Adventures in Theoretical Astrophysics

Thesis by

Alison Jane Farmer

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2005

(Defended May 13, 2005)

Acknowledgements

Thanks...

To everyone who made my PhD more than just an academic experience.

To my Caltech officemates, who have all managed to be successful despite my constant inane chatter: To Dawn Erb and Naveen “Dalek” Reddy for the best-decorated office in Robinson, to Margaret Pan for eating as many meringues as I did, and to Mike Santos for being as good at listening as he is at talking.[1]

To my big brothers at the IAS, for all the teasing.

To all the men I have wronged.

To everyone who beat me at sports.

To everyone who is in more than one of the above categories.

To California for my accent.

To New Jersey for the cicadas.

To db for feeding me like royalty while I wrote up this thesis.[2]

To Mr. Higgins of James Gillespie’s High School in Edinburgh for giving up his lunchtimes to talk physics.

To Sterl Phinney for converting me to theory. To Asantha Cooray and Eric Agol for their collaboration. To Re’em Sari for making me behave better, and for lunch.

To Peter Goldreich — my advisor, father-figure, grandfather-figure, personal trainer, chauffeur and friend — for everything. To Susan Goldreich for looking after him.

To my parents, Ann and John Farmer, for believing in me.

[1] Marill!

[2] Rat vit rôt, rôt tenta rat, rat mit patte à rôt, rôt brûla patte à rat.

Abstract

This thesis is a tour of topics in theoretical astrophysics, unified by their diversity and their pursuit of physical understanding of astrophysical phenomena.

In the first chapter, we raise the possibility of the detection of white dwarfs in transit surveys for extrasolar Earths, and discuss the peculiarities of detecting these more massive objects.

A population synthesis calculation of the gravitational wave background from extragalactic binary stars is then presented. In this study, we establish a firm understanding of the uncertainties in such a calculation and provide a valuable reference for planning the *Laser Interferometer Space Antenna* mission.

The long-established problem of cosmic ray confinement to the Galaxy is addressed in another chapter. We introduce a new wave damping mechanism, due to the presence of background turbulence, that prevents the confinement of cosmic rays by the resonant streaming instability.

We also investigate the spokes in Saturn's B ring, an electrodynamic mystery that is being illuminated by new data sent back from the *Cassini* spacecraft. In particular, we present assessments of the presence of charged dust near the rings, and the size of currents and electric fields in the ring system. We make inferences from the *Cassini* discovery of oxygen ions above the rings. In addition, the previous leading theory for spoke formation is demonstrated to be unphysical.

In the final chapter, we explain the wayward motions of Prometheus and Pandora, two small moons of Saturn. Previously found to be chaotic as a result of mutual interactions, we account for their behavior by analogy with a parametric pendulum. We caution that this behavior may soon enter a new regime.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Executive Summary

Astrophysics remains a field in which a theorist can possess expertise across the board. Many of the same principles are at work on scales from planets to pulsars, from cosmic rays to galaxy clusters. With the frontiers of astrophysics forever shifting, a mobile theorist is in an enviable position, always able to work on the most interesting problems.

This thesis charts my progress towards becoming a general purpose theoretical astrophysicist. The following chapters span white dwarfs, gravitational waves, cosmic rays, MHD turbulence, electrodynamics, orbital dynamics, chaos, and plasma physics. Learning must continue throughout the career of a general theorist, and there is much I have yet to learn. Over the course of the thesis my interests have widened but my approach has become more refined: reaching physical understanding through order of magnitude calculations has emerged as a strong theme in my work. Moving on from earlier work (chapters 2 and 3, and further papers that are not included in this thesis[1]) in which detailed calculations predicted future observations, my more recent work (chapters 4, 5, 6, and 7) is involved with investigating puzzles and obtaining familiarity with astrophysical environments.

In the following sections I introduce each chapter in turn, placing the work in context and outlining the main results. The chapters appear roughly in order of completion of their contents. Chapters 2, 3, and 4 have already appeared in print, and are reproduced here with the permission

[1] Cooray & Farmer, 2003, Occultation searches for Kuiper Belt objects, ApJL 587, 125; Cooray, Farmer & Seto, 2004, The optical identification of close white dwarf binaries in the Laser Interferometer Space Antenna era, ApJL 601, 47.

of the copyright holders. A bibliography is included at the end of each chapter.[2]

Chapter 2: Finding white dwarfs with transit surveys[3]

It has become fashionable to search for extrasolar planets via transit surveys. Most of these are designed to detect Jupiter-sized planets, which produce a ~ 1 % dimming as they pass in front of their parent stars. The proposed *Kepler* survey is more ambitious: this space-based mission will detect the ~ 0.01 % flux decrease due to transits of Earth-sized exoplanets.

While not denying the excitement of finding Earth analogs, we set out to investigate some other physics that could be done with the *Kepler* mission. We make use of the interesting coincidence that because white dwarfs are of similar size to the earth ($\sim$ 10,000 km), their transit signals are also detectable by *Kepler*. Further, we discover that microlensing can be significant in close white dwarf binaries, and that some transits will show flux *increases*. We demonstrate that, far from being a nuisance to the transit surveys, the study of white dwarf transit lightcurves could enable an accurate test of the white dwarf mass-radius relation predicted by Chandrasekhar. Most other tests rely heavily on atmospheric modeling.

We explore this possibility using binary star population synthesis to estimate that around 100 white dwarfs will show up in the *Kepler* survey as they transit their main sequence companions. We discover that, due to the effects of common-envelope evolution, there is little overlap in orbital period between habitable planets and white dwarfs. A complication in the white dwarf case is that the main sequence star can be tidally spun up by a massive white dwarf, leading to increased stellar variability and photometric noise. We quantify the problems this can cause for detecting transits in the shortest period systems. The *Kepler* mission is still in the planning stages, and as a result of our paper, efforts are being made (D. Sasselov, private comm.) to retain stars with large radial velocity variations in the transit sample, since these systems may harbor white dwarfs.

Topics: WDs, binary star evolution, planets, stellar variability, population synthesis, microlensing.

[2] In the style of the paper in which each chapter has or will be published

[3] Originally published as Farmer & Agol, 2003 ApJ 592, 1151 and reproduced with permission

Chapter 3: The gravitational wave background from cosmological compact binaries[4]

The *Laser Intereferometer Space Antenna* (*LISA*) will detect many types of gravitational wave signals, from such sources as the coalescence of supermassive black holes, the inspiral of compact objects into supermassive black holes, and perhaps even a background from inflation. The extent to which *LISA* can detect these is set both by instrumental limitations and by astrophysical backgrounds of (perceived) less interesting signals. These astrophysical backgrounds must be understood during the design phase, since there is little point in increasing instrumental sensitivity beyond the limits set by astrophysics.

The dominant background in most of the *LISA* frequency band (0.1–10 mHz) is from close double white dwarf (WD–WD) binaries in our Galaxy. This background is well-simulated and understood. Above ~ 3 mHz, there are few enough Galactic binaries that their signal can be spectrally resolved and hence subtracted. The important background in this part of the *LISA* band is then that due to the innumerable WD–WD binaries in other galaxies. The efforts to simulate these sources had met with limited success prior to our paper. In particular, the most detailed study had presented a most unexpected spectral shape. With this motivation, we calculate both analytically and numerically the expected spectrum from cosmological binaries. The numerical work consists of a population synthesis of all the close binary stars in the universe. A careful parameter variation is performed to determine the sensitivity of our calculations to poorly understood physics such as common-envelope evolution. Each of our models is tested against the observed population of WD–WD binaries in our Galaxy, and the origins of spectral variations are carefully understood.

We establish a new spectral shape for the background, acquired both numerically and analytically. The background amplitude is predicted to within an order of magnitude when all of the uncertainties are taken into account. Its level, $\Omega_{\rm gw} \sim 10^{-12}$ in a logarithmic frequency band around 1 mHz, is conveniently just below the planned instrumental sensitivity.

[4] Originally published as Farmer & Phinney, 2003 MNRAS 346, 1197 and reproduced with permission

This paper, the deepest study of its kind, has become the definitive work in the field.

Topics: gravitational waves, binary star evolution, cosmology, population synthesis.

Chapter 4: Wave damping by MHD turbulence and its effect on cosmic ray propagation in the ISM[5]

Relativistic cosmic rays are mostly of Galactic (as opposed to solar) origin. Of the many mysteries in cosmic ray (CR) astrophysics, one of the longest-standing is their confinement to the Galaxy. A relativistic particle can cross the disk of the Galaxy in 3,000 years, but radioactive dating places CR ages at $\sim$ 10 million years. Further, although they are expected to be produced inhomogeneously in the Galaxy, CR arrival directions are nearly isotropic, at least up to energies $\sim 10^{15}$ eV.

One way of impeding CR escape from the Galaxy is scattering off inhomogeneities (Alfvén waves) in the magnetic field lines on which they travel. Some inhomogeneities are always present, in the form of background turbulence (produced by supernova shocks, etc.). These waves are however elongated in a direction that makes them unsuitable for scattering CRs. A long-established alternative to scattering off pre-existing turbulence is for the CRs to create their own turbulent wakes, and to scatter off those. These waves, with which the CRs are resonant, are naturally suited to scattering CRs. If the waves are excited faster than any process can damp them, then the excitation and scattering increase exponentially. This *resonant streaming instability* (RSI) proceeds until the CRs are isotropized in the frame of the waves (i.e., that moving with the local Alfvén speed).[6] The CR fluid can then only stream out of the Galaxy at the Alfvén speed ($v_{\rm A} \sim 2 \times 10^7$ cm s^{-1}), explaining the long residence times.

The unresolved issue in this picture is whether any source of wave damping prevents the growth necessary to slow the CR streaming velocity. Damping via collisions with neutral atoms prevents the RSI from operating in the warm phase of the interstellar medium. In the hot cororal gas, collisions are infrequent and wave damping by nonlinear collisionless processes has been considered. In our

[5] Originally published as Farmer & Goldreich, 2004 ApJ 604, 671 and reproduced with permission

[6] The CRs do not lose a significant part of their energy; only the streaming velocity of the CR fluid is reduced, i.e., most of the fluid momentum is lost to the waves.

paper we introduce a *new* damping mechanism for CR-generated waves, which is operational in all phases of the ISM.

We demonstrate that background turbulence is not only ineffective at scattering CRs itself, but it also prevents the CRs from exciting waves via the RSI. The turbulent cascade shears and swallows the CR waves faster than they can be excited, precluding the operation of the RSI down to almost the lowest CR energies. We calculate the damping rate for the Goldreich-Sridhar model of turbulence, using a geometrical approach. Our paper does not resolve the issue of CR scattering, but finds that the RSI cannot be responsible for confinement of Galactic CRs. Our turbulent damping mechanism may also be of relevance to the acceleration of CRs at shocks, where operation of the RSI may be required.

Topics: turbulence, high energy particles, ISM, MHD.

Chapter 5: Ghosts of Saturn: studies for the ring spokes

Discovered by the *Voyager* spacecraft 25 years ago, the spokes in Saturn's rings are transient radial dust lanes. They are seen only near the place in the rings where the Keplerian angular velocity matches the rotation rate of Saturn. Spokes may form in as little as 5 minutes along their entire 10,000 km length, then they fade in a few hours. The mystery of the spokes gave birth to the whole field of dusty plasma physics, which has thus far failed to find an explanation for their formation.

There is however consensus on the principles at work behind the spokes: formation is believed to involve the sudden lifting of a radial lane of charged dust by electromagnetic forces. The dust is then "let go" by the fields, and falls back onto the ring plane.

Spurred on by the arrival of *Cassini* at Saturn in 2004 and the promise of new data, we embarked on a fresh study of the spokes. An unexpected development was the complete absence of spokes, whose lack of detection since 1996 was previously thought to be merely an effect of our viewing angle from Earth. The new data tell us that spokes only appear when the rings are close to edge on to the Sun, in Saturnian spring and autumn. Any successful theory must now also explain their current absence.

We endeavor to understand the environment in which the spokes form: how much dust is present in the rings, how it might become charged, and the magnitudes of the expected currents and magnetic field perturbations in the Saturnian system. Resulting from these studies, we present one potential spoke formation mechanism that selects the correct location in the rings.

An intriguing discovery made by the *Cassini* spacecraft was the existence of an oxygen ionosphere over the rings, hinting at the presence of a much greater amount of neutral O_2 in a ring atmosphere. We estimate the total amount of oxygen (neutral and ionized) that can be present, and assess its effect on the electrodynamics of the environment. We believe that the oxygen atmosphere/ionosphere may be the key that was missing in the understanding of the spokes. Regardless of its connnection to the spoke problem, however, this is a brand new and fascinating area of study.

Topics: electrodynamics, atmospheric physics and chemistry, impacts, plasma physics, dusty plasmas.

Chapter 6: Spoke formation under moving plasma clouds[7]

The most sophisticated and widely accepted spoke hypothesis is that of Goertz & Morfill (1983). In this theory, a meteoroid impact on the rings produces a dense plasma cloud that charges the rings negative, thus causing negatively charged dust grains to be repelled from the ring surface. The motion of these grains relative to the plasma produces an electric field, which causes the plasma cloud to drift in the radial direction. Goertz & Morfill claim that the plasma drifts, lifting more dust on its way, at ~ 30 km s^{-1}, fast enough to make a radial feature in the differentially rotating rings. In our paper (submitted to Icarus) we show that such drift velocities are impossible, so the Goertz & Morfill theory does not work. When we solve the self-consistent electrodynamics problem in which all currents close, we find that the drift velocity is at most ~ 2 km s^{-1}, far too slow to make a radial spoke.

Topics: electrodynamics, atmospheric physics, plasma physics.

[7] Submitted to Icarus as Farmer & Goldreich, 2005

Chapter 7: Understanding the behavior of Prometheus and Pandora

Prometheus and Pandora are two small moons orbiting at nearly the same distance from Saturn. Discovered by *Voyager*, they flank the narrow F ring, and are often known as the "F ring shepherds." Their orbits were fitted to *Voyager* data using ellipses that precess due to Saturn's oblateness.

In 1995, when Prometheus and Pandora were reobserved during the Saturn ring-plane crossing, they were not where they were expected to be. Each satellite was ~ 20 degrees from the predicted position in its orbit. Freely precessing ellipses were clearly not sufficient to describe their motions. After much speculation by others, Goldreich & Rappaport (2003) established that the discrepancies are due to the satellites' mutual interactions, involving a mean motion resonance. They also discovered that the satellites' motions are chaotic, and stated that intervals of unpredictability occur when the orbits are antialigned (once every 6.2 years).

