secular evolution, among published texts on galaxy dynamics there has not been a systematic treatment of the dynamical foundation of the secular evolution process. Past theoretical investigations in galactic dynamics focused on the construction of self-consistent galaxy models in steady states – that is, building up the masses of galaxies employing passive and mostly regular orbits obtained under a rigid, applied potential. These practices, while being successful at obtaining equilibrium models which can then be compared with the observed galaxy mass and light distributions, have been found wanting in predicting the more-detailed characteristics pertaining to the morphology, kinematics, and especially the long-term evolutionary behaviors of galaxies.

The advent of the density wave theory in the early 1960s was a significant step in our understanding of the dynamical processes in disk galaxies containing large-scale spiral and bar patterns. However, a well-known result of the older version of the density wave theory, that of no wave/basic-state interaction except at the wave-particle resonances for quasi-steady patterns, prohibited any serious consideration of a significant change of the morphologies of galaxies over their lifetime.

The original research described in this monograph was among the first to reveal that the suite of classical mechanical approaches traditionally employed in galactic dynamics is inadequate for the treatment of so-called "emergent" phenomena in open, many degrees-of-freedom, nonlinear, and far-from-equilibrium gravitational systems. The essence of these phenomena is the nonequilibrium phase transition process, whose proper treatment requires new dynamical tools properly matched to the emergent new physics.

The idea that new concepts and methods need to be devised to approach the so-called "problems of complexity" in nonlinear and many degrees-of-freedom systems has been around for quite a while. The spirit of this school has been eloquently summarized by the Nobel laureate solid-state physicist P.W. Anderson as "more is different" (Anderson 1972). Specifically, in dealing with complex systems, one can encounter paradoxical situations where "the whole is more than the sum total of its parts," in the sense that one cannot expect to simply pile together the component molecules to make a biological entity "come alive" on its own. An element of history in the development of the organism, which includes the sum total of its past complex interactions with its environment, as well as interactions among its constituent parts, will have to be taken into account. The history of interactions and the history of growth are partly what give an organism its structure and its dynamical vitality. The mystery of the "life-force" therefore originates in the self-organization process itself, the dynamics of which is irreversible and "emergent" from the underlying simple and reversible dynamical interactions.

In this monograph, we will explore the detailed and often subtle aspects of the self-organization process in the context of the dynamics and evolution of galaxies, which are assemblies of many stars plus the interstellar medium, often organized into coherent large-scale patterns now understood to be density waves. These waves, or wave modes, as we will show, not only give galaxies their breathtaking appearance, but also perform important functions in facilitating the long-term morphological evolution of their parent galaxies. The dynamical concepts and tools used for understanding the roles of these patterns, such as symmetry breaking, collective dissipation, entropy production and export, etc., have their analogues in other disciplines of physical and biological sciences as well. Of particular relevance are the studies of "dissipative structures" by Nobel laureate I. Prigogine and associates (Prigogine 1969, 1980; Glansdorff & Prigogine 1971; Nicolis & Prigogine 1977), as well as the studies of "synergetics" (Haken 2004; Fuller 1975). The author was fortunate to have met and discussed her work on applying the theory of dissipative structures to the problem of galaxy evolution with Prof. Prigogine in the year 2000, three years before his sad passing at the age of 86, on an occasion to give a physics department colloquium talk at the University of Texas at Austin, through a special invitation from Prof. Prigogine.

The current monograph serves to fill a void in the published texts on the internal dynamical mechanisms enabling the secular morphological evolution of galaxies. The past works on secular evolution were mostly performed within two kinds of theoretical framework: (1) treating the *gas* component, or the interstellar medium in galaxies, as the major driver of secular evolution (Kormendy 1979; Kormendy & Kennicutt 2004 and the references therein). This approach has its heritage in the linear response theory of nonequilibrium thermodynamics (de Groot & Mazur 1962 and the references therein), which is closely linked to the Navier–Stokes equation set, which in turn is obtained from the first-order expansion of the collisional Boltzmann equation in terms of the ratio of collisional mean free path versus the system size. Employing linear nonequilibrium thermodynamics gives us the various well-known transport processes such as thermal conduction, viscous dissipation, etc.. (2) When secular evolution in the *stellar* component of the galactic disk was considered, it was invariably in the context of the *resonant* interaction of the density waves and the disk stars (Sellwood 2014 and the references therein). This second approach also has a long heritage (i. e., the Lagrange–Laplace investigation of the stability of the solar system), and contemporary research in planetary science offers many good examples of its application.

The self-organized global instability patterns that emerge spontaneously in complex, many degrees-of-freedom, nonlinear, and far-from-equilibrium systems, however, have innate dynamical mechanisms that go far beyond the physical processes addressed by either the linear nonequilibrium thermodynamics of transport phenomena or the passive resonant interactions between an external driving agent and the orbits of the individual particles of the system under concern. The operation of these new dynamical mechanisms is closely linked to the *correlated fluctuations* that emerge and become amplified when a nonequilibrium phase transition takes place in open, many degrees-of-freedom systems. These correlated fluctuations violate one of the key assumptions in the derivation of the collisional Boltzmann equation (through, e. g., the rigorous BBGKY procedure, see Section 7.1.1 for further discussion), that is, the ansatz of "molecular chaos" which mandates that the fluctuations of component particles be totally uncorrelated. It is indeed this very element of correlated fluctuations that introduces the "history" element and gives "life-force" to a self-organized pattern. These correlations are anything but the random processes assumed in the derivation of the Boltzmann equation.

In the past galactic dynamics work, the most often employed fundamental kinetic equation is the collisionless version of the Boltzmann equation, also known as the Vlasov equation or the stellar dynamical equation. This version of the Boltzmann equation has served as the foundation of the early density wave theory explorations. Both versions of the Boltzmann equation, as well as the moment-equation derivatives of the collisional version of the Boltzmann equation (such as the zeroth-order Eulerian equation set, or the first-order Navier–Stokes equation set), are lacking the crucial ingredients that allow correlated fluctuations to spontaneously emerge – even though they can have indications of the necessity of such additional processes through the formation of unstable modes, which are not able to be stabilized within the range of validity of these (approximated) theories themselves. Thus a full understanding of the self-organization process cannot be based entirely on these *approximated* equation sets, even though they have been proven to serve our previous purposes of understanding the equilibrium or close-to-equilibrium phenomena.

The new approach we will adopt in the current work combines the analytical, N-body simulational, as well as observational aspects to bring about a new synthesis. The methodology we develop here may be generalized to the studies of self-organization phenomena in other physical sciences as well: this is so especially because most of the past studies of spontaneous symmetry-breaking were conducted in a model Lagrangian or Hamiltonian context, rather than self-consistently from first principles. Therefore the important process of interaction of the system with the environment, which leads to the irreversible long-term evolution of the environment, was ignored in most of these previous studies. We will show in the main body of this monograph that, as a result of collective processes due to unstable density wave modes, new kinds of closure relations are arrived at when the *global self-consistency* requirement is enforced at the quasi-steady state of the wave modes. The emergence of these new dynamical-equilibrium relations signals the failure of the classical mean-field approaches commonly used for treating the physics of close-to equilibrium (linear and quasi-linear) regimes. Furthermore, the need for replacement of the older governing laws with the new emergent closure relations can be understood by the fact that the fundamental nature of galactic density wave modes is akin to plasma collisionless shocks. It is well-known that traditional differential and mean-field treatments break down at the onset of nonequilibrium phase transitions or at the collisionless-shock discontinuities (Balogh & Treumann 2013), and interparticle correlations become the dominant factor governing the new physical processes that result, which the mean-field theories are known to ignore. We will show that the integral manifestation of secular shock dissipation at the density wave crest is *a characteristic radial distribution of the azimuthal phase shift between the potential and density wave patterns*, which allows the density-wave-assisted galactic secular evolution to proceed in a globally self-consistent and quasi-stationary manner.