Further simulations were carried out in preparation for the arrival of *Cassini.* The omission in all of these studies was a real understanding of the chaotic dynamics, not only an aesthetic issue but one of importance to making predictions. In our paper we show how to understand the system by analogy with that of a pendulum bob moving in a varying gravitational field (a parametric pendulum). Using the generic property that chaotic systems explore the whole chaotic region of phase space (ergodicity), we demonstrate that unpredictability can occur at any time throughout the precessional cycle, not just when the orbits are antialigned. The behavior of Prometheus and Pandora may change drastically in as little as 20 years. With the study in this paper, they will not take us by surprise again.

Topics: orbital dynamics, chaotic dynamics.

Chapter 2

Finding White Dwarfs with Transit Searches

Originally appeared as

Alison J. Farmer & Eric Agol,

“Finding white dwarfs with transit searches”

Astrophysical Journal, vol. 592, pp 1151–1155, 2003

Reproduced with permission.

Abstract

We make predictions for the rate of discovery of eclipsing white dwarf–main sequence (WD–MS) binaries in terrestrial-planet transit searches, taking the planned *Kepler* and *Eddington* missions as examples. We use a population synthesis model to characterize the Galactic WD–MS population, and we find that, despite increased noise due to stellar variability compared with the typical planetary case, discovery of $\gtrsim 10^2$ non-accreting, eclipsing WD–MS systems is likely using *Kepler* and *Eddington*, with periods of 2–20 days and transit amplitudes of $|\Delta m| \sim 10^{-(4\pm0.5)}$ magnitudes. Follow-up observations of these systems could accurately test the theoretical white dwarf mass–radius relation or theories of binary star evolution.

2.1 Introduction

As white dwarf stars are comparable in size to the earth, searches for terrestrial planets via transit of their parent stars should be capable of discovering white dwarf (WD) binaries as well. These can be distinguished from terrestrial planet systems because the WD will induce detectable Doppler shifts in the spectral lines of a main sequence (MS) companion, whereas an Earth-like planet will not, due to its smaller mass.

An interesting coincidence is that the radius of a white dwarf in a tight binary is comparable to its Einstein radius (Marsh, 2001), which means that both gravitational lensing and eclipse may be important during transit. When combined with other information about the binary, observation of the change of magnitude during transit may allow measurement of the mass and radius of the white dwarf, allowing a test of the relation originally predicted by Chandrasekhar (1935). Recent measurements of white dwarf masses and radii using atmosphere models (Provencal et al., 1998) have modeling uncertainties, which would be useful to test with another technique, and the range of WD masses available in binaries is larger than for single stars. The four known eclipsing WD–M-dwarf binaries are a good start (Marsh, 2000), but a larger, alternatively selected sample could allow measurement of the variation of the relation with WD mass, age, or composition.

This paper addresses the question of how many white dwarf–main sequence (WD–MS) binaries might be found in searches for transiting planets, with application to the *Kepler*[1] (Borucki et al., 1997) and *Eddington*[2] (Horne, 2002) missions. We concern ourselves with white dwarfs in detached binaries, since accreting white dwarfs can be found by other means. In § 2 we discuss the various modes for discovery of white dwarfs and the surveys that might detect them, then in § 3 we describe the expected properties of WD–MS binaries in our Galaxy. In § 4 we estimate the number of binaries that might be found in the surveys mentioned above, in § 5 we discuss what one might do with any binaries discovered via this method, and in § 6 we conclude.

During the preparation of this paper, Sahu & Gilliland (2003) published a study of near-field WD microlensing for the *Kepler* mission. In their paper, numerical lightcurves for fiducial systems

[1] See http://www.kepler.arc.nasa.gov/

[2] See http://astro.esa.int/SA-general/Projects/Eddington/

are presented, similar to the analytic lightcurves presented in Agol (2002); however, they did not attempt to accurately estimate the number of detectable systems, the goal of this paper.

2.2 White dwarf binary variations

First we consider the case in which the WD passes in front of the MS star (primary transit). The Einstein radius is $R_{\rm E} = [4R_G a]^{1/2}$, where $R_G = GM_{\rm WD}/c^2$ is the gravitational radius for a lens of mass $M_{\rm WD}$ and a is the semimajor axis of the binary. A white dwarf in a binary system has a size, $R_{\rm WD}$, which is comparable to the Einstein radius

$$\frac{R_{\rm WD}}{R_{\rm E}} \simeq 0.7 \left(\frac{M_{\rm WD}}{M_\odot}\right)^{-1/2} \left(\frac{a}{0.1{\rm AU}}\right)^{-1/2} \left(\frac{R_{\rm WD}}{0.01R_\odot}\right). \tag{2.1}$$

Thus microlensed images (which will appear a distance $\sim R_{\rm E}$ from the center of the WD) may be occulted by the WD, so that we may have either dimming (favored if $R_{\rm WD} \gg R_{\rm E}$, i.e., small $M_{\rm WD}$, a) or amplification (if $R_{\rm WD} \ll R_{\rm E}$, i.e., large $M_{\rm WD}$, a).

If the occulting body (here the WD) is much smaller than the occulted body (MS star), the microlensing plus occultation equations take on an exceptionally simple form (Agol, 2003), with the dimming or amplification dependent only on the surface brightness immediately behind the WD, so long as the WD is away from the edge of the MS star. The fractional change in flux during a transit is then given by

$$\Delta f_1 \equiv \left(\frac{2R_{\rm E}^2 - R_{\rm WD}^2}{R_{\rm MS}^2}\right) \left(\frac{F_{\rm MS}}{F_{\rm MS} + F_{\rm WD}}\right) \frac{I(r)}{\langle I \rangle} \Theta(R_{\rm MS} - r), \tag{2.2}$$

where $F_{\rm MS,WD}$ are the MS and WD fluxes, respectively, $R_{\rm MS}$ is the radius of the MS star, r is the projected distance of the WD from the center of the MS star, $I(r)/\langle I \rangle$ is the limb-darkened intensity profile of the source normalized to the flux-weighted mean intensity, and Θ is the step function.

For close ($a \sim 0.1$ AU) WD–MS systems, $|\Delta f_1| \sim 10^{-4}$, since $R_{\rm WD} \sim R_{\rm E} \sim 10^{-2} R_{\rm MS}$, similar to the transit depth of a terrestrial planet, though the effect will be a flux increase when $R_{\rm WD}/R_{\rm E} < 2^{-1/2}$. In a system of random inclination i on the sky, the probability of transits along our line of sight is $\simeq R_{\rm MS}/a$, and for a small transiting body, the transit duration will be $T_{\rm tr} \simeq 2R_{\rm MS}\sin\theta/v_{\rm orb}$, where $R_{\rm MS}\cos\theta = a\cos i$ and $v_{\rm orb}$ is the relative orbital velocity.

The fractional change of flux when the MS star passes in front of the WD (secondary transit) is simply given by $\Delta f_2 \equiv -F_{\rm WD}/(F_{\rm WD}+F_{\rm MS}) < 0$. The transit will obviously be deeper for younger, more luminous WDs. The luminosity of a typical WD 1–10 Gyr after birth is of the order $10^{-3}-10^{-5}$ $\rm L_\odot$, so that terrestrial planet searches should be well-suited to detecting these events too. Note that these will be flat-bottomed transits, but will have the same durations and transit probabilities as the complementary primary transit. If the orbit is circular, as will be the case for close systems, a given lightcurve will exhibit either both primary and secondary transits (though the transit depths Δf_1 and Δf_2 may be very different) or no transits at all.

There are other types of variations in WD–MS lightcurves, including fluctuations as a result of tidal effects, both directly, due to tidal distortion of the MS star, and indirectly, due to increased MS rotation rate and hence increased stellar variability, which is described in detail in § 2.4.1. In addition, we may see variations due to irradiation of the MS star by a hot WD in the very youngest WD systems, neglected here, or due to flickering if the WD is accreting at a low level from the MS stellar wind, also discussed in § 2.4.1.

Since the predicted transit amplitudes are of the same order of magnitude as those for terrestrial planet transits ($\Delta f \simeq |\Delta m| \sim 10^{-4}$ magnitudes), surveys designed to search for other Earths ought to pick up WD–MS systems too. To illustrate this, we use as examples the *Kepler* (Borucki et al., 1997) and *Eddington* (Horne, 2002) missions. These satellites will continuously monitor the brightness of many stars at high photometric precision to search for periodic dimming characteristic of a transiting planet.

Kepler will monitor $\sim 10^5$ dwarf stars with $9 < V < 14$ in a $10^\circ \times 10^\circ$ field centered on Galactic coordinates $(l,b) = (69.6^\circ, 5.7^\circ)$, for at least four years. The proposed *Eddington* design is for a smaller $3^\circ \times 3^\circ$ field, with deeper magnitude limits ($11 < V < 18$) and a lifetime of 3 years. In the absence of published coordinates for the *Eddington* field, we use here the same (l,b) as *Kepler*. *Kepler* will have a broad band pass, extending from ~ 400 to 850 nm, while *Eddington* may have two-color information (Bordé et al., 2003). *Kepler* will read out fluxes every 15 minutes, and will have a fiducial sensitivity (similar to that of *Eddington*) of 2×10^{-5} for a 6.5-hour exposure of a $V = 12$ G2V star.

2.3 White dwarfs in binaries

2.3.1 Evolution to the WD–MS stage

The evolution of a zero-age main sequence (ZAMS) binary system to a white dwarf–main sequence (WD–MS) system proceeds via one of two main pathways, according to whether or not Roche lobe overflow occurs en route, the critical initial orbital separation being at $a_{\rm crit} \sim 10^3 - 10^4 {\rm R}_\odot$; in both cases, the more massive star has evolved into a WD, while the other is still on the MS. The lifetime of this phase depends on the difference in main-sequence lifetimes of the two stars.

If the primary fills its Roche lobe on the RGB or AGB, then the ensuing mass transfer will most likely be dynamically unstable and a common envelope phase will result (see e.g., Iben & Livio, 1993), in which the envelope of the evolved star is heated by friction as the secondary and the core of the giant orbit inside it. This ends when either the stars coalesce or the envelope is heated sufficiently that it escapes the system, leaving a WD–MS binary with greatly *reduced* orbital separation compared with the initial ZAMS system. The orbit will be circular, since tidal circularization will have occurred in any system in which a giant star is close to overflow.

If the orbital separation is sufficiently large that the primary does not overflow on the RGB or AGB, then it is affected only as the primary loses its envelope in the planetary nebula phase, leading to an *increase* in the orbital separation. We therefore expect that the Galactic WD–MS population will consist of two *distinct* groups of sources in P-space, the "short" period systems from systems in which overflow has occurred, and the "long" period systems in which no Roche lobe has been filled.

2.3.2 Population synthesis

A population synthesis approach was used to quantify the properties of the Galactic WD–MS population. Evolutionary tracks were computed using the rapid evolution BSE code (Hurley, Pols & Tout, 2000). Following the preferred Model A of Hurley et al. (2000), we distribute the primary mass according to the initial mass function (IMF) of Kroupa, Tout & Gilmore (1993), while the secondary mass distribution we choose to be flat in the mass ratio $q = M_2/M_1$, for $0 < q < 1$. The initial orbital semi-major axis distribution is flat in $\log a$. These initial conditions, along with, most notably, a constant star formation rate over the age of the Galaxy, a binary fraction of 100%, solar metallicity across all stars, a common envelope efficiency parameter[3] $\alpha = 3.0$, and zero initial orbital eccentricity, were found by Hurley et al. (2000) to best reproduce the observed numbers of double degenerate, symbiotic, cataclysmic variable and other binary star populations in the Galaxy. Here, we evolve pairs in which both stars have masses between 0.1 and 20 $\mathrm{M}_\odot$, and the initial semi-major axis distribution extends from 2 to 10^5 ZAMS stellar diameters. Note that the BSE code uses the Nauenberg (1972) mass-radius relation for WDs.

We distribute this population according to a double exponential disk model, with scale length 2.75 kpc. The scale height h is chosen to vary according to stellar age t, with $h \propto t^{1/2}$, set equal to 100 pc for stars born today and to 300 pc for the oldest stars in the disk. We use the extinction corrections of Bahcall & Soneira (1980), and thus model the stars in the *Kepler* field of view within the magnitude limits of the survey. We normalize to the number of dwarf stars counted in this field ($\sim 136,000$ with $9 < V < 14$, from *Kepler* Web page; results can be rescaled if this number is later revised). We estimate that the total number of WD–MS systems within this sample of stars is $\sim 15,000$, while the *Eddington* sample should contain $\sim$ 35,000 WD–MS systems. The resulting orbital period distribution of target WD–MS systems is plotted in Figure 2.1, and is seen to display the double-peaked structure expected from § 2.3.1, with a relative dearth of systems with periods

[3] Note that the common envelope formalism used in Hurley et al. (2000) defines the efficiency parameter α as $E_{\mathrm{bind,i}} = \alpha(E_{\mathrm{orb,f}} - E_{\mathrm{orb,i}})$, where $E_{\mathrm{bind,i}}$ is the binding energy of the envelope, and $E_{\mathrm{orb,i,f}}$ are the initial and final orbital energies, respectively. A smaller α corresponds to a lower ejection efficiency and hence a larger loss of orbital energy, and greater orbital shrinkage. Here we use $\alpha = 3.0$, since this is found by Hurley et al. (2000) to best reproduce the Galactic double degenerate population.