The current text is based mainly on the original work by the author and her collaborators in the past two and a half decades, starting from her Ph.D. dissertation work at the University of California, Berkeley in the late 1980s, through her postdoc years at the Harvard-Smithsonian Center for Astrophysics, and up until her most recent collaborative work with Prof. Ron Buta of the University of Alabama on the observational verifications of theoretical predictions. Most of the original contributions described in this book had previously appeared in refereed journals and conference proceedings. When deciding what to include in this book from the author's previously published work, priority is given to including enough details of derivations so that a reader can follow the logical flow of the argument without needing to refer back to the original papers; whereas more extended applications to large samples of galaxies are only briefly mentioned, and interested readers can find these in the referenced papers. Permissions to use material from these previously published works have been obtained from the original publishers. Due to the differing publisher requirements, for some works the permissions are printed explicitly (i.e. for works published in the AAS journals associated with the IOP), some referenced with its DOI (i.e., for work published in the PASP journal), some referenced with the conference proceedings citation link, and still others with journal citation together with a Science Direct link (i.e., for works published in the Elsevier journals). A number of N-body simulations are rerun with the same global parameters as the previously published ones for the purpose of generating higher-quality graphics to meet the publisher's requirement. Even though the random number seeds for initial condition assignment will not be exactly duplicated, this does not affect the global modal characteristics obtained. Like the subject of the collective effects itself, here we hope the whole of the completed manuscript will be more than the sum total of its parts, through the context provided by the organization of the material, and the logical links among the different components. Additional published results of other researchers are discussed as relevant, though by the nature of this monograph a thorough survey has not been attempted. The study of secular evolution of galaxies is still an emerging discipline, so new understandings and new results are expected to occur as the field develops and matures. It is the hope of the author though, that through this monograph a coherent framework is provided for understanding the dynamical principles underlying the secular morphological evolution of galaxies. This in turn should stimulate and facilitate further advances in this area.

Throughout the development of this work numerous colleagues have provided support, help, and feedback in various forms. The author would like to thank especially the following individuals for helpful interactions: Ron Allen, Hong Bae Ann, Lia Athanassoula, Steve Balbus, James Binney, Alberto Bolatto, Ron Buta, Gregory Bothun, Gene Byrd, Francoise Combes, George Contopoulos, Warrick Couch, Victor Debattista, Tim De Zeeuw, Bruce Elmegreen, Bob Ehrlich, Henry Ferguson, Jackie Fischer, Roger Fux, Ken Freeman, Daniel Friedli, Oleg Gnedin, Rosa González-Lópezlira, Preben Grosbol, Jin-Lin Han, Carl Heiles, Richard Hills, Luis Ho, Paul Jaminet, Agris Kalnajs, Sasha Kashlinsky, Rob Kennicutt, Ivan King, John Kormendy, Eija Laurikainen, Youngung Lee, Chia-Chiao Lin, Xiao Liu, Steve Lubow, Shude Mao, Eric Martínez-García, Colin Masson, Stacy McGaugh, Richard Miller, Colin Norman, Panos Patsis, Daniel Pfenniger, Alice Quillen, Alessandro Romeo, Heikki Salo, Paolo Salucci, Shobita Satyapal, Paul Schechter, James Shombert, Jerry Sellwood, Frank Shu, Antony Stark, Mario Tafalla, Peter Teuben, Magnus Thomasson, Scott Tremaine, Alar Toomre, Keiichi Wada, Simon White, Jack Welch, Robert Wilson, Mel Wright, Anatoly Zasov, as well as several referees of the journal articles as well as the current book. She thanks the management at the George Mason University for the Affiliate Faculty position which allowed her the freedom to pursue independent research. She also thanks her husband Brian and her extended family for the support and encouragement throughout the years. Finally, she thanks the staff at the De Gruyter Publishers, especially acquisition manager Chao Yang and editorial managers Astrid Seifert & Nadja Schedensack, as well as production editor Sabina Dabrowski and the production team, for their valuable help in making this monograph possible.

Contents

1 Introduction

During the past few decades, the dominant view was that the structural properties of galaxies remain largely unchanged over time, unless galaxies were perturbed by violent events such as major or minor mergers (Toomre & Toomre 1972). This view is particularly favored in the currently popular hierarchical clustering/Lambda cold dark matter (LCDM) paradigm of structure formation and evolution (Ostriker & Steinhardt 1995; Liddle 2003 and the references therein; Mo, van den Bosch, & White 2010 and the references therein). Even though the role of mergers during the early phase of galaxy assembly at high redshifts could indeed be important, growing evidence has shown that at least since redshift $z \sim 1$ the rate of merger appears to have been significantly reduced (Cohen 2002; Conselice et al. 2003; López-Sanjuan et al. 2009. See also Conselice et al. 2016 which indicates that the role of merger was already insignificant starting from $z = 2$, when the universe was only a quarter of its current age), and subsequent galaxy morphological transformation is likely to be dominated by the slower internal secular evolution process.

In the late 1970s, photometric and kinematic evidence in the bulges of late-type galaxies, as well as hints from N-body simulations of barred galaxies which incorporated a dissipative gas component, prompted several investigators to speculate that the morphology of late-Hubble-type disk galaxies may be transformed into that of earlier Hubble-types by gas accretion under barred potential (Kormendy 1979, 1982; Combes & Sanders 1981). These initial speculations have since been developed into one version of the secular evolution scenario, which emphasizes the role of dissipative gas accretion in the formation of the so-called pseudo bulges (or disky-bulges) in late-type galaxies (Kormendy & Kennicutt 2004 and the references therein). More careful examination, on the other hand, has since demonstrated that the relevant gas-dynamical process is not powerful enough to transform the morphology of intermediate- to early-Hubble-type disk galaxies, considering the observed fraction of gas and the rate of star formation in these galaxies (Kormendy & Kennicutt 2004).

Beginning in a series of papers published in the late 1990s (Zhang 1996, 1998, 1999), the present author showed that the stellar component, which constitutes the major portion of the galaxy mass in all but the very late-type galaxies, in fact is the main contributor of the secular morphological evolution of galaxies, through a collective interaction process between the spontaneously formed density wave modes and the basic state of the galactic disk[†]. This process had previously been overlooked in galactic dynamical analyses employing the stellar dynamical equation (which is the same as the collisionless Boltzmann equation, or the Vlasov equation), as well as

[†] The term *basic state* refers to the axisymmetric galaxy disk from which a density wave mode emerges. It is usually specified as the radial distribution of the galaxy-disk surface density, circular velocity, and velocity dispersion.

DOI 10.1515/9783110527421-001

the moment equation descendant of the collisional Boltzmann equation, the Eulerian equation set, because the Boltzmann equation itself was derived in a context that ignored collective interactions and interparticle correlations (see the detailed derivation of Boltzmann equation in the Appendix of this monograph), thus is not suitable for applications that focus on the self-organization and collective dissipation processes such as the study of the secular morphological evolution of galaxies mediated by spontaneously formed density wave modes.

In the course of presentation of this monograph, we will make extensive use of the phrases “collective dissipation” or “collective effects.” These phrases describe a general tendency for nonequilibrium, complex systems to spontaneously form global patterns, which lead to emergent new dynamics through *coherent* interactions of their component parts. Other phrases used in this context in the literature, ranging in applications from condensed matter physics, fluid dynamics, economics, biological and social sciences, include “cooperative effect,” “coupling and interaction among the degrees-of-freedom of a complex system,” and “synergetics.”