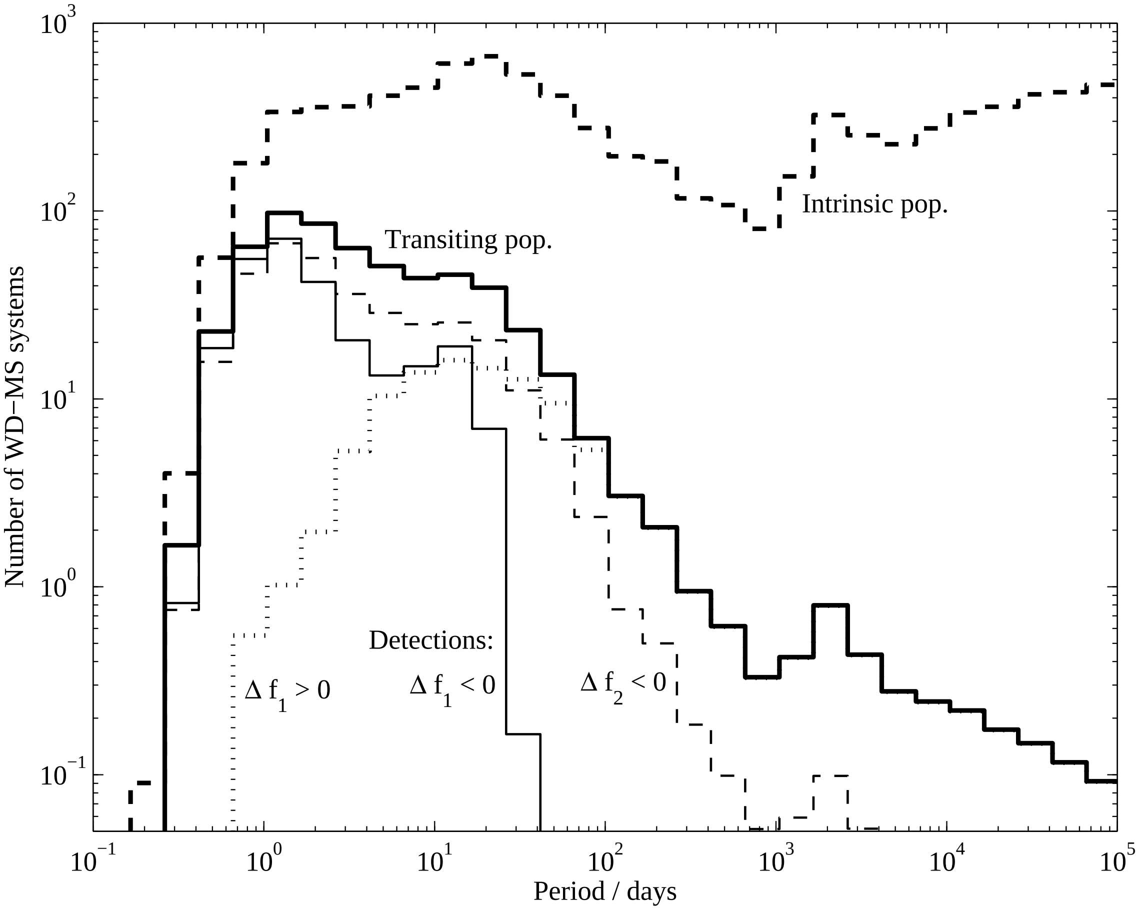

Figure 2.1: The period distribution of WD–MS systems for *Kepler*: thick dashed line: intrinsic WD–MS population in the field of view (generated as described in § 3.2); thick solid line: intrinsic transiting population along our line of sight; thin lines: detectable population using *Kepler* (assuming no stellar variability) — solid line: $\Delta f_1 < 0$; dotted line: $\Delta f_1 > 0$, dashed line: $\Delta f_2 < 0$. Note the dominance of microlensing at longer periods, for which the Einstein radius is larger. The introduction of stellar variability (described in § 2.4.1) limits detectable periods to $\gtrsim$ 2d, and hampers the detection of shallower transits.

from months to years. A similar distribution was predicted by de Kool & Ritter (1993).

2.4 Detecting white dwarfs

The detectability depends on the signal-to-noise ratio of all transits combined in the time series for a given system. We require that the transit duration be at least 1 hour (four *Kepler* time-samples) for detection.

We assume photon counting statistics typically dominate the noise, but also add in a fractional

Mission	Comment	$\Delta f_1 < 0$	$\Delta f_1 > 0$	$\Delta f_2 < 0$
Kepler	Radiative	85	75	190
	Convective	178	26	154
	Conv.+ var.	35	5	28
Eddington	Radiative	19	12	81
	Convective	397	25	299
	Conv.+ var.	81	4	64

Table 2.1: The table gives a summary of detections for *Kepler* and *Eddington*, for survey parameters described in § 2, and assuming broad-band photometry for both. For each mission, expected numbers of WD–MS systems detectable are given separately for systems containing MS stars with radiative envelopes, $M \gtrsim 1.6\mathrm{M}_\odot$, and for systems containing MS stars with convective envelopes, $M \lesssim 1.6\mathrm{M}_\odot$, assuming no stellar variability contribution to lightcurve noise; and for the same convective systems but including the stellar variability noise prescription as described in § 4.1 (Conv.+ var.). In each case the number of systems detectable via the three transit types is given.

instrumental noise of 10^{-5}. For simplicity, we neglect limb-darkening. Shorter period systems are doubly favored, since the time spent in transit and the probability of a transit in a system with random orientation both scale as $P^{-2/3}$. For this reason, *Kepler*'s sensitivity — designed to detect the longer-period terrestrial systems — is easily good enough to detect a large fraction of the transiting WD–MS systems within the magnitude limits at good signal to noise. The same is not true of *Eddington*, with its deeper magnitude limits, which restrict most detectable transits to lower-mass MS primaries. The properties of the detectable systems, assuming an 8σ detection threshold, and the parent population from which they are drawn, are illustrated in figures 2.1, 2.2, and 2.3 for *Kepler*. The total numbers of WD–MS systems detectable are summarised for both *Kepler* and *Eddington* in Table 1. We split the transits into the three modes of detection described in § 2.2. Note that many systems have both detectable primary and secondary transits. For a blind search, a number of transits ($\gtrsim 3$) would need to be seen for detection as a periodic signal amongst the noise, but if radial velocity information were additionally available, it is possible that an identification could be made based on fewer transits. For this reason, systems at all periods are included as detectable (weighted by their transit probabilities and appropriate signal-to-noise ratios); if no such additional information is available, the curves in Figure 1 should be cut off at about $P \sim 1.3$ years, though it can be seen that this makes little difference to the overall number of detectable systems, as a result of the underlying WD–MS period distribution.

We see from Table 1 that several hundred transiting WD–MS systems are in principle detectable with each mission. However, the most serious (and least well-defined) limitation is still to be added: that of stellar variability noise, which is discussed in the following section.

2.4.1 Stellar variability

The issue of stellar microvariability is of concern in terrestrial planet transit surveys (e.g., Jenkins, 2002; Batalha et al. 2002; Aigrain et al. 2002). Variability levels are higher, and hence more of a problem for WD–MS systems. This is because the majority of detectable systems (Fig. 2.1) have short orbital periods, $P \lesssim 30$d, and at these short periods, tidal effects due to the WD are significant (note that the BSE code follows tidal effects in detail). Synchronization of the MS star's rotation with the orbital period is rapid if the MS star has a convective envelope ($M_{\rm MS} \lesssim 1.6{\rm M}_{\odot}$) and $P \lesssim 10$d. This is the case for about half of the transits in principle detectable with *Kepler* (see Fig. 2.2), and most for *Eddington.* Rapidly rotating late-type stars display increased starspot activity (Messina, Rodonò & Guinan, 2001), and hence greater photometric variability, due to these starspots rotating into and out of sight. Individual spots persist for only a few rotational cycles, meaning that over time the variability has a random nature. The amplitude of this variability scales with rotational period approximately as $\sigma \propto P_{\rm rot}^{-1.5}$. Examination of the power spectrum of solar irradiance variations (Fröhlich et al., 1997; Jenkins, 2002) shows that this starspot noise is present up to frequencies $\sim 25/P_{{\rm rot},\odot}$.

Of most importance in transit detection is the stellar noise on the timescale of a transit, since one can in principle filter out variability on other timescales (e.g., Jenkins, 2002; Aigrain et al., 2002). For WD–MS systems with $P \sim 1-10$d, we have $T_{\rm tr} \sim 1-3$h. If a WD transit has $P_{\rm rot} \sim P_{\rm orb} < 25T_{\rm tr}$, then the fractional starspot noise will be $\gg 10^{-4}$, and will drown out almost all WD transit signals, even if there are many transits during the survey lifetime. Effectively this places a lower limit on the orbital period of detectable systems (of around $P \sim 2$d for typical systems). At frequencies $\gtrsim 25/P_{\rm rot}$, the noise is governed by convective (super-)granulation, has an amplitude $\sim 5 \times 10^{-5}$ in the Sun on the appropriate transit timescale, and is likely unaffected by rotation rate. This noise

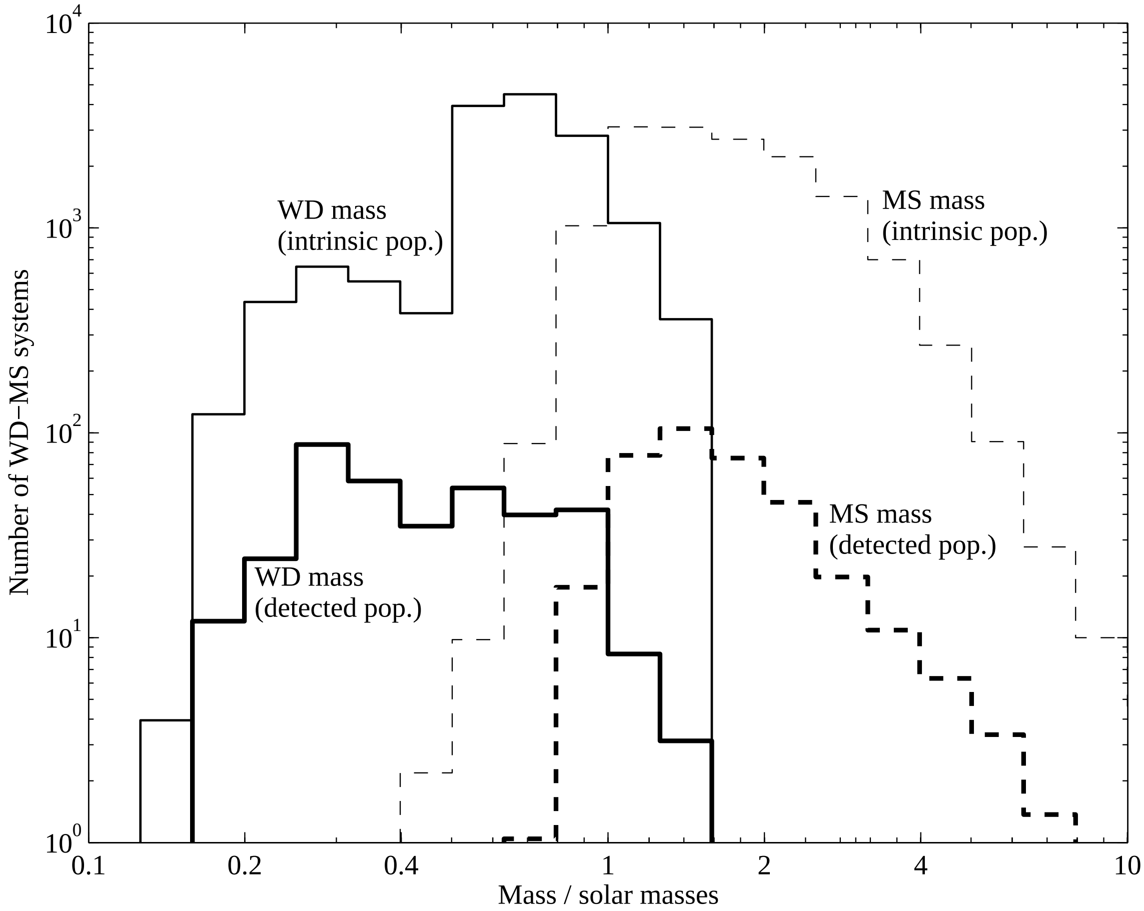

Figure 2.2: The distribution of MS and WD masses for *Kepler*: thin solid and dashed lines: intrinsic distributions in WD and MS masses, respectively, for the *Kepler* field WD–MS population. Thick solid and dashed lines: distributions in detected systems of WD and MS masses, respectively, where detection is of primary transit (WD in front of MS star, Δf_1 positive or negative). Secondary transit distributions are not significantly different. Note that low-mass WDs are preferentially detected, both due to their preferential formation as short-period systems and their larger physical size, making occultations deeper. The MS mass distribution peaks around 1 $M_\odot$, a region traditionally difficult to survey for WD secondaries.

.

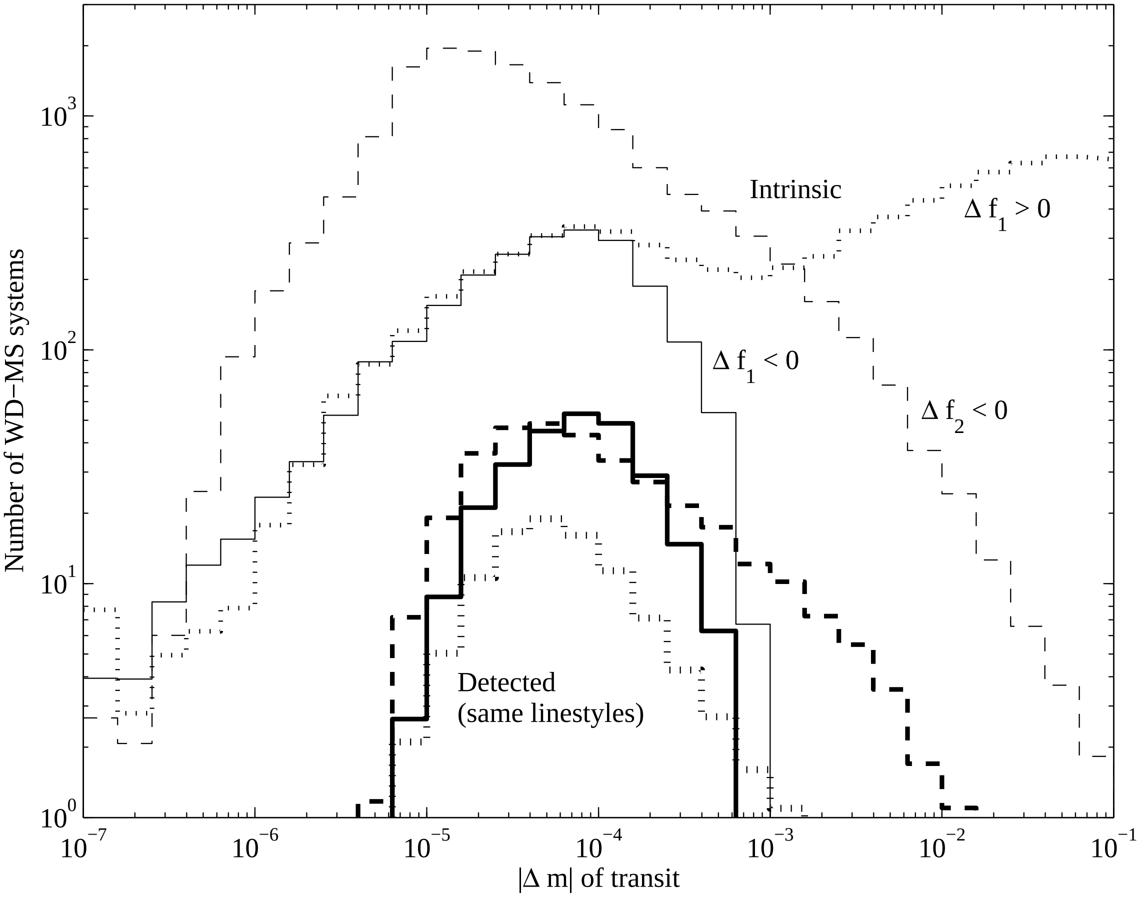

Figure 2.3: The distribution of transit depths for *Kepler*: thin solid, dotted, and dashed lines show intrinsic WD–MS population displaying $\Delta f_1 < 0$, $\Delta f_1 > 0$, and $\Delta f_2 < 0$, respectively (without taking into account transit probability along our line of sight). Thick lines show detectable population, with the same linestyles, assuming no stellar variability. Note the sensitivity cutoff at $\sim 10^{-5}$ magnitudes' flux variation for all transit types, the tail of transits of WD by MS to high $|\Delta m|$ due to young WDs, and the non-detectability of the large $|\Delta m|$ values from microlensing of MS by WD, since these are long-period systems for which transit probabilities and rates are low.

will, however, reduce detectability of faint transits.

We approximate all convective stars as solar in these respects, and in Table 1 show the sizable effects of adding this variability noise upon the detection rates for *Kepler*, again requiring 8σ detections. The *Eddington* mission, which will observe in two colors, may be able to use the color signature of the stellar variability to enhance detection probabilities (Bordé et al., 2003), in which case our predictions without stellar variability may be more appropriate.

More massive MS stars have radiative envelopes, and smaller $|\Delta f|$ for transits, since their radii are larger. For these stars, there is less detailed literature available on stellar microvariability, so we do not attempt to calculate its expected impact on the WD–MS detection rate. We do, however, note that these stars can be intrinsically quite variable. Since the radiative tide is weaker, systems are most likely asynchronous, but Zaqarashvili, Javakhishvili & Belvedere (2002) suggest that in this case, tides can excite the fundamental mode of pulsation of the star, potentially leading to oscillations on roughly the timescale of a transit. We note also that some observations (e.g., Dempsey et al., 1993) suggest that stars in close binaries display greater activity than single stars of the same rotation rate. Thus, the numbers of detectable systems given in Table 1 for MS primaries with radiative envelopes are likely to be reduced by variability, but we have not attempted to quantify this reduction. Extensive data on these topics may only be acquired once space-based transit searches fly.

An additional source of microvariability in the lightcurve may be from flickering as the WD accretes at a low level from the MS star's wind. To have an accretion luminosity $L_{\rm WD} \sim 10^{-4}{\rm L}_\odot$, the WD needs to accrete at a rate $\sim 10^{-13}{\rm M}_\odot{\rm yr}^{-1}$, cf. the solar mass loss rate $\sim 10^{-14}{\rm M}_\odot{\rm yr}^{-1}$. Only a fraction of the mass lost from the MS star will be accreted by the WD, though we note that stellar wind mass loss may be enhanced in close binaries.

2.5 Discussion

As can be seen from Figure 2.1, it is unlikely that WDs will provide a significant source of spurious earth-like transits at $P \sim$ years, given the dearth of WD–MS systems and (hypothesized) large numbers of terrestrial planets at these periods. In addition, primary transits will be microlensing events at these periods. Although WD companions are easily distinguished using radial-velocity observations, it is useful to know that terrestrial planet signatures will not be swamped by those of WDs.

A large sample of close WD–MS systems would enable useful tests of binary star evolution theories (such as common envelope evolution). Also, if the mass and radius of the WD can be separately determined, then the WD mass-radius relation could be tested. The transiting WDs cannot, in general, be observed other than by their dimming effect upon the MS star, so that all properties must be inferred. This approach is, however, independent of WD atmosphere modeling.

The *Kepler* and *Eddington* missions easily have the sensitivity to produce high-quality lightcurves of short-period WD–MS systems, since they are designed to search for longer-period terrestrial planets. The inclusion of stellar variability may affect this somewhat. However, there is still a population of $\gtrsim 50$ WD–MS systems in principle detectable with each, given adequate signal processing power. Many of these systems display both detectable primary and secondary transits, which can be distinguished using their transit profiles. WD transits are longer in duration than, and shaped differently from, grazing MS–MS transits of the same depth. Radial velocity measurements should eliminate blending (dilution of a larger transit depth to the expected WD–MS level due to the presence of a brighter star within the same resolution element) as a source of confusion. This is simpler than in the planetary case, since WDs induce larger radial velocity variations on the orbital timescale. If variability or ellipsoidal modulation of the MS star flux, radial velocity variations, or characteristic accretion luminosity from the WD should draw attention to a system as a candidate close WD–MS system, then transit searches could also be targeted towards these sources, since the geometric transit probability can be of order 10% or more. Provided none of these is a significant noise source on the transit timescale, discovery of the transits in the lightcurve is possible, and system parameters therefore extractable. Although it has been proposed that systems displaying large radial velocity variations be left out of the *Kepler* survey (D. Sasselov, private communication, 2003), the findings of this paper strongly argue for the inclusion of candidate WD–MS systems in the target lists of transit surveys.

It has been noted by Gould, Pepper & DePoy (2002) that the sensitivity of *Kepler* to habitable planets could be significantly increased by pushing the magnitude limit for red MS stars to $V = 17$. Such a change would also be expected to increase the survey's sensitivity to WD–MS systems, since the typical orbital periods of transiting systems are shorter, and so the required signal to noise is more easily achieved. In addition, we expect there to be a large underlying population of late MS star–WD systems available for detection. However, it should be noted that later-type stars tend to be more photometrically variable, which may complicate the situation.

We need 8 quantities to fully characterize a given system: $M_{\rm MS}, R_{\rm MS}, M_{\rm WD}, R_{\rm WD}, P, i, L_{\rm WD}$, and $L_{\rm MS}$. Orbits are expected to be circular. The effect of microlensing complicates the parameter extraction in some sense, since the primary transit depth depends on both WD mass and radius. We have a number of observables from the transit lightcurves: $P, T_{\rm tr}$, and the primary and secondary transit magnitude changes, Δm_1 and Δm_2. Ingress/egress times, $\sim$ minutes in duration, will be unmeasurable if the proposed 15 minute exposures of *Kepler* are used. With radial velocity follow-up, we can also measure $v_{\rm orb,MS}$. More information is clearly necessary to solve for all system parameters. MS modeling can give us $M_{\rm MS}, R_{\rm MS}$, and $L_{\rm MS}$, at least in principle, though this may not be accurate for particularly active stars, or those that have passed through common envelopes. If the distance to the system is known (for example from *Space Interferometry Mission*[4] parallax measurements) then the situation may be improved. If one is able to model the limb-darkening of the MS star in the primary transit then perhaps more information might be extracted. The potentially large numbers of such systems available in transit surveys make feasible statistical tests of WD and binary star theory.

2.6 Conclusions

Since the *Kepler* and *Eddington* missions are designed to detect terrestrial planets, they are ideal for detecting white dwarfs as well. White dwarfs have more modes of detection than planets due

[4] See http://sim.jpl.nasa.gov

to their larger masses (and luminosities), but their detection is complicated by the (tidal) effects of these larger masses. In principle, at least 50 new WD–MS systems might be unambiguously detected with either mission, while at most 500 could be detected if stellar variability were less significant than our estimates. Follow-up observations of the systems might yield mass and radius estimates for the (unseen except by transit) WDs, and hence give a test of the WD mass-radius relation, or of theories of binary star evolution.

Acknowledgements

We thank J. Hurley for providing the BSE code, and S. Seager for an idea leading to this work. Support for E. A. was provided by NASA through Chandra Postdoctoral Fellowship award PF0-10013. We are grateful to the referee, A. Gould, for helpful and detailed comments.

Bibliography

Agol, E., 2002, ApJ, 579, 430

Agol, E., 2003, ApJ, submitted (astro-ph/0303457)

Aigrain, S., Gilmore, G., Favata, F. & Carpano, S., 2002, to appear in Conference Proceedings "Scientific Frontiers in Research on Extrasolar Planets," ed. Drake Dreming (astro-ph/0208529)

Bahcall, J. N. & Soneira, R. M., 1980, ApJS, 44, 73

Batalha, N. M., Jenkins, J., Basri, G. S., Borucki, W. J. & Koch, D., G., 2002, Proc. 1st *Eddington* Workshop, eds Favata, F., Roxburgh, I. W. & Galadí-Enríquez, D.

Bordé P., Leger A., Rouan D., Cameron A. C., 2003, submitted to A&A, astro-ph/0301430

Borucki, W. J., Dunham, E. W., Koch, D. G., Cochran, W. D., Rose, J. A., Cullers, K., Granados, A. & Jenkins, J. M., 1997, ASP Conf Ser, 119, 153, ed. Soderblom, D.

Chandrasekhar, S., 1935, MNRAS, 95, 207

de Kool, M., & Ritter, H., 1993, A&A, 267, 397

Dempsey, R. C., Bopp, B. W., Henry, G. W. & Hall, D. S., 1993, ApJS, 86, 293

Fröhlich, C. et al., 1997, SoPh, 170, 1

Gould, A., Pepper, J., DePoy, D., L., 2003, ApJL, submitted (astro-ph/0211547)

Horne, K., 2002, Proc. 1st *Eddington* Workshop, eds Favata, F., Roxburgh, I. W., & Galadí-Enríquez, D.

Hurley, J. R., Pols, O. R. & Tout, C. A., 2000, MNRAS, 315, 543

Jenkins, J. M., 2002, ApJ, 575, 493

Kroupa, P., Tout, C. A., & Gilmore, G., 1993, MNRAS, 262, 545

Iben, I., Livio, M., 1993, PASP, 105, 1373

Marsh, T. R., 2000, NewAR, 44, 119

Marsh, T. R., 2001, MNRAS, 324, 547

Messina, S., Rodonò, M. & Guinan, E. F., 2001, A&A, 366, 215

Nauenberg, M., 1972, 175, 417

Provencal, J. L., Shipman, H. L., Hog, E., & Thejll, P., 1998, ApJ, 494, 759

Sahu, K. & Gilliland, R. L., 2003, ApJ, 584, 1042

Zaqarashvili, T., Javakhishvili, G. & Belvedere, G., 2002, 579, 810

Chapter 3

The Gravitational Wave Background from Cosmological Compact Binaries

Originally appeared as

Alison J. Farmer & E. S. Phinney,

“The gravitational wave background from cosmological compact binaries”

Monthly Notices of the Royal Astronomical Society, vol. 346, pp 1197–1214, 2003[1]

Reproduced with permission.

[1] MNRAS is a British journal and so the article is reproduced here in British English and adheres to the typesetting conventions of that journal.

Abstract

We use a population synthesis approach to characterise, as a function of cosmic time, the extragalactic close binary population descended from stars of low to intermediate initial mass. The unresolved gravitational wave (GW) background due to these systems is calculated for the 0.1–10 mHz frequency band of the planned Laser Interferometer Space Antenna (LISA). This background is found to be dominated by emission from close white dwarf–white dwarf pairs. The spectral shape can be understood in terms of some simple analytic arguments. To quantify the astrophysical uncertainties, we construct a range of evolutionary models that produce populations consistent with Galactic observations of close WD–WD binaries. The models differ in binary evolution prescriptions as well as initial parameter distributions and cosmic star formation histories. We compare the resulting background spectra, whose shapes are found to be insensitive to the model chosen, and different to those found recently by Schneider et al. (2001). From this set of models, we constrain the amplitude of the extragalactic background to be $1 \times 10^{-12} \lesssim \Omega_{\rm gw}(1\ {\rm mHz}) \lesssim 6 \times 10^{-12}$, in terms of $\Omega_{\rm gw}(f)$, the fraction of closure density received in gravitational waves in the logarithmic frequency interval around f.

3.1 Introduction

Except at very low radio frequencies, most electromagnetic telescopes have good angular rejection, so that faint sources and backgrounds can be seen by looking between bright sources. In contrast all currently implemented gravitational wave detectors, and most of those envisaged for the future, simultaneously respond to sources all over the sky, modified only by a beam pattern of typically quadrupole form. It is therefore important to understand the brightness of the gravitational wave sky, since this will limit the ultimate sensitivity attainable in gravitational wave astronomy. One immediate pressure to understand this background comes from the need to set design requirements for the ESA/NASA Laser Interferometer Space Antenna (LISA) mission (LISA mission documents and status may be found at http://lisa.jpl.nasa.gov/ and http://sci.esa.int/home/lisa/).

In this paper, we attempt to predict the gravitational wave background produced by all the binary stars in the universe, excluding neutron stars and black holes. This is believed to be the principal source of gravitational wave background in the frequency range $10^{-5} < f < 10^{-1}$ Hz. Below 10^{-5} Hz, the background is probably dominated by merging supermassive black holes, and above 10^{-1} Hz, it is probably dominated by merging neutron stars and stellar mass black holes (whose complicated and poorly understood formation histories and birth velocities make predictions more uncertain, cf. Belczynski, Kalogera & Bulik 2002).

Besides the extragalactic background, there is also a Galactic background produced by the binary stars in our Milky Way (Evans, Iben & Smarr 1987; Hils, Bender & Webbink 1990; Nelemans, Yungelson & Portegies Zwart 2001c). Although the Galactic background is many times larger in amplitude than both the extragalactic background and LISA's design sensitivity, the individual binaries contributing to it can be (spectrally) resolved and removed at frequencies above $\sim 3 \times 10^{-3}$ Hz (Cornish & Larson, 2002). Below this frequency they cannot be removed (at least in a mission of reasonable lifetime $\sim$ 3 years), but the unresolved Galactic background will be quite anisotropic. As the detector beam pattern rotates about the sky, the Galactic background will thus be modulated, while the isotropic (or nearly so; see Kosenko & Postnov 2000) distant extragalactic background will not. Modelling of the angular distribution of the Galactic background using both *a priori* models and the observed distribution of higher frequency resolved sources will thus allow the Galactic background to be subtracted to some precision (Giampieri & Polnarev 1997). In addition, there will be anisotropies due to the distribution of local galaxies, at the level of 10 per cent of the distant extragalactic background from the LMC, and at the per cent level from M31 or the Virgo cluster (see also Lipunov et al. 1995).

The immediate motivation for this work is a design issue for LISA. One of LISA's major science goals (see the LISA Science Requirements document at http://www.tapir.caltech.edu/listwg1/) is the detection of gravitational waves from compact objects spiralling into supermassive black holes (Finn & Thorne 2000; Hils & Bender 1995), since these can provide precision tests of strong field relativity and the no-hair theorem (Hughes, 2001). However, these signals are weak, and their templates not yet fully understood. It has thus been proposed that LISA should be designed with somewhat greater sensitivity to increase the probability that these signals are detected. However, this would be pointless if the principal background were cosmological rather than instrumental. As we shall see (Fig. 3.16), we find that this is most probably almost, but not quite the case at the relevant frequencies ($4-10$ mHz). So there would be a point to increasing LISA's sensitivity in the $4-10$ mHz range, but not to increasing it by more than a factor of 3 in gravitational wave amplitude h (9 in $\Omega \propto f^2h^2$).