In the case of a galaxy possessing a density wave pattern, the traditional approach employed in constructing galaxy models displays a top-down type of organization, meaning that the stellar orbits in such galaxy models respond *passively* to an *applied* potential. Though still adequate for the study of the density wave modal emergence phase, the top-down approach is entirely powerless to deal with the self-consistent evolution of the galaxy basic state together with the corresponding self-limiting density wave modal set. The density wave modes of galaxies are able to maintain their quasi-steady amplitude *only at the expense* of the dissipative secular evolution of the mass distribution of the basic state of the galactic disk. *So the long-term survival of the mode and the secular evolution of the basic state of the disk are closely linked*.

The secular evolution process, which slowly transforms the morphology of a galaxy over its lifetime, could naturally account for the observed properties of the great majority of physical galaxies, if both the stellar and the gaseous accretion processes are taken into account. As an emerging paradigm for galaxy evolution, its dynamical foundation has been gradually established in the past few decades, and its astrophysical consequences are just beginning to be explored. In the current monograph, we seek to establish that the secular evolution scenario provides a coherent framework for understanding the extraordinary regularity and the systematic variation of galaxy properties along the Hubble sequence.

1.1 Observational Background

Observational data on the characteristics of galaxies have been accumulating for more than a century, but evidence for the importance of internal secular evolution processes has only become apparent in the recent decades, largely as a result of the increased sensitivity, spectral coverage, and angular resolution of the space and ground-based telescopes. The supporting evidence ranges from the characteristics of individual galaxies both in isolated and in group/cluster environments, as well as the statistical properties of galaxy populations as a whole.

The most detailed characteristics of an individual galaxy come from the observation of our own Galaxy, the Milky Way. It has long been known that the observed kinematics of the different age groups of stars in the Milky Way Disk differ systematically, manifesting as the well-known age–velocity dispersion relation of the solar neighborhood stars (Wielen 1977). It is obvious that there exists a dynamical mechanism that heats the Disk stars secularly as they age, which operates smoothly across the entirety of a Hubble time (Gilmore, King, & van der Kruit 1990). Furthermore, the stellar population in the Galactic Bulge region has the well-known stratified distribution, with younger populations closer to the Galactic central region. This stratification trend also extends to the Thick and Thin Disks away from the Bulge region (Gilmore et al. 1990). These distributions also hint at a secular evolution origin for their formation.

Recent deep surveys have found that galaxies in the general field environment similar to that occupied by the Milky Way have undergone significant morphological transformation over the cosmic time. It is found that more field galaxies are of earlier Hubble types in the nearby universe than at higher redshifts (Lilly et al. 1998). There exist also the so-called faint-blue galaxies, which are in fact L_* galaxies having luminosities and sizes similar to the Milky Way, which are found at the intermediate redshifts, but which have all but disappeared in the nearby universe (Ellis 1997). Since the total number density of galaxies of all Hubble types have not evolved significantly between redshifts $z = 1$ and $z = 0$ (Cohen 2002; Conselice et al. 2016), and the merger fraction since $z = 1$ is low (Conselice et al. 2003; López-Sanjuan et al. 2009), the most likely explanation for these observed statistical differences in galaxies between the higher redshifts and the nearby universe is the internal morphological (as well as the accompanying stellar population and color) evolution of galaxies.

Dense cluster was the environment where morphological transformation of galaxies was first hinted at through the so-called Butcher–Oemler effect (Butcher & Oemler 1978a, 1978b). When it was discovered, the Butcher–Oemler effect referred to the bluer colors of galaxies in dense clusters at the intermediate redshifts, compared to similar-density clusters in the local universe which contain mostly red, early-type galaxies. Subsequent Hubble Space Telescope (HST) observations (Couch et al. 1994; Dressler et al. 1994) have been able to resolve the morphology of these intermediate-redshift Butcher–Oemler galaxies, and show that they are mostly late-type disks; therefore the Butcher–Oemler effect is now considered not only a color evolution effect but also a morphological transformation effect. Because of the high-speed nature of encounters of galaxies in dense clusters, mergers are known to be infrequent in the virialized regions of dense clusters, so once again internal dynamical mechanisms are likely to have played a prominent role in the transformation of Butcher–Oemler galaxy morphologies between the intermediate redshifts and the nearby universe.

In the late 1970s, from observational studies of late-type galaxy bulges, J. Kormendy concluded that many of these bulges appear to have disk-like morphological and kinematic characteristics. He subsequently proposed (Kormendy 1979, 1982) that these late-type bulges are evolutionarily linked to disks, and named these bulges "pseudo bulges" to distinguish them from "true" (or "classical") bulges in early-type disk galaxies, which were believed to be formed either primordially, or else through galaxy mergers.

Since it was at the time universally believed that stars "do not dissipate", or that their orbital behavior is adiabatic under a steady potential (see, e. g. Binney & Tremaine 2008), Kormendy proposed that the formation of pseudo-bulges and the associated mild Hubble-type evolution among late-type galaxies were made possible through gas accretion in barred potentials, and these inwardly accreted gas subsequently formed new bulge stars. This proposed pathway for secular evolution, though very influential since then, has a number of problems from comparison with the observed properties of galaxies, as well as from detailed analysis of the physics of the accretion process. First of all, as emphasized by Andredakis, Peletier, & Balcells (1995), the continuity of galaxy properties across the entire Hubble sequence as highlighted for example by the smooth variation of the Sersic index n in fitting the bulge surface density profile, indicates that there is not an apparent break in the formation mechanism between late-type and early-type bulges. However, because of the paucity of gas compared to stars in most galaxies of earlier Hubble types, there simply is not enough of a reservoir of gas to build up the bulges of galaxies such as our own (of type Sb), which has its Bulge mass comparable to the Disk mass, not to say for galaxies of even earlier Hubble types. Secondly, bulges of intermediate- to early-type galaxies, including our own, consist mostly of stars of very old ages (Jablonka, Gorgas, & Goudfrooij 2002), despite also possessing a stratified color distribution (Wirth & Shaw 1983), and could not have been built up by an *extended* process of secular *gas* accretion which subsequently formed stars. A good fraction of these apparently old bulges have stellar kinematics which are also rotation-dominated and are related to the kinematics of their disks, hinting at a secular evolution origin for their formation (Kormendy & Kennicutt 2004 and the references therein). Therefore, a dynamical mechanism which could channel the pre-existing disk stars from the outer to the inner region of a galaxy is needed to explain these rotation-dominated older bulges. Thirdly, recent results of the ATLAS^{3D} team (Cappellari et al. 2013) have shown that the morphologies and other internal properties of spirals, S0s, *as well as disky ellipticals*, form a continuous trend of evolution, which also coincides with the trend of aging of stellar population of their galactic disks. This provides further support for a unified formation history of the majority of the Hubble sequence galaxies through internal processes, which necessarily involves the participation of the stellar component. Fourth, recent results of the COSMOS team (Cisternas et al. 2011a, 2011b) showed that the establishment and evolution of the well-known black-hole-mass/bulge-mass correlation since $z = 1$, on the bulge-building side, was mainly due to the radial mass accretion process on pre-existing *stellar* disks, which were already in place by $z = 1$. The COSMOS team had also excluded merger as a significant contributor to the build-up of the black-hole-mass and bulge-mass correlation since $z = 1$. This further motivates a serious consideration of the internal secular mass redistribution process that *emphasizes the role of stellar mass accretion in bulge building*[†]. Finally, as we will show later in this monograph, even the gas contribution to the secular mass accretion process is through the same collective dissipation process involving large-scale, coherent density wave modes, rather than through the particle-level viscosity commonly attributed to gas dissipation. The mean free path of the collision/scattering of the galactic molecular clouds in a density wave crest region is similar to that of the stars, and both components should be considered together in supporting a star-gas combined two-fluid instability. This common mean free path of the two-fluid instability is many orders of magnitude larger than the microscopic collisional mean free path of the gas. So the roles of stars and gas are now parallel, rather than distinct, in the collective process that leads to the secular morphological transformation of galaxies.

From the above arguments, plus more that will be discussed in the main text, we see that internal secular evolution processes involving the participation of both the stellar and gaseous components appear to be crucial to the transformation of galaxy morphologies along the Hubble sequence at least since $z = 1$, and possibly since $z = 2$. The admittance of the stellar component into the secular mass redistribution process in galaxies, however, cannot be accomplished without a major breakthrough in the dynamical foundation of the theories of galactic structure and evolution.

1.2 Theoretical Background

The ultimate engine for the morphological evolution of galaxies is in a self-gravitating system's tendency to increase its entropy with time. It has long since been known (e. g., Antonov 1962; Lynden-Bell & Wood 1968) that the direction of this entropy evolution for self-gravitating systems is toward configurations with ever more centrally concentrated cores, together with increasingly extended outer envelopes.

[†] Incidentally, the known weaker correlation between the masses of the *late-type* bulges and their corresponding central black holes at any given epoch (Kormendy, Bender, & Cornell 2011) may be on the one hand a result of the quicker onset of accretion events in certain black-hole/active galactic nuclei (AGN) accretion disks compared to the onset of significant mass accretion in their outer galactic disks for bulge-building since the local dynamical time scale is shorter for the smaller AGN accretion disks; on the other hand some galaxies may form shallow pseudo-bulges without a deep central potential well to form a nuclear AGN accretion disk and to build a central black hole, during the early stages of a late-type galaxy's life. So the weaker correlation for late-type bulges and black holes reflect more the inhomogeneity of the galaxy formation process (i. e., the inhomogeneity of the *initial conditions of secular evolution*), rather than the inhomogeneity of the secular mass flow process itself.

Even with the above understanding, it was long held that the intrinsic speed of this evolution process is extremely slow. This slowness is due partly to the well-known pressure and angular momentum barriers in the galactic systems. Through an order of magnitude calculation, one can show that the natural speed of galactic evolution through microscopic transport processes results in a time scale for energy and angular momentum redistribution which is several orders of magnitude longer than the age of the universe (Zhang 1992).

It is clear that in order for secular morphological transformation of galaxies to be a relevant process in the observed history of the universe, so as to explain statistical evolution of galaxy properties with redshifts, dynamical mechanisms other than diffusion or dynamical friction have to be identified. An analogy here is the atmospheric heat flow (or Rayleigh–Bénard convection) problem (Kreuzer 1981). It is well known that when the natural speed of heat conduction is too slow in an atmospheric layer possessing a temperature gradient, an organized macroscopic flow pattern, facilitated by hexagonal convection cells, will spontaneously develop when the temperature gradient exceeds a certain threshold. These convection flow patterns greatly accelerate the speed of reducing the original temperature nonequilibrium in the atmospheric distribution, compared to the speed due to the conduction/diffusion process alone. As we will demonstrate in subsequent text, the new dynamical mechanism in galaxies operates in a similar fashion to the convection process in atmospheric flow. This mechanism will be shown to be closely related to the emergence of density wave patterns (or more specifically, unstable density wave *modes*) in galaxies, and to the maintenance of these modes as a dynamical equilibrium state, balanced by the opposing tendencies of the spontaneous growth of the unstable mode and the irreversible local dissipation process, with the latter process both damps the growth of the wave mode and simultaneously leads to the secular evolution of the basic state of the galactic-disk mass distribution.

Bertil Lindblad was the first astronomer to come up with the idea that spiral patterns in galaxies may be waves of density enhancement. In the 1950s, he made a series of numerical studies using an assembly of point-particles to simulate the appearance of spiral galaxies, but failed to obtain self-sustained patterns. C. C. Lin and F. Shu (1964, 1966) were the first to succeed in obtaining linear perturbative solutions of self-consistent spiral patterns in the tightly wrapped, or the so-called WKBJ (which stands for Wentzel, Kramers, Brillouin, & Jeffries) regime of the density wave parameter space, and the same WKBJ approximation allowed them to simplify the global problem of the large-scale wave pattern on a galaxy disk into a local problem of plane waves. At almost the same time, A. Toomre (1964) obtained the local stability criterion for axisymmetric oscillations in the disk, and A. Kalnajs (1965) independently obtained the dispersion relation for tightly wrapped waves as obtained by Lin & Shu (1964). In Gilmore et al. (1990) Figure 9.3, an illustration originally due to Lin (1967) summarizes the essence of the density wave approach as was practiced at that time: (a) a spiral-formed potential perturbation is introduced onto (i. e., superposed on) the differentially rotating basic state of the disk. (b) This spiral-formed potential perturbation requires a spiral-formed density perturbation in the stellar and gaseous disks to support it, through the Poisson equation (under the local WBKJ approximation, this spiral-density perturbation needed by the Poisson equation is exactly in-phase, or is proportional to, the spiral-potential perturbation). (c) On the other hand, the spiral-formed potential perturbation also leads to the perturbation in stellar orbits and in gas streamlines, which also leads to perturbed spiral-density distributions in the stellar and gaseous disks. (d) By enforcing the equality of the density perturbations from (b) and (c), a so-called *dispersion relation* is obtained, which guarantees (in this case, under the linear and local approximation) the self-consistency of the spiral wave solution.

Subsequent workers began testing the predictions of the newly developed mathematical theories of the density waves against observations, including the observed features of the density wave arms in our own Galaxy (Lin, Yuan, & Shu 1969). These comparisons were limited both by the then poorly known details of the Galactic spiral structure, the poor spatial resolutions obtainable in external galaxies, as well as by the intrinsic limitations of the WKBJ solution itself. When the spiral density wave theory was first proposed, Lin and Shu had sought a *neutral wave solution* (similar to a soliton) that retains a steady amplitude. The lowest-order WKBJ solution is indeed a neutral-wave solution. Toomre (1969), on the other hand, found that the Lin–Shu density waves in fact possess a radial group velocity and (the short-trailing wave branch) will inevitably propagate toward the central region of a galaxy and be absorbed at the inner Lindblad resonance (ILR). There is thus the need to re-generate these transient spiral waves. Lynden-Bell & Kalnajs (1972) discovered that the trailing form of spiral density wave, which is the predominant form among observed wave patterns in galaxies, actually transports angular momentum outward through the combined actions of gravitational and advective torques. Since a density wave rotates slower, or has negative angular momentum density with respect to the basic state of the galactic disk, inside the corotation radius[†], the removal of angular momentum by the wave from the inner disk and the transport of it by the same wave to the outer disk, will promote the spontaneous growth of a density wave train. This was the reason that Lynden-Bell and Kalnajs had named their 1972 paper "On the generating mechanism of spiral structure." They concluded in the same paper, however, that a steady wave exchanges angular momentum with the basic state of the galactic disk only at the wave-particle resonances, therefore the wave maintains a *constant* angular momentum flux *en route* of the outward angular momentum transport between the

[†] The corotation radius is a radial location where the density wave pattern rotates at the same speed as the underlying differentially rotating disk matter. For a regular disk galaxy which has higher angular speed in the inner disk and progressively lower angular speed in the outer disk (i. e., a galaxy that possesses *differential rotation*), the corotation radius (designated as r_{co} or CR in the following text) of the dominant mode of the density wave usually occurs somewhere in the mid-disk.

inner and outer Lindblad resonances[‡]. Thus the conclusion from these early density wave studies is that the major portion of the basic state of the galactic disk does not exhibit secular mass redistribution if it contains a density wave of steady amplitude. Lynden-Bell and Kalnajs, like most of their peers at the time, sought the generating mechanism of the wave trains and any associated secular effect only in the resonant interaction between the waves and the particles.