A second motivation for this work comes from the fact that this background is an astrophysical *foreground* to searches (both with LISA and with future detectors with extended frequency range and sensitivity) for backgrounds produced in the very early universe. Gravitational waves from bubble walls and turbulence following the electroweak phase transition are expected to be in the LISA frequency band, with amplitude that could be well above LISA instrumental sensitivity (Kamionkowski, Kosowski & Turner 1994; Kosowsky, Mack & Kahniashvili 2002; Apreda et al. 2002). Another potential source of isotropic gravitational waves in the LISA band are those produced when dimensions beyond the familiar four compactified, which occurred when the universe had temperature $kT >$TeV (Hogan, 2000).

Note that detection of a gravitational wave background can possibly be made even if it is considerably below the noise limit of the LISA detectors shown in our Fig. 3.16. This can be done by comparing the signals from Michelson beam combinations (sensitive to instrument noise and gravitational waves) with Sagnac beam combinations (sensitive to instrument noise, but insensitive to gravitational waves), thus calibrating the instrumental noise—cf. Tinto, Armstrong & Estabrook (2001), Hogan & Bender (2001).

Gravitational waves are the only directly detectable relic of inflation in the early universe, and their detection over a range of frequencies would provide a valuable test of models of inflation (Turner, 1997). It has been proposed that advanced space-based gravitational wave detectors might search for the background of gravitational waves from inflation. The gravitational waves from slow-roll inflation models contribute to the critical density in the universe $\Omega_{\rm gw} < 10^{-15}$ per octave of frequency. We shall see (Fig. 3.8, 3.17) that the gravitational wave background from cosmological binaries makes such detection impractical except at frequencies below 10^{-5} Hz (where supermassive black holes continue to make it impossible), or above 0.1 Hz.

A third motivation is that a detection of the extragalactic binary background, e.g. by LISA, would set an independent (and unaffected by dust extinction) constraint on a combination of the star formation history of the universe and binary star evolution.

There have been previous estimates of the extragalactic binary background. Hils, Bender & Webbink (1990) made detailed estimates of the Galactic binary background, and estimated that the extragalactic background from close double white dwarf pairs should be about 2 per cent (in flux or Ω units) of the Galactic background. This estimate was refined, using more modern star formation histories, by Kosenko & Postnov (1998), who found instead a level of ~ 10 per cent. Schneider et al. (2001) used a descendant of the Utrecht population synthesis code to estimate the extragalactic binary background as a function of frequency, and claimed that the background should have a large peak at $\sim 3 \times 10^{-5}$ Hz, just below the frequency at which typical binaries have a lifetime that equals the age of the Universe.

We have followed the spirit of this previous work, but with an independent binary population synthesis code. More importantly, we have devoted much effort to the normalisation of the background, to understanding the contributions of different types of binaries and their formation pathways to the background, and to estimating the uncertainties in all of these, so that we can have a better idea of the sources and level of uncertainty in the predicted background.

The paper is organised as follows: In section 3.2, we describe the gravitational wave (GW) emission from a binary system, then in section 3.3 we outline the main evolutionary pathways to the close double degenerate (DD) stage, which we shall see is the dominant source of GW background in the LISA band. In section 3.4, we use the preceding sections to make some simple analytic arguments about the nature of DD inspiral spectra. We describe the use of the BSE code in our population synthesis, in section 3.5, then go on to construct a set of synthesis models whose results we test against the observed Galactic DD population. We also motivate some modifications made to the prescription for the evolution of AM CVn stars in the BSE code. In section 3.6, we present the cosmological integrals used in the code, along with the cosmic star formation history and overall normalisation chosen. Section 3.7 is devoted to a discussion of the GW background spectra produced by our code, in terms of the systems contributing to the background and the progenitors of these sources. We also discuss the differences between our population synthesis models. In section 3.8, we place limits on the maximum and minimum expected background signals, and compare these with the LISA sensitivity and in section 3.9 with previous work. In section 3.10 we summarise and conclude.

3.2 Gravitational waves from a binary system

A binary system of stars in circular orbit with masses M_1 and M_2 and orbital separation a emits gravitational radiation, at the expense of its orbital energy, at a rate given by (Peters & Mathews, 1963)

$$\begin{aligned} L_{\rm circ} &= \frac{32}{5}\frac{G^4}{c^5}\frac{(M_1M_2)^2(M_1+M_2)}{a^5} \\ &\simeq 1.0\times 10^{32}\frac{(M_1'M_2')^2(M_1'+M_2')}{(a')^5}\ \mathrm{erg\ s}^{-1}, \end{aligned} \tag{3.1}$$

where primes denote quantities expressed in solar units, i.e. $M/\mathrm{M}_\odot$, $a/\mathrm{R}_\odot$. The gravitational radiation is emitted at twice the orbital frequency ν of the binary, $f_{\rm circ} = 2\nu = \Omega/\pi$.

If the binary is eccentric with eccentricity e, this expression must be generalised to include emission at all harmonics n of the orbital frequency, $f_n = n\nu = n\Omega/2\pi$, where $\Omega = (a^{-3}G(M_1 + M_2))^{1/2}$. The luminosity in each harmonic is given by

$$L(n,e) = g(n,e)L_{\rm circ}, \tag{3.2}$$

where $L_{\rm circ}$ is the luminosity of a circular binary with separation a, as given in Eq. 3.1, where a is now the relative semi-major axis of the eccentric orbit, and the $g(n,e)$ are defined in eq. (20) of Peters & Mathews (1963). The total specific luminosity $L_f = dL(f)/df$ of the system is then a sum over all harmonics:

$$L_f(e) = L_{\rm circ}\sum_{n=1}^{\infty} g(n,e)\delta(f - n\nu). \tag{3.3}$$

The total luminosity is

$$L = L_{\rm circ}\sum_{n=1}^{\infty} g(n,e) = \frac{1 + \frac{73}{24}e^2 + \frac{37}{96}e^4}{(1-e^2)^{7/2}}L_{\rm circ}. \tag{3.4}$$

For eccentric orbits, the emission spectrum of Eq. 3.3, $g(n,e)L_{\rm circ}$ as a function of $f = n\nu$ consists of points along a skewed bell-shaped curve with maximum near the relative angular velocity at pericentre, where the greatest accelerations are experienced ($2\pi\nu \sim \Omega_p$, where Ω_p is the angular velocity of the relative orbit at pericentre, v_p/r_p). In terms of harmonic number, a good approximation for all e (becoming very good for $e > 0.5$) is that L_f peaks at $n = 1.63(1-e)^{-3/2}$, and fL_f peaks at $n = 2.16(1-e)^{-3/2}$.

3.3 Evolution to the DD stage

We shall see that the GW background is dominated by the emission from close double degenerate (DD) binaries at frequencies $10^{-4} \lesssim f \lesssim 10^{-1}$ Hz. In this work, the term DD will refer to WD–WD pairs and loosely to WD–naked helium star pairs, i.e. we exclude neutron stars from our definition.

In this section we describe the two main evolutionary pathways from the zero-age main sequence (ZAMS) to the close DD stage. The route followed depends mainly on the initial orbital separation of the ZAMS stars. Similar descriptions can be found in e.g. Webbink & Han (1998).

We begin with an intermediate-mass ZAMS binary system with primary mass M_1, secondary mass M_2 ($< M_1$), semi-major axis a and eccentricity e. The orbit may evolve somewhat due to tidal interactions between the stars, particularly if they have convective envelopes. When the primary evolves off the main sequence and swells in size, it may fill its Roche lobe and start to transfer matter on to the secondary. The stability of this mass transfer determines which of the two main pathways to the DD stage is commenced.

3.3.1 CEE+CEE

If the primary fills its Roche lobe when it has a deep convective envelope (i.e. on the red giant branch (RGB) or asymptotic giant branch (AGB)), then for mass ratios $M_1/M_2 \gtrsim 0.6$, the ensuing mass transfer is dynamically unstable (for conservative transfer). The envelope of the primary spills on to the secondary on a dynamical timescale, leading to the formation of a common envelope, inside which orbit the secondary and the core of the primary. The envelope is frictionally heated at the expense of the stars' orbital energy, until eventually either they coalesce, or the envelope is heated sufficiently that it is ejected from the system, leaving the primary's core (a hot subdwarf that will rapidly cool to become a WD, or if the primary was on the RGB and had mass $M_1 \gtrsim 2\ \mathrm{M}_{\odot}$, then a helium star that will evolve to the WD stage). The basic idea of the common envelope phase is well-accepted and observationally motivated, though not well-simulated (see e.g. Livio & Soker 1988; Iben & Livio 1993; Taam & Sandquist 2000). Several formalisms have been proposed to model it in population synthesis studies. The evolution code used here (see section 3.5.1) follows closely the prescription of Tout et al. (1997) (originally from Webbink 1984), in which

$$E_{\mathrm{bind,i}} = \alpha(E_{\mathrm{orb,f}} - E_{\mathrm{orb,i}}), \tag{3.5}$$

where $E_{\rm bind,i}$ is the initial binding energy of the envelope of the overflowing giant star (or the sum of both envelopes' binding energies if both stars are giants), parametrized by $E_{\rm bind,i} = -G/\lambda(M_1 M_{\rm env,1}/R_1)$, where λ is of order unity, and is calculated in the BSE code (see section 3.5.3). $E_{\rm orb,i}$ and $E_{\rm orb,f}$ are, respectively, the initial and final orbital binding energies of the core-plus-secondary system, and α is the so-called common envelope efficiency parameter, also of order unity, usually taken to be a parameter to be fitted to observations. Variations to this prescription will be considered in sections 3.5.2 and 3.5.3.

Continuing with the system's evolution, the secondary star later evolves off the main sequence, and a second common envelope phase is likely to occur, leading to further orbital shrinkage. If once again the stars do not coalesce then we will be left with a close(r) pair of remnants, one or both of which may be helium stars, which in time will evolve to the WD stage. (It is not uncommon for either helium star to overflow its Roche lobe upon leaving the helium main sequence; this can lead to either stable mass transfer or to a further common envelope phase.) In this picture, the second-formed WD will be the less massive of the pair, since the giant star from which it descended had a smaller core mass when its core growth was halted as it lost its envelope.

3.3.2 Stable RLOF+CEE

If Roche lobe overflow occurs when the primary is in the Hertzsprung gap, that is after the primary has exhausted its core hydrogen and before it has developed a deep convective envelope and ascended the giant branch, then Roche lobe overflow may be dynamically stable for moderate mass ratios, and a phase of stable but rapid mass transfer can occur. In this way, the primary transfers its envelope to the secondary, leaving a compact remnant, and a common envelope phase is avoided, since by the time the primary evolves to the giant branch, the mass ratio has been sufficiently inverted that mass transfer remains dynamically stable. The orbital separation will typically have increased during this phase (for conservative mass transfer at least), since much of the transfer was from the less-massive to the more-massive star. When the secondary evolves off the main sequence, it will most likely fill its Roche lobe on the RGB, so that a common envelope phase ensues, and a close DD is born, provided that the resulting orbital shrinkage does not lead to coalescence. The second-formed WD will this time be the more massive, since its progenitor was the more evolved at the time of its overflow.

The initial conditions for this route occupy a smaller range in initial orbital semimajor axis than the CEE+CEE route, but as it results in the injection of DD systems only at very short periods, we expect both pathways to be significant contributors to the close DD population, i.e. those systems contributing to the GW background in the LISA waveband. We note also that both routes ought to lead to the production of DDs with circular orbits, even if the ZAMS eccentricity was non-zero, since tidal circularisation is rapid when a system contains a near-Roche lobe-filling convective star.

3.4 Analytic arguments about spectral shape

Given only the above, we can make some predictions as to the shape of the GW spectrum seen today. A somewhat analogous treatment is given in Hils et al. (1990). We consider the evolution under GW emission of a population of DDs after creation as in section 3.3, with circular orbits. We deal here with detached systems; the spectral shape due to interacting pairs is discussed in section 3.7.1.2.

Here and throughout, we use ν for orbital frequencies and f for gravitational wave frequencies. For circular orbits, $f = 2\nu$.

The number density $N(\nu, t)$ of binary WDs per unit orbital frequency interval at time t must obey the continuity equation

$$\frac{\partial N}{\partial t} + \frac{\partial}{\partial \nu}(\dot{\nu} N) = \dot{N}_{\mathrm{b}}(\nu, t), \tag{3.6}$$

where $\dot{N}_{\mathrm{b}}(\nu, t)$ is the birth rate (after nuclear evolution and mass transfer) of WD–WD systems per unit frequency. Now for a given source, we know that $\dot{E}_{\mathrm{orb}} = -L_{\mathrm{gw}}$, and using Eq. 3.1 along with

$E_{\rm orb} = -GM_1M_2/(2a)$ and Kepler's law, we obtain

$$\begin{aligned} \dot{\nu} &= \frac{96}{5}(2\pi)^{8/3}\left(\frac{G\mathcal{M}}{c^3}\right)^{5/3}\nu^{11/3} \\ &\equiv K\nu^{11/3} \\ &\simeq (3.7\times10^{-6}{\rm s}^{-2})(\mathcal{M}/{\rm M}_\odot)^{5/3}\nu^{11/3}, \end{aligned} \tag{3.7}$$

where we have used the definition of the chirp mass $\mathcal{M}$,

$$\mathcal{M} \equiv \frac{M_1^{3/5}M_2^{3/5}}{(M_1+M_2)^{1/5}}. \tag{3.8}$$

We solve Eq. 3.7 to give the evolution $\nu(t)$ for $\dot{N}_{\rm b} = \delta(t-t', \nu-\nu')$, i.e. for a single source injected at frequency ν' at time t',

$$\nu(t)^{-8/3} - \nu'^{-8/3} = 8K(t'-t)/3. \tag{3.9}$$

The corresponding source number density (Green's function for Eq. 3.6) $N_G(\nu, t; \nu', t')$ as a function of time is given by

$$\begin{aligned} N_G\, d\nu &\propto dt(\nu) \\ &\propto \frac{d\nu}{\mathcal{M}^{5/3}\nu^{11/3}}\,\delta\left(t - \left[t' + \frac{3}{8K}(\nu'^{-8/3} - \nu^{-8/3})\right]\right), \end{aligned} \tag{3.10}$$

since, as the system traces out a path in ν, it spends a time at each point inversely proportional to its velocity $\dot{\nu}$ through frequency space.

We then consider a real injection spectrum $\dot{N}_{\rm b}(\nu', t')$, for $\nu_{\rm min} < \nu' < \nu_{\rm max}$. The resulting number density $N(\nu, t)$ is given by

$$N(\nu,t) = \int_{\nu_{\rm min}}^{\nu_{\rm max}}\int_0^t \dot{N}_{\rm b}(\nu', t')\, N_G(\nu, t; \nu', t')\, dt'\, d\nu'. \tag{3.11}$$

Since $L_{\rm gw} \propto \nu^{10/3}\mathcal{M}^{10/3}$, we can then construct the GW emission spectrum by taking $F_{\rm gw}(f,t) \propto f^{10/3}\mathcal{M}^{10/3}N(f,t)$.

The choice of DD injection spectrum is therefore instrumental in determining the shape of the GW emission spectrum. We can estimate its shape as follows: we will later choose to distribute ZAMS orbital semimajor axis uniformly in $\log a$, i.e. also uniformly in $\log \nu$, for given initial M_1 and M_2. We suppose that, for at least the CEE+CEE route (see section 3.3), the common envelope phases lead to some mean orbital shrinkage factor, so that WD–WD pairs at their birth are also distributed roughly uniformly in $\log \nu$. We then have $\dot{N}_b(\nu') \propto 1/\nu'$, from some $\nu_{\rm min} \ll \nu$ of interest, up to $\nu_{\rm max}$ (see also fig. 1 of Webbink & Han 1998). This is the maximum orbital frequency at which a system can exit a common envelope phase and survive to become a WD–WD pair. Upon CE exit, the newly exposed stellar core will be a hot subdwarf, larger than the WD it will cool to become, or it could be a naked helium star, which will eventually evolve to the WD stage. The maximum injection frequency at WD–WD birth is set by the minimum orbital separation that will keep this object (and the first-formed WD) from overflowing its Roche lobe on the way to the WD stage, whether this is at the exit of common envelope or (applicable to the helium star case) as its radius changes due to nuclear evolution.

For illustration, we compute the emergent spectrum for a fiducial population of 0.5 $\rm M_\odot$ WD–WD pairs. The radius of a 0.5 $\rm M_\odot$ naked helium star does not exceed ~ 0.13 $\rm R_\odot$ on its way to the WD stage, which sets $\nu_{\rm max} \sim 0.7$ mHz. If we then assign a constant pair formation rate, so that $\dot{N}_b(\nu',t') = \dot{N}_b(\nu')$, and perform the integral in Eq. 3.11, we obtain the spectral shape shown in Fig. 3.1. Note that the spectrum is truncated at a frequency above which the inspiralling WDs would undergo Roche lobe overflow and merge, $f_{\rm merge} = 2\nu_{\rm merge} \simeq 40$ mHz.

If instead we only inject sources for $0 < t < \tau$, and look at the spectrum obtained for $t > \tau = 1$ Gyr, (Fig. 3.2), we see that the basic spectral shape is little affected.

Because of the strong dependence of $\dot{\nu}$ on ν, a given system of specified age will either have merged or will have remained at essentially constant separation. Thus there are two clear physical regimes displayed in the spectra, separated by the injection frequency from which a source could

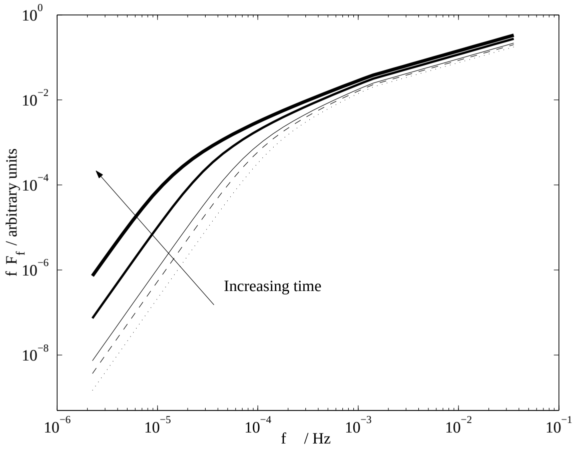

Figure 3.1: Gravitational wave spectrum arising from constant WD–WD formation rate, at times 2, 5, 10, 100 and 1000 Gyr, increasing in the direction of the arrow shown.

have reached contact due to GW losses in the time t since its birth, $\nu_{\rm crit} \simeq 0.03$–$0.04$ mHz for $t \sim 5$–10 Gyr. (In all relevant situations for us, $\nu_{\rm crit} < \nu_{\rm max}$.)

At $f < 2\nu_{\rm crit}$ lies a "static regime", in which losses due to GW are negligible in the time available, giving $N(\nu) \propto \nu^{-1}$ and hence $fF_{\rm gw} \propto f^{10/3}\mathcal{M}^{10/3}$. For $f \gtrsim 2\nu_{\rm crit}$, we are in the "spiral-in" regime. In the case of a burst of DD formation (Fig. 3.2), sources simply sweep through this region on the way to merger, so that we have $N(\nu) \propto \nu^{-11/3}\mathcal{M}^{-5/3}$, giving $fF_{\rm gw} \propto f^{2/3}\mathcal{M}^{5/3}$. If we have a constant DD formation rate (Fig. 3.1), then for $2\nu_{\rm crit} < f < 2\nu_{\rm max}$, merging systems are continually being injected, so that $N(\nu)$ is less steeply decreasing than $\nu^{-11/3}$ in this region. For $f > 2\nu_{\rm max}$ the spectral slope is again $2/3$. Reality will be some combination of these histories.

We therefore expect the cosmological spectrum we calculate later (section 3.6) to be composed of a superposition of curves of these shapes, modified for chirp mass variations, redshift effects and time delay between progenitor star formation and DD formation. The detailed calculations described in following sections follow in detail the evolution of all sources from ZAMS to merger, and do not rely upon approximate treatments of the kind given above. Simple estimates of the background amplitude are discussed in section 3.7.

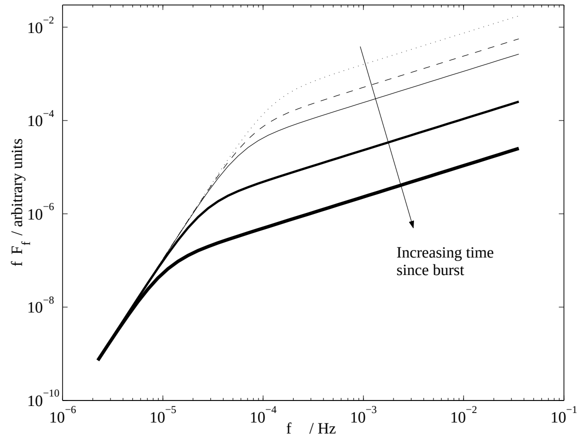

Figure 3.2: Gravitational wave spectrum arising from a burst of WD–WD formation between 0 and 1 Gyr. Curves plotted are spectra at times 2, 5, 10, 100 and 1000 Gyr, increasing in the direction of the arrow shown.

3.5 Model construction

3.5.1 The *BSE* code and population synthesis

The rapid evolution code BSE (Hurley, Tout & Pols, 2002) is used throughout this work whenever a binary system is evolved. This code is a fit to detailed models of stellar evolution, and produces an evolutionary time-sequence $x(t_j)$ of the properties x of any input ZAMS binary system. The code's time-resolution adapts to the shortest current timescale for change of the system components and orbit, due to e.g. nuclear evolution, angular momentum loss or mass transfer, which are all treated iteratively and have finite duration. In this way, even the most fleeting of evolutionary phases is captured in detail, without requiring excessive time-resolution during long phases in which little changes. This is especially useful in the study of gravitational waves, since the majority of the GW emission from a given system occurs over an inspiral timescale much shorter than the nuclear timescales of the binary's parent ZAMS stars. Some of the most relevant features of the BSE code will be described in the following section; see Hurley et al. (2002) for full details.

The output $x(t_j)$ from the code can be used to construct a stellar population at time T as follows. This method is similar to that used by Hurley et al. (2002) to characterise the Galactic binary population.

We describe the ZAMS binary parameter space in terms of the primary (larger) mass M_1, secondary mass M_2 (or mass ratio $q = M_2/M_1 \leq 1$), orbital semi-major axis a and orbital eccentricity e. We divide this space into grid boxes, and from each box k, we randomly choose a ZAMS system to represent the evolution of all sources in that box.

The number $\mathcal{P}_k$ of sources born into box k per unit binary system realised is determined by probability distributions $A(a)$, $\Xi(e)$ and $\Phi(M_1, M_2) = \Phi_1(M_1)f(q)$ in the ZAMS system properties described above (see section 3.5.3). $\mathcal{P}_k$ is obtained by integrating the product of these distribution functions over the extent of box k.

We wish to construct the population of sources present at time T. For each output timestep t_j, the system with properties $x(t_j)$ can be viewed as a system born between times $(T - t_{j+1})$ and $(T - t_j)$. If at this point the star formation rate was $\mathcal{R} = \mathcal{R}(T - t_j)$ (expressed as a number of binary systems born per unit time), then the number of systems with properties $x(t_j)$ we expect to see at time T is given by

$$\mathcal{N}_{j,k}(T) = (t_{j+1} - t_j)\mathcal{R}(T - t_j)\mathcal{P}_k, \tag{3.12}$$

so long as $T > t_j$, so that stars were not born before time began. We perform this calculation for all boxes k and all timesteps j, so that the total population at time T is given by the combination of all $\mathcal{N}_{j,k}(T)$.

This method of population synthesis ensures that sources from even unlikely regions of ZAMS parameter space are represented, weighted by their low formation probability. Coupled with the adaptive time-resolution of the BSE output, and a sufficiently fine grid spacing, this technique allows the synthesis of a statistically reasonable population in a modest amount of computing time. Alternatively, statistical accuracy can be ensured with a Monte Carlo approach by simply generating a large enough number of stars under the initial distribution functions (see Belczynski et al. 2002).

Our grid extends from 0.08 to 20 M$_\odot$ in the mass of the primary M_1, and from 0.08 M$_\odot$ to M_1 in the secondary's mass. The initial separation is gridded from $2(R_1 + R_2)$ to $10^5(R_1 + R_2)$, where R_1 and R_2 are the ZAMS radii of the primary and secondary respectively. We find that our background fluxes are statistically accurate to around one per cent if we choose grid spacings of 0.05 in $\ln M$ for each mass and 0.1 in $\ln a$ for the separation. This corresponds to evolving $\sim 7 \times 10^5$ binaries. For the Galactic tests described in section 3.5.3 we find that it is sufficient to use a grid spacing twice as large in each dimension.

The BSE code has previously been tested against various Galactic populations of binary stars (Hurley, 2000). A set of input parameters and distributions is recommended for use with the code, to best reproduce the observed Galactic binary population as a whole. However, in this work we are keen to quantify the effects of astrophysical uncertainties upon population synthesis calculations of the GW background, and so in the following subsections we construct a set of models that differ in their choice of input parameters but produce specifically a Galactic DD population not in conflict with observations. The current observational uncertainties about DDs admit a range of models. This set of models is then considered representative of the population synthesis uncertainties affecting the GW background.

3.5.2 The state of observations

The observations of DD stars are currently undergoing a revolution. Full results of this revolution have not yet been published, so the detailed comparison of synthesised populations with observations is still difficult.

Marsh (2000) reported on the 15 then known DDs with measured periods, six of which had measured component mass ratios (Maxted, Marsh & Moran, 2002). Searches for DDs have mainly focussed on low-mass WDs, $M_{\rm WD} \lesssim 0.5$ $\mathrm{M}_\odot$ (e.g. Marsh, Dhillon & Duck 1995), since these must have formed through giant stars losing their envelopes in binary systems, before the helium burning that would inevitably occur in a single star. Maxted & Marsh (1999) determined that the fraction of DDs among these DA WDs is between 1.7 and 19 per cent, with 95 per cent confidence. Statistical comparisons with population synthesis models are thus difficult, given the sample size and level of bias, but there are some notable disagreements between observations and theory that are not easily explained in terms of selection effects. The first of these is the lack of observed very low-mass He WDs ($M \sim 0.25$ $\mathrm{M}_\odot$). Theory predicts an abundance of such sources. Nelemans et al. (2001b) suggest that this can be explained by a more rapid cooling law for low-mass WDs than is commonly used. The second discrepancy is in the distribution of known DD mass ratios, which is seen to peak near unity (Maxted et al., 2002). Even considering selection effects (Nelemans et al., 2001b), this is difficult to explain in terms of either standard DD formation route, since as described in Section 3.3, the WD masses are expected to differ significantly.

This prompted Nelemans et al. (2000) to suggest an alternative scenario in which a common envelope phase between a giant and a main sequence star of similar mass does not result in a substantial spiral-in of the orbit, meaning that the second common envelope phase does not occur until the secondary's radius is larger (relative to that of the primary when it filled its own Roche lobe) than in the standard CEE+CEE picture, so that the second WD formed is more massive, closer to the mass of the first-formed WD. They motivate this choice by parametrizing in terms of an angular momentum, rather than an energy balance (cf. section 3.3).

The observational sample of DDs is currently being substantially increased by the SPY project (Napiwotzki et al., 2002), a spectroscopic study of $\sim$ 1500 apparently single WDs (not restricted to low mass) to search for radial velocity variations indicative of binarity. Napiwotzki et al. (2002) report that of the 558 WDs surveyed so far, 90 (16 per cent) show evidence for a close WD companion. Of these, mass-ratio determinations are reported for three DDs (Karl et al., 2002), these three continuing the observed trend of mass ratios near unity.

The results of the SPY project, once analysed fully, will help to constrain DD population synthesis calculations in a greatly improved way. However, given the preliminary and partial nature of the results so far, we can make only rather broad statements about their compatibility with any given synthesised Galactic population. This process is described in the next section.

3.5.3 Candidate models

Our fiducial population synthesis model (Model A) is similar to the preferred model suggested by Hurley (2000) (also his Model A): we use the initial mass function (IMF) of Kroupa, Tout & Gilmore (1993) (KTG) for $\Phi(M_1)$, we distribute M_2 uniformly in the mass ratio $q = M_2/M_1$, $f(q) = 1$, and we start with a flat distribution in $\log a$, choosing our limits as $2(R_1 + R_2) < a < 10^5(R_1 + R_2)$, where R_1 and R_2 are the ZAMS radii of the primary and secondary, respectively. We have tidal effects switched "on", we use $\alpha = 3.0$ for the common envelope efficiency parameter, and we assign all stars solar metallicity, $Z = 0.02$. For the Galaxy, we adopt the constant star formation rate $\mathcal{R}$ over the past 10 Gyr which gives a stellar disk mass of 6×10^{10} $\mathrm{M}_\odot$ today.