For growing wave disturbances as a result of its continued outward angular momentum transport, there is the need to counter the growth tendency in order to clamp the wave amplitude at a finite value, that is, there is the need of a damping mechanism. Kalnajs (1972), and Roberts & Shu (1972) proposed that the gaseous density waves often found to accompany stellar density waves in galaxies could serve to damp the growing stellar wave, owing to the intrinsic dissipation in the gas component and the resulting azimuthal phase shift between the stellar and the gaseous density wave patterns. However, the asymmetric roles assigned to stellar and gaseous waves meant that gas will be preferentially accreted to the galaxy center compared to the distribution of stars, contrary to what is observed (i. e., early type galaxies, considered to be in the late stages of secular morphological evolution, are known to contain far less gas than late-type galaxies which are the precursors of the secular evolution process). The gas damping mechanism also has difficulty in explaining the so-called anemic galaxies, which have coherent spiral patterns but little or no gas. It seems highly unlikely that the anemic spirals are stabilized through a different dynamic mechanism than those operating in gas-rich disk galaxies.

The problem of damping of the growing wave amplitude to achieve quasi-steady pattern was exacerbated after the discovery of over-reflection mechanisms near the corotation region, that is, the so-called WASER (which stands for Wave Amplification by Stimulated Emission of Radiation) mechanism by Mark (1974, 1976), which is responsible for the amplification of the normal spiral patterns, as well as the SWING mechanism by Toomre (1981) and Zang, which is responsible for the amplification of bar-like patterns, following the earlier work of Goldreich & Lynden-Bell (1965) and Julian & Toomre (1966). These over-reflection mechanisms at corotation, coupled with the wave-reflection/transmission mechanisms at the galaxy center, set up feedback loops for the radially propagating wave trains in the galactic "resonant cavity" between the central region and the corotation region, with a positive amplification factor for each round-trip of wave propagation. Figure 1.1 presents an illustration for the relatively milder version of the over-reflection mechanism WASER

‡ An inner Lindbald resonance is a galactic radial location where the pattern speed of the density wave Ω_p, the circular speed of the particles according to the galactic rotation curve Ω, and the local epicycle frequency κ satisfy $\Omega_p = \Omega - \kappa/2$ (for a two-armed pattern, which is the predominant form among the observed grand-design density wave patterns). Similarly, an outer Lindblad resonance is where $\Omega_p = \Omega + \kappa/2$. For the dominant pattern of the density wave the inner Lindblad resonance is usually located in the central bulge region if it exists at all, and the outer Lindblad resonance is usually located in the outer disk region external to the corotation radius of the same wave train or mode.

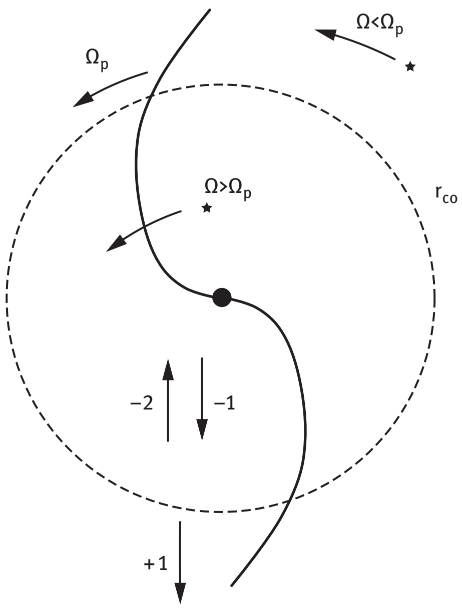

Figure 1.1: Schematic of the wave propagation and amplification in the galactic resonant cavity to form infinitely-growing density wave modes, through the WASER mechanism (Mark 1974, 1976). The symbols Ω and Ω_p denote the galaxy (or disk matter) circular/angular speed and wave pattern speed, respectively. The numbers next to the arrows indicate the normalized amount of wave-train angular momentum relative to the basic state (adapted from Zhang 2016).

(Mark 1974, 1976). For the SWING mechanism (Toomre 1981), the over-reflection factor can be significantly bigger than the factor of 2 illustrated here. The SWING and WASER mechanisms differ by whether the outward-propagating wave train toward corotation is that of the leading or the trailing type, respectively, produced in turn by either the inward-propagating trailing wave tunneling through the central region of the galaxy when no Q-barrier[†] exists, and emerging as a leading wave; or else by the inward-propagating trailing wave being reflected by the Q-barrier when one exists, which produces an outward-propagating trailing wave. The SWING mechanism tends to produce bar-like modes, and the WASER mechanism spiral modes. In this way the superposition of radially oppositely propagating density wave trains produces indefinitely growing *modes* (Lin & Lau 1979 and the references therein).

In addition to the approach for obtaining a modal solution through the superposition of oppositely propagating wave trains, the spontaneously growing spiral/bar modal solutions can also be obtained directly, through solving the density wave perturbation as a normal mode problem under the boundary condition of a given basic state (Lin & Lau 1979). Leaving the general problem of the wave-damping mechanisms undecided, Lin and collaborators went on to solve a complete series of modal morphologies under systematically varying basic state boundary conditions, and the resulting modal series largely reproduce the spiral morphology variation along the Hubble sequence, so the density wave pattern morphologies can now be

[†] A Q-barrier is a rapid rise in stellar velocity dispersion, normally occurs in the galactic bulge region, that reflects an inward-propagating density wave train into an outward-propagating wave train.

tied to the corresponding basic state characteristics of the parent galaxies (Bertin et al. 1989a, 1989b), achieving in part the explanation of the observed correlation of these two sets of characteristics which underlies the original Hubble classification scheme.

Throughout the first three decades of density wave studies, the possibility that quasi-steady density wave modes could lead to significant secular morphological evolution of the basic states of galaxy disks was never seriously discussed, even though mild secular changes in certain over-stable basic states possessing *transient* spiral waves were found by, e. g., Sellwood & Carlberg (1984). This impression of the immutable-basic-state for the majority of observed galaxies was further supported by the well-known classical result of the conservation of the Jacobi integral for a single stellar orbit under an enforced, steady, spiral or bar potential (see, e. g. Binney & Tremaine 2008), otherwise known as the adiabatic condition of the stellar orbit. Other relevant studies in this context include that of Goldreich & Tremaine (1979), as well as Goldreich & Nicholson (1989), which found that there is no wave/basic-state interaction in the Eulerian fluid solution up to second order of nonlinear approximation. Bertin (1983) reached a similar conclusion that any potential secular evolution of the basic state induced by the spiral wave happens on a time scale much longer than the age of the universe.

However, these classical work on the density wave theories ignored important collective effects responsible for the long-term evolution of galaxies. It has long been known that the relaxation and dynamical evolution of physical systems governed by long-range forces are determined chiefly by collective effects. The physical behavior of such systems is not completely reflected in the result of passive orbit calculations. Collective effects were known to enhance the speed of relaxation in many plasma systems (Kulsrud 1972; Balogh & Treumann 2013 and the references therein). Collective effects in the context of disk galaxies invariably involve local or global instabilities, and they operate by forcing a particle to systematically “collide with” or scatter off an aggregate of other particles collected together by the instabilities. The effective impact parameter in such a “collision” process is of the order of the size of the instability-feature formed. In effect, besides experiencing the smooth part of the potential, a particle which participates in a collective process also bounces off the local short-to-intermediate-range grainy scattering-potential. These collisions or scatterings are not random but rather are highly coordinated, and in the case of disk galaxies they enable the formation and stabilization of global density wave modes, as well as the long-term secular morphological evolution of their parent galaxies.

A collective dissipation mechanism responsible for the secular evolution of the disks of spiral galaxies was first proposed and analyzed in Zhang (1996), based on the more qualitative studies in Zhang (1992). It was shown that there exists a characteristic radial distribution of an azimuthal phase shift between the perturbation potential and density pairs of a spontaneously formed spiral or bar *mode*, such that the density leads the potential inside the corotation radius, and lags the potential outside the corotation. Due to the resulting secular torque between the density wave and the disk matter implied by the potential-density phase shift, the disk matter inside the corotation radius loses energy and angular momentum to the density wave and accretes inward, and the matter outside corotation gains energy and angular momentum from the wave and excretes. This happens for the entire disk surface, not just at the wave-particle resonances as predicted by older analyses. As a result, the disk surface density becomes more and more centrally concentrated, together with the build-up of an extended outer envelope, consistent with the direction of entropy evolution in self-gravitating systems. Furthermore, a local physical mechanism was also found to account for the collective 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, which makes spiral and bar modes akin to the collisionless shocks in plasma physics (Balogh & Treumann 2013). Subsequent work by Zhang (1998, 1999, 2008, 2016), Zhang & Buta (2007, 2012, 2015), and Buta & Zhang (2009) further established the relevance and importance of this collective dissipation process to the observed structure, kinematics, and secular morphological evolution of galaxies.

The conceptual support for these results also comes from the connection to theories of "dissipative structures" (Prigogine 1969, 1980; Glansdorff & Prigogine 1971; Nicolis & Prigogine 1977). Prigogine and coworkers have found that the large-scale coherent patterns present in complex, nonlinear, and far-from-equilibrium systems were not merely impressive veneers, they serve the important dynamical role of greatly accelerating the speed of entropy evolution of the underlying nonequilibrium systems. At the close-to-equilibrium regime, and under certain restrictive conditions, entropy production was found to be a minimum through transport processes (Prigogine 1969); whereas at the far-from-equilibrium regime, through the formation of dissipative structures, local entropy production is greatly accelerated, yet the global pattern is not destroyed in the process because the locally produced entropy is exported to the environment through the operation of the very same dissipative structures, such that the local entropy density can remain at a quasi-steady value. The accelerated entropy evolution refers to the system *plus* the environment, the paradox of entropy increase in the face of the formation of highly organized structures is thus solved. Nonequilibrium can become the source of order.

For self-gravitating systems, as we have commented before, the direction of entropy evolution is toward configurations of ever-increasing central concentration, together with an extended outer envelope (Antonov 1962; Lynden-Bell & Wood 1968). In the case of galaxies, this is the same as the direction of morphological evolution along the Hubble sequence from the late to the early Hubble types. Density wave patterns accelerate the local entropy production (the conversion of the organized orbital-motion energy into random motions of the stars and gas), but possess the ability to transport energy, angular momentum, and entropy to the environment through the coordinated interaction of the wave and the basic state.

1.3 Organization of the Material

The content of this monograph covers the analytical, numerical, and observational aspects of secular morphological evolution of galaxies driven by an internal dynamical process mediated by density wave modes. The choice of material reflects not only the author's past experience (as well as its unavoidable limitations), but also the fact that in the study of collective phenomena, a purely deductive analytical approach (or a "reductionist approach") is in principle not viable: the very essence of the self-organization process is such that the causes and effects become intermingled. This special feature in the collective global instabilities may be partly the reason that throughout the past centuries when classical mechanics made headway, only very slow progress had been made in the area of collective interactions within a complex system: because the necessary experimental, observational, and simulational techniques were not available. In cases where significant progress had been made before, such as Kolmogorov's derivation of the scaling laws for fully developed turbulence (Kolmogorov 1941a, 1941b, 1941c; Frisch 1995), a global, synthetic, balance-equation approach had been employed, rather than a deductive analytic approach[†]. Another example is the study of hydrodynamic shocks, where integral forms of the mass, energy and momentum balance-equations (i.e., the so-called Rankine-Hugoniot conditions) across the shock discontinuity need to be employed, together with the equations of state, to deal with the fact that a continuum (or differential) approach is no longer viable to aid with the task of crossing the shock front.

To properly employ a synthetic approach requires us to incorporate constraints from observational data of diverse range, as well as to use the results of numerical simulations to verify the products of logical reasoning. In a way, the approach adopted for writing this book mirrors the way progress had been made, and will continue to be made, in the study of complex and self-organized systems. Despite its employment of analytical equations and modern N-body simulations, in spirit this book has its kinship to Charles Darwin's *On the Origin of Species* (Darwin 1859) or Alfred Wegener's *The Origin of Continents and Oceans* (Wegener 1929), in that it is an explorer's account of how original discoveries were made, as well as a presentation of the diverse suite of evidence in support of a potential paradigm shift. The aim of the book is more to inspire a new generation of explorers to set foot onto this new frontier of galactic research, rather than to deliver the final words on a mature subject.

The organization of the chapters is as follows. In Chapter 2, we will present analytical derivation of the dynamical behaviors and emergent laws in disk galaxies containing spontaneously-formed density wave modes. Note that on a first reading, it is OK to skim through some of the more lengthy proofs in this chapter, as long as the central train-of-thoughts is followed. In Chapter 3, we will verify the theoretical

[†] In some sense the assumption of fully developed turbulence is like our assumption of the quasi-steady state of the wave mode, where the juggling act of energy injection and dissipation has been accomplished by nature, and the global self-consistency requirement is satisfied.

results through N-body simulations. Numerical techniques most relevant to the simulation of collective effects in disk galaxies will be highlighted. This will be followed by discussions of astrophysical implications and observational confirmations of the theoretical results in Chapter 4. In Chapter 5, we put together everything we have learned so far to infer further characteristics of the self-organized instabilities (including the relation of quasi-steady density wave modes to fully developed turbulence, as well as the origin of irreversible behavior in a nominally reversible dynamical system), as well as implications of this work on the cosmological evolution of galaxies. This will be followed by a brief Conclusions chapter (Chapter 6). The Appendix chapter (Chapter 7) discusses the link between the dynamical theory of nonequilibrium phase transition, to the relevant classical treatments in kinetic theory and fluid mechanics.

The current monograph focuses on the latest research results, and by itself does not serve as a complete course on galactic dynamics. The author encourages the reader to make full use of the classical texts (e. g., Rohlfs 1977; Gilmore et al. 1990; Shu 1992, especially Chapters 11 and 12, which together form a streamlined introduction to density wave theory; Binney & Tremaine 2008; Bertin 2014), in conjunction with the study of the current text, in order to arrive at a personal perspective. This was also the initial path that the author had taken in the late 1980s as a graduate student. Another piece of advice to the newcomers: take your own ideas and intuitions seriously, and follow them through with a lifetime of dedicated exploration.

2 Dynamical Drivers of Galaxy Evolution

We begin the formal presentation with a chapter devoted to the analytical formulation of the dynamical drivers for the secular morphological evolution of galaxies containing density wave modes. The formulation is built on the foundation of the density wave theory of galaxies as developed in the past few decades, which itself is built on the foundations of classical kinetic theory and fluid dynamics. As the presentation proceeds, however, we will point out that within this older theoretical framework, we cannot self-consistently accommodate a density wave mode of finite amplitude. The application of the continuum-based density wave theory leads either to neutral-wave solutions (as in the local WKBJ treatment), to linear and infinitely growing density wave modes (as in the fluid-disk treatment of global modes), or to non-self-consistent nonlinear solutions (as in the response of a gaseous disk to stellar density wave forcing). This sets the background for us to explore the new elements needed to construct a globally self-consistent theory that leads both to a self-sustained density wave modal-set in disk galaxies, as well as to the secular morphological evolution of its parent-disk mass distribution.