We differ from Hurley's Model A in three main ways: first, we assign an initial binary fraction of 50 per cent (cf. Hurley's 100 per cent) since this is observed locally to be the case (Duquennoy & Mayor, 1991) and we evolve a set of single stars alongside the binaries, distributed according to the same IMF as the binary primaries. Second, we assign a ZAMS orbital eccentricity e to all systems, according to a thermal distribution $\Xi(e) = 2e, 0 < e < 1.0$. Hurley (2000) finds that an $e = 0$ model gives a somewhat better fit to observations (though he finds that the numbers of *close* ($P < 10$ d) DD systems produced are not affected); we will also test a model of this type as part of our parameter variation (see below). Lastly, Hurley's Model A assumed the envelope binding energy parameter $\lambda = 0.5$ for all stars, whereas here we allow this parameter to be calculated in the code (values of λ are from fits to detailed models of stellar evolution by O. Pols and are an addition to the code described in Hurley et al. 2002; J. Hurley, private communication, 2003), and in addition we include 50 per cent of the envelope's ionisation energy in its binding energy.

We test our synthesised Galactic populations against observations in a necessarily simple way. The aim is to reject models in clear conflict with the observed population of double degenerate stars, and to admit all others as representative of the uncertainties in DD population synthesis. Since the overall normalisation for the cosmological integral will be entirely separate from that used for the Galaxy, we choose primarily to compare relative populations as opposed to absolute numbers of Galactic sources. An ideal criterion is the fraction among field WDs of close DD binaries, which currently available SPY results place at 16 per cent. Since the sample size is substantially larger than that of Maxted & Marsh (1999), we adopt the SPY data, despite their incompleteness. We assume a negligible false-positive rate for SPY, and approximate the survey as magnitude-limited ($V < 15$) for the purposes of comparison. The somewhat approximate Galactic model and star formation history used here are sufficient, given the generosity of our selection criteria and the fact that we compare fractional quantities wherever possible.

We distribute all stars according to a simple double exponential Galactic disk model (scale height 200 pc, scale radius 2.5 kpc), then calculate the fraction of WDs with $V < 15$ expected to be members of DD binaries with $P < 100$ d. We then require that this calculated fraction be at least 10 per cent, if a given model is to be accepted. We assign a lower limit only, since our calculated binary fractions are likely to be overestimates, for several reasons. First, 100 d is a generous upper limit to the orbital periods detectable with SPY; second, we do not address the issue of the substantial lack of observed low-mass (hence binary-member) WDs found in other population synthesis studies; and finally, the cooling curves used are the simple Mestel curves from Hurley et al. (2002); if we instead use the "modified Mestel cooling" from Hurley & Shara (2003), which better fits the theoretical curves of Hansen (1999), then our calculated binary fraction decreases by a few percent. For our fiducial Model A, with Hurley et al. (2002) cooling, we find that 18 per cent of field WDs will show up as DDs in such a survey, in reasonable agreement with the SPY results.

We also find a local total space density of WDs of 9×10^{-3} pc^{-3}, and compare this with observational values, which range from $\sim 4 - 20 \times 10^{-3}$ pc^{-3} (Nelemans et al., 2001b, and references therein). We do not attempt to compare to distributions in mass, mass ratio or period in detail: the observed distributions are subject to complex selection effects, and turn out often to be most constraining for WD cooling models (e.g. Nelemans et al., 2001b), whose development is beyond the scope of this paper. We note however that in a volume-limited sense, the mean mass ratio (where $q < 1$ by definition) for detached WD–WD pairs is $\langle q \rangle = 0.62$, not in good agreement with observations, but in common with other studies.

We then go on to consider adjustments to our model, varying the initial distributions and mass transfer prescriptions. In all respects other than those mentioned below, these models are identical to Model A.

In Models B, C and D, we use common envelope efficiency parameters α of 1.0, 2.0 and 4.0 respectively, while Model E uses the angular momentum formalism proposed by Nelemans et al.

(2000) for the first phase of spiral-in, with their recommended value of γ, and with $\alpha = 4.0$.

In models N, O, P and W, we also perturb the common envelope phase. In Model N, we include all of the envelope's ionisation energy (a positive quantity corresponding to the energy released when the ionised part of the envelope recombines) in its binding energy, meaning that envelopes will be less strongly bound and hence their removal will require less orbital shrinkage. This effect becomes important for stars on the AGB. Model O, on the other hand, does not include any of the ionisation energy.

Determinations of λ from stellar modelling are found to depend on the definition of the core-envelope boundary (Tauris & Dewi, 2001) in giant stars. Because of this uncertainty, we also evolve models W and P, in which we fix $\lambda = 0.5$, with $\alpha = 3$ and $\alpha = 4$, respectively.

In Model F, we choose the primary mass from the IMF of Scalo (1986), as in Schneider et al. (2001). Then in Model G we select both M_1 and M_2 independently from the KTG IMF, as suggested by Kroupa et al. (1993). We also evolve a Model K, in which initial orbital eccentricities are set to zero.

Models L and M alter the production of DDs via the RLOF+CEE route described in section 3.3. It has been suggested (Han, Tout & Eggleton, 2000) that Roche lobe overflow may be stable until later in the Hertzsprung gap (HG) than happens using the BSE code, so a Model with enhanced HG overflow was added (Model L). Model M has semiconservative overflow during this stage, to emphasise the uncertainties associated with HG mass transfer.

The Galactic DD population was simulated using each model in turn; the results of this exercise are summarised in Table 3.1. Imposing the criterion given above, we eliminate Models B, G and W based on their underproduction of DDs. If we increase the binary fraction to 100 per cent, this tends to underproduce single WDs, leading to an especially high DD fraction and a low overall WD space density. Note that the table also contains a Model H, which is in agreement with observations and is described in the next section.

Thus the models A, C, D, E, F, H, K, L, M, N, O and P progress to the next round, as representative of reasonable astrophysical uncertainties in our population synthesis calculations.

Model	% DD	$\rho_{\mathrm{WD},\odot}$ $(10^{-3}\mathrm{pc}^{-3})$	$\langle q \rangle_{\mathrm{vol}}$	Acceptable?
A	18	9	0.62	Yes
B	7	8	0.68	No
C	13	9	0.63	Yes
D	20	9	0.63	Yes
E	24	9	0.75	Yes
F	22	6	0.64	Yes
G	6	6	0.58	No
H	18	9	0.62	Yes
K	17	9	0.63	Yes
L	18	9	0.63	Yes
M	17	9	0.62	Yes
N	17	9	0.63	Yes
O	20	9	0.59	Yes
W	9	8	0.62	No
P	12	8	0.62	Yes

Table 3.1: Properties of Galactic DD models; details of models given in section 3.5.3. % DD is the percentage of field WDs in a magnitude-limited survey that will have a WD companion in an orbit with $P < 100$ d. $\rho_{\mathrm{WD},\odot}$ is the local space density of WDs (single and double). $\langle q \rangle_{\mathrm{vol}}$ is the volume-limited average detached DD mass ratio q, where $q \leq 1$ by definition.

Three further models are added later (section 3.6.4); these vary in their cosmic star formation and metallicity histories, and so cannot be tested against the Galactic DD population.

3.5.4 Interacting DDs and modifications made to the *BSE* code

Some modifications were made to the BSE code regarding the treatment of accreting DD systems. In this we mainly follow the recommendations made in the detailed population synthesis work of Nelemans et al. (2001a).

AM CVn stars are mass-transferring compact binaries in which the transfer is driven by gravitational radiation, and in which the accretor is a white dwarf and the donor is a Roche-lobe filling star, which could be another (less massive) white dwarf, or a helium star. For a review, see Nelemans et al. (2001a) and references therein. While not expected to be the dominant source of the Galactic gravitational wave background (Hils 1998; Hils & Bender 2000), some of these systems will be useful as "verification" sources for LISA, with large, predictable gravitational wave amplitudes.

We include in our definition of AM CVns all systems in which a helium star or WD is transferring mass on to a WD, including those systems in which the donor star is a CO or ONe WD.

The WD family When the donor star is a white dwarf, the orbital separation at initial Roche lobe overflow is around 0.1 $R_\odot$, which is often sufficiently small that the accretion stream impacts directly on the accretor's surface, so an accretion disc is not expected to form. This has implications for the orbital evolution of the mass-transferring binary. When an accretion disc is present, tidal torques on the outer edge of the disc return to the orbit the angular momentum carried away from the donor by the accretion stream. In the absence of such a mechanism for restoring the orbital angular momentum, the criterion for stable mass transfer becomes much more stringent, and in most cases an AM CVn star will not form, precluding the existence of the WD family. Here we take the optimistic view (as in model II of Nelemans et al. 2001a) that, even if no disc is present, some tidal mechanism has an equivalent effect and that all WD–WD systems for which the mass ratio is < 0.628 (Hurley et al., 2002) will commence stable mass transfer upon Roche lobe overflow. We modify the BSE code accordingly. This optimism is perhaps warranted, since we *do* see WD family AM CVn systems, e.g. Israel et al. (2002), which reports on the discovery of a helium-transferring compact binary with orbital period (321 s) too short to involve a (non-degenerate) helium star donor.

The helium star family In this case, the donor star is a helium star, produced when a star with mass $\gtrsim 2$ $M_\odot$ loses its envelope on the RGB. Since these stars can live for a rather long time compared with the main sequence lifetimes of their progenitors, there is a significant chance that through GW losses (or sometimes radial evolution) they will commence mass transfer before evolution to the WD stage. Here we shall employ the same condition on the dynamical stability of this mass transfer as Nelemans et al. (2001a): $q = M_{\mathrm{nHe}}/M_{\mathrm{wd}} < 1.2$ (we use "nHe" to denote (naked) helium star, to avoid confusion with helium-core WDs). Stellar modelling (Savonije, de Kool & van den Heuvel, 1986) indicates that rapid mass transfer forces the helium star out of thermal equilibrium, increasing the thermal timescale beyond a Hubble time. The star cannot ever regain thermal equilibrium, and becomes semi-degenerate (as opposed to fully degenerate) as its mass falls. This results in a negative exponent in the mass-radius relation, so that the orbital separation then increases as the helium star stably loses mass, i.e. an AM CVn system is formed. Note that at the onset of Roche lobe overflow, helium stars are always large enough that an accretion disk can form.

The standard BSE code does not incorporate the possibility of these semi-degenerate helium stars, so this was added. Here we adopt the same semi-degenerate mass-radius relation as in Nelemans et al. (2001a) (in solar units):

$$R_{\rm nHe} = 0.043\, M_{\rm nHe}^{-0.062}, \tag{3.13}$$

and switch between this and the regular non-degenerate relation by selecting the larger of the two radii when the helium star is transferring mass on to a WD companion. In our code, this changeover occurs at $M_{\rm nHe} \sim 0.29$ M$_\odot$. We also modify the mass transfer rate prescription in the code, in order that the transfer responds more quickly to the initial overflow, so that the helium star does not hugely overhang its Roche lobe, and we halt further helium burning, so that the star cannot evolve to the WD stage during transfer, due to its long thermal timescale. We note that this modification is fairly crude, but ought to give a good indication of the relative importance of helium star AM CVn systems as sources of the GW background.

A further issue in the formation of any helium-transferring system is that of edge-lit detonations (ELDs), which are believed to occur after a layer of helium has built up in the surface of an accreting CO WD. The BSE code detonates CO WDs in this way after the accretion of 0.15 M$_\odot$ of helium. We evolve separately a model (Model H) in which this is increased to 0.3 M$_\odot$, as in Model II of Nelemans et al. (2001a).

3.6 Cosmological equations

In this section we describe our calculation of the cosmological background. We adopt a standard lambda-cosmology, with $\Omega_{\rm m} = 0.3$, $\Omega_\Lambda = 0.7$ and $H_0 = 70$ km s^{-1} Mpc^{-1}. This means that the current age of the universe, $T_0 = 13.5$ Gyr. We assume isotropy throughout; for an analysis of the small anisotropy due to the localisation of binary stars in galaxies that follow the large scale structure of the universe, see Kosenko & Postnov (2000).

3.6.1 Basic equations

The specific flux $F_{f_\mathrm{r}} = dF(f_\mathrm{r})/df_\mathrm{r}$ received at frequency f_r from an object at redshift z with specific luminosity L_{f_e} is given by (e.g. Peacock, 1999)

$$F_{f_\mathrm{r}} = \frac{L_{f_\mathrm{e}}}{4\pi d_\mathrm{L}(z)^2}\left(\frac{df_\mathrm{e}}{df_\mathrm{r}}\right), \tag{3.14}$$

where $f_\mathrm{e} = (1+z)f_\mathrm{r}$, $d_\mathrm{L}(z) = (1+z)d_\mathrm{M}(z)$ is the luminosity distance to redshift z and d_M is the proper motion distance (cf. section 5 of Hogg (2000), which is also $1/(2\pi)$ times the proper ("comoving") circumference of the sphere about the source which passes through the earth today).

If the radiation comes from a large number of sources spread over redshift and isotropically distributed on the sky, we can write $dL_{f_\mathrm{e}}(z) = \ell_{f_\mathrm{e}}(z)dV(z)$, where $\ell_{f_\mathrm{e}}(z)$ is the comoving specific luminosity density (say in erg s^{-1} Hz^{-1} Mpc^{-3}), $dV(z) = 4\pi d_\mathrm{M}^2 d\chi$ is the comoving volume element and χ is the comoving distance.

We can then write the specific flux received in gravitational waves today as

$$\begin{aligned} F_{f_\mathrm{r}} &= \int_{z=0}^{\infty} \frac{\ell_{f_\mathrm{e}}}{4\pi d_\mathrm{L}^2(z)}\left(\frac{df_\mathrm{e}}{df_\mathrm{r}}\right) dV(z) & (3.15)\\ &= \int_{T=0}^{T_0} \frac{\ell_{f_\mathrm{e}}(T)}{(1+z(T))}\left(\frac{df_\mathrm{e}}{df_\mathrm{r}}\right) c\, dT, & (3.16)\end{aligned}$$

using $d\chi = -(1+z)c\,dT$, where T is cosmic time.

This is the basic equation on which the code is based. The equation is discretised in f_r, T and ℓ as described in section 3.6.2.