2.1 Motivation and Outline for the Theoretical Approach

In order to obtain sufficient radial mass flux to transform the Hubble types of galaxies over the age of the universe (i. e., within the so-called Hubble time, or 13.7–13.8 billion years), with the understanding that a significant fraction of galaxies may have formed much later, thus their ages are only a fraction of the Hubble time, common sources of microscopic viscosity were long since known to be drastically inadequate. From the analogy with other types of self-organized global patterns, such as the hexagonal convection cells in atmospheric flow (i. e., the Bénard problem), we naturally suspect that the strikingly beautiful density wave patterns in disk galaxies, such as spirals, bars, rings, and lenses, could perform a similar function of greatly accelerating the speed of entropy evolution and mass redistribution in their parent systems. After all, the whirlpool appearances of some of these galaxies (such as the namesake "Whirlpool Galaxy" M51) give us the impression that the orbiting stars and gas on the disks of these galaxies were spiraling into their central region. In fact, this very inkling was what had set the current author out on this strenuous quest in the late 1980s, when she was just beginning her astronomy graduate study at UC Berkeley[†].

[†] A popular account of the genesis of the spiral density wave theories, including the part that the author's work had played, can be found in an article by Jack Lucentini in September 2002 issue of the *Sky & Telescope* magazine.

DOI 10.1515/9783110527421-002

The situation with the galaxies turns out to be infinitely more complicated than the flow of water in the whirlpool, which at first glance may seem to underlie what is going on in a spiral galaxy. The main difference is that the process in the whirlpool is not a *self-organized* process, it is instead what we would call a *driven* or a *passive* process. The gravitational potential difference between the surface and the drain of a typical whirlpool coerces the accumulated water to spiral down the drain. The stars and gas in a galaxy, however, had achieved their quasi-equilibrium configuration, that is, that of the mass and velocity distributions on the galactic disk, through negotiating a delicate balance between gravity and pressure forces[†]. Unless they are perturbed by external driving forces (such as the tidal forces of a companion), there is not an easy "path of least resistance" for the galactic mass to follow, in order to overcome the angular momentum barrier due to their orbital motion.

Here is where the analogy with other types of self-organized dissipative structures comes into play. If the spirals and bars in galaxies are indeed another instance of dissipative structures, they will generate emergent physical characteristics to speed up entropy production and export, and to allow the original diffusion-type angular momentum transport to be replace by a convective (or advective) type of angular momentum transport. Since the density wave patterns are perturbations on the basic state mass distribution, and since most of the galaxy mass resides in the basic state, the key to this wave-assisted advective evolution process is to be found in a set of physical mechanisms which are able to (1) load the angular momentum from the basic state onto the wave within the inner disk, (2) allow the wave to transport the angular momentum to the outer disk, and (3) unload the transported angular momentum by the wave back onto the basic state at the outer disk region. This way the wave can be thought of as a kind of "lorry," performing the task of secular angular momentum transport while itself remains quasi-steady.

In order to load the angular momentum from the basic state onto the wave in the inner disk and unload it from the wave to the basic state at the outer disk, there needs to be an irreversible/dissipative interaction between the basic state and the wave. For classical WKBJ type (tightly wound) density waves, past studies have shown that there is no wave and basic state interaction for steady-amplitude waves (Goldreich & Tremaine 1979; Goldreich & Nicholson 1989). Therefore, to enable

[†] We can regard the stellar circular motion, or the so-called angular momentum barrier, as a kind of generalized pressure-force as well, when considering the system in a virial-equilibrium configuration. Incidentally, in Toomre (1964)'s derivation of the disk-galaxy local stability condition, it is the pressure *force* (or the local velocity dispersion) that enters into the stability criterion. However, in dealing with the global stability of a galactic system, when considering the stability of each galactic annulus separately, it is rather the pressure *gradient*, as well as centrifugal force of the stellar orbital motion, that balance the inward gravitational pull. Furthermore, if the global stability problem is considered from the point of view of the virial equilibrium of the galaxy as a whole, then once again the pressure force itself enters the consideration, and the orbital angular momentum can be viewed as just another form of stellar motion.

wave/basic-state interaction, one must look to the open types of spiral waves (i. e., those that have nonzero pitch angles).

What is the new physical element added by the open waves? As we all know, gravitational interaction is long range, meaning that the potential of a mass distribution is obtained from the Poisson integral of the mass distribution, and the resulting potential does not have to look like the mass distribution itself. In the case of a mass distribution in the shape of a spiral, the Poisson integral in general produces a potential spiral that is *phase shifted* in azimuth from the density spiral that generates the potential field. This phase shift is of vital importance to the dynamical mechanism that enables the secular mass redistribution in galaxies since, as we will show below, in the disk geometry in order for a *quasi-steady* wave pattern to exchange angular momentum with the basic state mass distribution secularly, a phase shift between the perturbation potential and density distributions is the *necessary and sufficient condition*.

But would the radial distribution of the azimuthal phase shift be of the correct sense to allow angular momentum loading in the inner disk, and unloading in the outer disk? As it turns out, this is exactly the case for spontaneously formed density wave *modes*. The types of potential and density distributions for modes are such that the phase shift is positive (means density leads potential in the azimuthal direction of galactic rotation) inside the corotation radius, and negative outside. This pattern of phase shift distribution is equivalent to the torquing by the density wave on the basic state matter in each annular ring of the galactic disk in just the correct sense to produce a secular mass redistribution trend as dictated by the entropy law of self-gravitating systems. Therefore, nature seems to have engineered a mechanism in the precise fashion to allow a disk galaxy to achieve the goal of accelerated entropy evolution. The same phase shift distribution will be shown to be responsible also for the spontaneous emergence of the wave mode in the linear regime, and for its stabilization to a constant amplitude at the quasi-steady state.

With the mechanism for loading and unloading of the angular momentum now available, and with the mechanism for outward transport of the angular momentum previously found by Lynden-Bell & Kalnajs (1972)[†], we are in possession of a dynamical process for the secular evolution of the disk-galaxy mass distribution. However, much more subtle intricacies of the process still await discovery. The phase-shift/torque/angular-momentum-transport process is a global description. There is the need to explore the local dynamical processes that allow these global relations to be fulfilled. This is shown to be related to the local instability condition at the density

[†] Lynden-Bell and Kalnajs had advocated for a *constant* angular momentum flux for a steady wave train, due to their *a priori* assumption of no-wave-basic-state interaction. Whereas with the continuous angular momentum loading within the inner disk, and unloading in the outer disk, which we will demonstrated in this work, the angular momentum flux turns out to be of a characteristic bell shape, with the peak of the bell at the corotation radius.

wave crest of an open spiral (or skewed bar) mode. This instability condition, coupled with the demonstration of the transonic velocity jump across the spiral arm (see Section 3.3), shows that the density wave modes possess the essential features of plasma "collisionless shocks" (Zhang 1996; Balogh & Treumann 2013).

The collisionless shock at the density wave crest is a formal singularity of the solutions of the underlying fluid (or stellar dynamical) equations used to model the finite-amplitude density wave modes in galactic disks. We will show that within such formal singularities of the dissipationless equation sets, which are used to model the spontaneously formed density wave modes, the differential form of the Poisson equation is no longer valid (in fact, the continuum formulation itself is invalid), which is to be expected since a formal singularity of the solutions of the governing differential equation set implies precisely the breakdown of the differential forms of governing equations, or the breakdown of mean-field/continuum treatment. So we will need to seek a genuine particle treatment of the problem to repair this "breakdown," and to put the different pieces of the self-organization process back together.

As part of this process of putting together, as we will show in Chapter 3, when advancing articles on their trajectories in N-body simulations, we in fact are making use of velocities generated from the force field of the matter distribution that incorporated the multitude of *correlations* among disk particles. The kinematics and the particle mass distribution together enable the self-organization process, and together they encode particle correlations. This is why a proper treatment of the gravitational viscosity generated by self-organized density wave patterns requires a genuine N-body simulation, which naturally models particle-correlation-induced gravitational viscosity through the use of the integral form of the Poisson equation, whereas schemes such as the smoothed-particle-hydrodynamics (SPH, Monaghan (1992) and the references therein), which is based on the mean-field differential formulation, will necessarily require the specification of gravitational viscosity as an artificial input parameter.

We emphasize at the outset that a *proper basic state choice* that allows genuine unstable (or self-organized) modes to spontaneously emerge is crucial for the attainment of the correct secular evolution behavior of the basic state of the disk. Only such basic states that admit intrinsic unstable modes display *correlated fluctuations* that produce the kind of wave/basic-state interactions that lead to galaxy evolution along the Hubble sequence. Over-stable disks that allow only *transient wave trains* to emerge will not serve this purpose. We will address this point further in Section 5.1.2.

To follow the intricate details of the self-organization process, as we have mentioned before, a particle approach is required, which we will defer the discussion to Chapter 3. In this chapter, we present analytical derivations of the relevant secular evolution dynamics mostly from a global closure-relation point of view. We will make use of the conservation requirements of the various physical entities to derive new closure relations at the quasi-steady state of the wave mode, which turn out to be closely related to the potential-density phase shift we had mentioned earlier. The process through which the older closure relations in the angular momentum transport and exchange processes is replaced by emergent new closure relations is similar in spirit to the mechanism of spontaneous breaking of gauge symmetry in high-energy physics to arrive at new dynamical laws, when the energy scale of fundamental physical processes is traversed. In fact, nature appears to organize the hierarchies of physical laws in the different regimes of physical parameters through invoking just such symmetry-breaking processes, both in the high-energy and low-energy physics. We will come back to the discussion of this topic toward the end of this monograph.

2.2 Density Wave Crest as the Site of Gravitational Instability†

Much of the new dynamics responsible for enabling the secular morphological evolution of galaxies originates from the operation of collective effects due to the multitudes of mutually interacting stars and molecular complexes in the galactic disk, especially near the potential minimums of the density waves (Zhang 1996).

For collective effects to operate in a spiral or a barred galaxy, individual stars have to be *aware* of their “neighbors” directly, besides experiencing the smoothed axisymmetric plus 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 2008). Under this circumstance, the scattering of a star off its neighboring stars can only happen when the disk is locally gravitationally unstable – even if just marginally so. Therefore, 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 in an originally marginal stable disk.

We point out at the outset that in the derivations in this section, even though expressions for WKBJ waves were initially used, the final conclusions of collective dissipation in self-organized modes depend on the patterns being *open*, or have finite pitch angles. *The WKBJ formulation is used here only to illustrate its own inadequacy* – that is, we will arrive at the conclusion that only by going *beyond* the WKBJ approximation in the density wave theory, can we expect to arrive at the true source of local gravitational instability, as well as collective dissipation and secular morphological evolution of galaxies. Furthermore, even though we start with a generic density *wave* formulation, in later sections of this chapter we will show that the global-self-consistency requirement dictates that the kind of open waves that allow the secular and quasi-steady mass redistribution need to be *unstable wave modes* of the underlying basic state of the galactic disk.

† Portions of this section used material previously published in Zhang (1996), reproduced with modifications from The Astrophysical Journal @ AAS. Reproduced with permission.

2.2.1 Local Stability Condition at the Spiral Arm and Interarm Region

By considering the competing influence of pressure-force and rotational stabilization effects in a stellar disk configuration, Toomre (1964) arrived at the following well-known local stability condition against axisymmetric type of instabilities in a disk geometry:

$$Q \equiv \frac{\sigma_r \kappa}{3.36 G\Sigma} > 1, \tag{2.1}$$

where σ_r is the radial velocity dispersion, κ is the epicycle frequency, Σ is the surface density of the disk, G is gravitational constant and Q is Toomre's stability parameter. For a fluid disk, the factor 3.36 in the denominator is changed to π.

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 derive the variations of these parameters with the phase of the spiral and calculate how these variations influence the value of the instability parameter Q at the spiral arm and the 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.

In the following we adopt the simpler Eulerian fluid formulation (rather than the stellar dynamical equation used in the original Lin–Shu theory) for the discussion of the stability condition in the spiral-arm region. For an m-armed spiral density wave of pattern speed Ω_p, the gravitational potential at the disk location (r, ϕ) and time t can be written as (Rohlfs 1977; Shu 1992)[†]

$$\mathscr{V}(r, \phi, t) = \mathscr{V}_0(r) + A(r) \exp\{i[m\Omega_p t - m\phi + \Phi(r)]\}, \tag{2.2}$$

where $|A| \ll |\mathscr{V}_0|$ and where $\Phi(r)$ is related to the pitch angle i and the wavenumber $k = \lambda/2\pi$ of the spiral through

$$\frac{d\Phi}{dr} = \frac{m}{r \tan i} = k, \tag{2.3}$$

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

$$|kr| \gg 1, \tag{2.4}$$

[†] Despite using somewhat different notations, the formulations of Rohlfs (1977) and Shu (1992), when both using Eulerian fluid equation set, are completely equivalent. For example, the solution for the perturbation spiral density under the linear and WKBJ approximation, is represented by eq. (68) in Rohlfs (1977), and by eq. (11.44) in Shu (1992). These two solutions can be shown to be identical after straightforward variable substitutions.

or that the wavelength of the wave is much smaller than the system dimension under concern, so locally the WKBJ waves in the disk geometry approximate plane waves.

The solution for the azimuthal velocity of the streaming stars within the WKBJ approximation is

$$\mathrm{v}(r,\phi,t) = \mathrm{v}_\mathrm{c}(r) + i\frac{kA}{2\Omega_0}\frac{1}{1-\nu^2+x}\exp\{i[m\Omega_\mathrm{p}t - m\phi + \Phi(r)]\}, \tag{2.5}$$

where ν is the normalized encounter frequency of streaming stars with respect to an m-armed spiral pattern, and

$$\nu = m(\Omega_\mathrm{p} - \Omega_0)/\kappa_0, \tag{2.6}$$

$$x = k^2\sigma_{r0}^2/\kappa_0^2, \tag{2.7}$$

and where v_c, $\sigma_{\mathrm{r}0}$, 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, that is, $\mathrm{v}_\mathrm{c}(r) = \mathrm{v}_\mathrm{c}$ is a constant. Since we are considering WKBJ waves, which is a local plane-wave approximation of the density waves, the constant circular velocity assumption will not increase the level of error within the WKBJ approximation.

The corresponding density variation is

$$\Sigma(r,\phi,t) = \Sigma_0(r) - \Sigma_0(r)\frac{k^2A}{\kappa_0^2}\frac{1}{1-\nu^2+x}\exp\{i[m\Omega_\mathrm{p}t - m\phi + \Phi(r)]\}. \tag{2.8}$$

Since

$$\kappa^2 = 2r\Omega\frac{d\Omega}{dr} + 4\Omega^2, \tag{2.9}$$

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

$$\Omega(r,\phi,t) = \frac{\mathrm{v}(r,\phi,t)}{r} = \Omega_0(r) + i\frac{kA}{2\Omega_0 r}\frac{1}{1-\nu^2+x}\exp\{i[m\Omega_\mathrm{p}t - m\phi + \Phi(r)]\}. \tag{2.10}$$

Therefore,

$$\frac{d\Omega}{dr}(r,\phi,t) = -\frac{\mathrm{v}_\mathrm{c}^2}{r^2} - \frac{k^2A}{2\Omega_0 r}\frac{1}{1-\nu^2+x}\exp\{i[m\Omega_\mathrm{p}t - m\phi + \Phi(r)]\}. \tag{2.11}$$

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

$$\kappa^2(r,\phi,t) = \kappa_0^2\left\{1 - \frac{k^2A}{\kappa_0^2}\frac{1}{1-\nu^2+x}\exp\{i[m\Omega_\mathrm{p}t - m\phi + \Phi(r)]\}\right\}, \tag{2.12}$$