3.6.2 Computational equations

In the code, we bin the received gravitational waves in frequency. To calculate the flux received in a frequency bin with limits $f_{\rm r1}$ and $f_{\rm r2}$, we integrate Eq. 3.16 between these limits:

$$\begin{aligned} F_{f_{\rm r1}\to f_{\rm r2}} &= \int_{f_{\rm r1}}^{f_{\rm r2}} \int_{T=0}^{T_0} \frac{\ell_{f_e}(T)}{(1+z(T))} \left(\frac{df_e}{df_r}\right) cdT\, df_r \\ &= \int_{T=0}^{T_0} \int_{(1+z)f_{\rm r1}}^{(1+z)f_{\rm r2}} \frac{\ell_{f_e}(T)}{(1+z(T))}\, df_e\, cdT, \end{aligned} \tag{3.17}$$

i.e. we integrate only over those emitted frequencies that will have been redshifted to arrive in this frequency bin today. The bin size was chosen to be 0.1 in $\log_{10}(f_{\rm r})$.

Clearly, to calculate F, we need to know the comoving luminosity density ℓ_{f_e} in gravitational radiation at frequency $f_{\rm e}$ as a function of cosmic time.

We first obtain the source population at a given cosmic time T_i, by simply generalising Eq. 3.12, so that now

$$N_{k,j}(T_i) = (t_{j+1} - t_j)\mathcal{R}_c(T_i - t_j)\mathcal{P}_k, \tag{3.18}$$

where $\mathcal{R}_c(T)$ is the cosmic star formation rate at time T, expressed as a number of binary stars born per unit time per unit volume, and $N_{k,j}(T_i)$ is the number density of binaries with parameters k, j at cosmic time T_i, and where we require $T_i \geq t_j$.

The gravitational wave luminosity density at time T_i is then given by

$$\ell_{f_{\rm e}}(T_i) = \sum_{k,j} N_{k,j}(T_i) L_{k,j}(f_{\rm e}), \tag{3.19}$$

i.e. we simply sum over the emission at frequency $f_{\rm e}$ from all sources k, j present at that time, weighted by their space densities.

Since each binary source s emits radiation at only specific frequencies $f_n = n\nu_s$ (where ν_s is the orbital frequency of binary s) at a given time (Eq. 3.3), this sum can be expressed as

$$\ell_{f_{\rm e}}(T_i) = \sum_{k,j} N_{k,j}(T_i) \sum_n L_{\rm circ,k,j}\, g(n, e_{k,j}) \delta(f_{\rm e} - n\nu_{k,j}). \tag{3.20}$$

We then have

$$F_{f_{\rm r1} \rightarrow f_{\rm r2}} = \sum_i \sum_{k,j} \sum_{n_{\rm min}}^{n_{\rm max}} \frac{N_{k,j}(T_i) L_{\rm circ,k,j}\, g(n, e_{k,j})}{(1+z_i)}\, c\Delta T, \tag{3.21}$$

where we have also discretised the integral over cosmic time T, as a sum over i intervals ΔT, and where n is an integer, with the limits $n_{\rm min}$ and $n_{\rm max}$ defined by $f_{\rm r1} < \frac{n\nu_{j,k}}{1+z_i} < f_{\rm r2}$. At a given redshift $z_i(T_i)$, we just sum over those harmonics of those sources that will lead to emission at frequencies $f_{\rm e}$, with $f_{\rm r1}(1+z_i) < f_{\rm e} < f_{\rm r2}(1+z_i)$, and hence reception in the $f_{\rm r1} \rightarrow f_{\rm r2}$ frequency bin today.

The integration timestep ΔT must be sufficiently small that the emitting source population does not change significantly on timescales shorter than this, i.e. we assume a quasi-steady state population during this interval, so that our snapshot of the population at time T_i is representative of the whole timestep ΔT. A value of $\Delta T = T_0/50$ was used throughout. We checked that timesteps smaller than this did not yield noticeably different results. Individual sources may evolve significantly within this timestep, but the characteristic emission of the population will be unchanged. It should also be noted that the evolutionary timesteps taken for the binary stars are independent of this integration timestep (see section 3.5.1), so that ΔT may be made much larger than the timescales of the evolutionary processes of interest, so long as the population is roughly steady-state over ΔT.

Equation 3.21 is the sum performed by the code written for this paper, for a large number of received frequency bins over the range $10^{-6} < f_{\rm r} < 10^0$ Hz. For practical purposes, the sum over harmonics is truncated when $g(n, e)$ drops below 10^{-3}, well beyond the peak in the emitted spectrum at $n_p = 1.63(1-e)^{-3/2}$. For typical $e < 0.95$, our numerical cutoff at $g = 10^{-3}$ corresponds roughly to including only $n < 5n_p$. The higher values $n > 5n_p$ contribute less than 1 per cent of the total gravitational wave luminosity.

3.6.3 Quantities used

Some quantities commonly used in gravitational wave astronomy are $F_{f_\mathrm{r}}(f_\mathrm{r})$, 4π times the specific intensity; $\Omega_\mathrm{gw}(f_\mathrm{r})$, the fraction of closure density per logarithmic GW frequency interval; and the power spectral density $S_h(f_\mathrm{r})$.

The first of these, $F_{f_\mathrm{r}}(f_\mathrm{r})$, can be calculated from

$$F_{f_\mathrm{r}}(\overline{f_{12}}) = \frac{F_{f_{\mathrm{r}1}\rightarrow f_{\mathrm{r}2}}}{(f_{\mathrm{r}2} - f_{\mathrm{r}1})}. \tag{3.22}$$

The second, $\Omega_\mathrm{gw}(f_\mathrm{r})$, is the fraction of closure energy density contained in gravitational waves received in the logarithmic frequency interval around f_r, i.e.

$$\Omega_\mathrm{gw}(f_\mathrm{r}) = \frac{1}{\rho_c c^2}\frac{f_\mathrm{r} F_f(f_\mathrm{r})}{c}, \tag{3.23}$$

where ρ_c is the critical mass density of the Universe; $\rho_c = (3H_0^2/8\pi G) \simeq (1.88 \times 10^{-29})\, h_{100}^2$ g cm^{-3}, where $H_0 = 100\, h_{100}$ km s^{-1} Mpc^{-1}. In terms of computational quantities,

$$\begin{aligned}\Omega_\mathrm{gw}(\overline{f_{12}}) &= \frac{1}{c^3\rho_c}\frac{F_{f_{\mathrm{r}1}\rightarrow f_{\mathrm{r}2}}}{\Delta(\ln f_\mathrm{r})} \\ &\simeq 0.0175\, F_{f_{\mathrm{r}1}\rightarrow f_{\mathrm{r}2}} \end{aligned} \tag{3.24}$$

(where $F_{f_{\mathrm{r}1}\rightarrow f_{\mathrm{r}2}}$ is in erg s^{-1} cm^{-2}) since $\Delta(\ln f_\mathrm{r}) = \ln(10^{0.1}) \simeq 0.23$ and $h_{100} = 0.7$.

The power spectral density $S_h(f_\mathrm{r})$ is given by

$$S_h(f_\mathrm{r}) = \frac{4G}{\pi c^3}\frac{1}{f_\mathrm{r}^2} F_{f_\mathrm{r}}(f_\mathrm{r}). \tag{3.25}$$

Usually this is plotted as $S_h^{1/2} \simeq (5.6 \times 10^{-20})\frac{F_{f_\mathrm{r}}^{1/2}}{f_\mathrm{r}}\mathrm{Hz}^{-1/2}$, where F_{f_r} is in erg s^{-1} cm^{-2} Hz^{-1}, and f_r is in Hz.

3.6.4 Cosmic star formation history

As pointed out by Schneider et al. (2001), most determinations of cosmic star formation history are based on the UV emission from massive stars (e.g. Madau et al. 1996; Steidel et al. 1999), and use an assumed single-star IMF (commonly that of Salpeter 1955) to convert observed UV flux into a star formation rate as a function of redshift. This type of rate is inconvenient here for two reasons: first, a non-trivial factor is required for conversion to a *binary* star formation rate (for an assumed binary fraction), because of the need to correct for the observed flux from companion stars; and second, the total star formation rate is pivoted on the high-mass end of the stellar distribution, while here we are interested in studying the remnants of low- to intermediate-mass stars. This results in a crucial dependence on the choice of stellar IMF.

Schneider et al. (2001) overcome the first problem by assuming the measured shape of the cosmic SFH as a function of time, but normalising its amplitude to the local rate of core-collapse supernovae. This Type Ibc/II SN rate is a more easily calculated quantity for a given (binary or single) IMF than is the UV luminosity density. Since Schneider et al. (2001) are also concerned with neutron stars in their study, this is a reasonable choice. However, the second problem remains when one is concerned with WDs; and in addition, not only does the normalisation pivot on the high mass stars, but it also depends crucially on the ratio of local to peak cosmic SFR. We also note that the minimum mass of star producing a core-collapse supernova explosion is uncertain (e.g. Jeffries, 1997).

For our normalisation, we use instead the observed local stellar mass density Ω_*, as derived from the local near-IR luminosity function by Cole et al. (2001). This quantity is most sensitive to stellar masses near the MS turnoff in old populations, $M \sim 0.8 - 1.0$ M$_\odot$, and thus is more closely related than the SNIbc/II rate to the DD progenitor population. We convert between their assumed single star IMF and our binary star IMFs by keeping constant the mass in stars in this range. We then use the recycled fraction $R = 0.42$, as for the Kennicutt (1983) IMF used in Cole et al. (2001), to convert stellar density today to total mass of stars ever formed, $\Omega_{*,\rm tot}$ (the time-integral of the cosmic star formation rate). Doing this, we obtain $\Omega_{*,\rm tot} = 5.0 \times 10^8\,M_\odot$Mpc^{-3} for the KTG IMF, while for the Scalo IMF this figure is $\Omega_{*,\rm tot} = 4.0 \times 10^8\,M_\odot$Mpc^{-3}. Due to this rather crude conversion, the

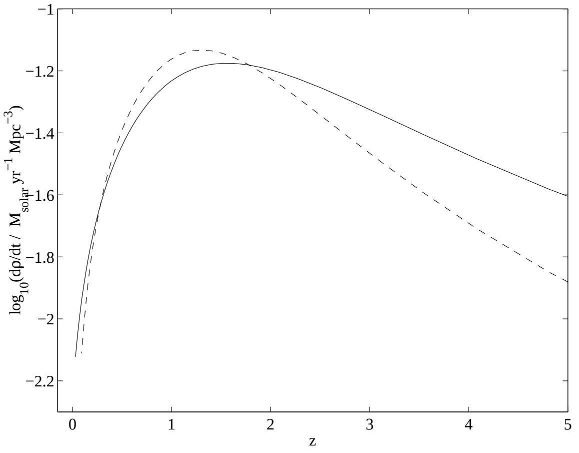

Figure 3.3: Two possible cosmic star formation histories, plotted as a function of redshift z and parametrized according to the smooth curve fits given in Cole et al. (2001). Dashed line: no extinction correction made, used in Model J. Solid line: extinction corrected with $E(B-V) = 0.15$, used in all other models. The time integral of each rate is fixed using the appropriate $\Omega_{*,\mathrm{tot}}$ derived from Cole et al. (2001). Curves shown are for the KTG IMF.

uncertainty in these figures will be greater than the 15 per cent quoted by Cole et al. (2001) for Ω_*; we estimate the resulting uncertainty to be $\sim$ 30 per cent.

Cole et al. (2001) note that their calculated stellar densities are most consistent with UV-derived star formation rates if the extinction corrections used in these methods are moderate. However, we would like to assess the effects of uncertainty in the shape of the cosmic star formation history. We therefore select both a history with large extinction corrections and one with none, keeping the integral over time fixed to $\Omega_{*,\mathrm{tot}}$ for each. The corresponding curves are plotted in Fig. 3.3. We use the extinction-corrected rate, favoured by Steidel et al. (1999), in Model A and all other models except for Model J, which uses the uncorrected rate (but is identical to Model A in all other respects). We also introduce Models Q and R, whose metallicity histories differ from that of Model A: in Model Q, stars born during the first Gyr have metallicity 1/20 solar, while stars born later have solar composition; in Model R, all stars have metallicity $Z = 0.01$, i.e. half-solar.

3.7 Basic results

The GW background spectrum received in the frequency range $10^{-6} < f_{\rm r} < 10^{-1}$ Hz, generated using our fiducial Model A, is plotted in Fig. 3.4. The total amplitude is broken down into separate contributions from four main evolutionary stages: main sequence–main sequence (MS–MS), WD–MS, WD–WD and WD–helium star (WD–nHe) binaries, and plotted in terms of each of $F_{f_{\rm r}}$, $\Omega_{\rm gw}$ and $S_h^{1/2}$ described in section 3.6.3. The unitless $\Omega_{\rm gw}$ will be our preferred quantity for the remainder of the paper.[2].

The four component spectra are plotted in Fig. 3.5 for all of the models evolved, to illustrate that the spectral shapes are largely unaffected by any of the changes made. A summary of important quantities for each model (to be discussed later) is given in Table 3.2. For reference, we also list a Model A′, identical in parameters to Model A, to demonstrate the typical level of statistical variation in the results. This is clearly at the 1 per cent level in flux, so that variations larger than this between models can be ascribed to parameter, and not statistical, variations.

Throughout we will focus on the properties of the spectrum around 1 mHz, in the centre of the LISA band and of the spiral-in regime. We will also compare with the spectral properties at 10 mHz, at which frequency lower-mass WD–WD pairs can no longer be present and at which point this extragalactic WD–WD background will be the dominant LISA background source (see Fig. 3.16).

It is clear that the signal in the LISA frequency band ($0.1 \lesssim f_{\rm r} \lesssim 10$ mHz) is dominated by the WD–WD component, as expected. Neither the MS–MS nor the MS–WD binaries can radiate at frequencies above the bottom of this band pass, since even the lowest mass MS stars come into contact at frequencies below 1 mHz. WD–nHe pairs can contribute to a somewhat higher frequency due to the smaller radii of helium stars, but still come into Roche lobe contact at $f_{\rm e} \sim 1$ mHz.

The WD–WD component clearly displays the spectral shape predicted in Section 3.4 ($\Omega_{\rm gw} \propto f_{\rm r} F_{f_{\rm r}}$, plotted in Figs. 3.1 and 3.2), with a clear separation between static and spiral-in regimes at around 10^{-4} Hz. The slope in the static regime suggests that sources are injected with a spectrum closer to $\dot{N}_{\rm b}(\nu) \propto \nu^{-2/3}$ than to ν^{-1}, but agreement to this level is encouraging. The spiral-in slope

[2] Note that since Cole et al. (2001) quote $\Omega_{*} h$ in their paper, and we use this quantity to normalise our star formation rate, our calculated $\Omega_{\rm gw}$ also scales as h^{-1}. We use $h = 0.7$.)

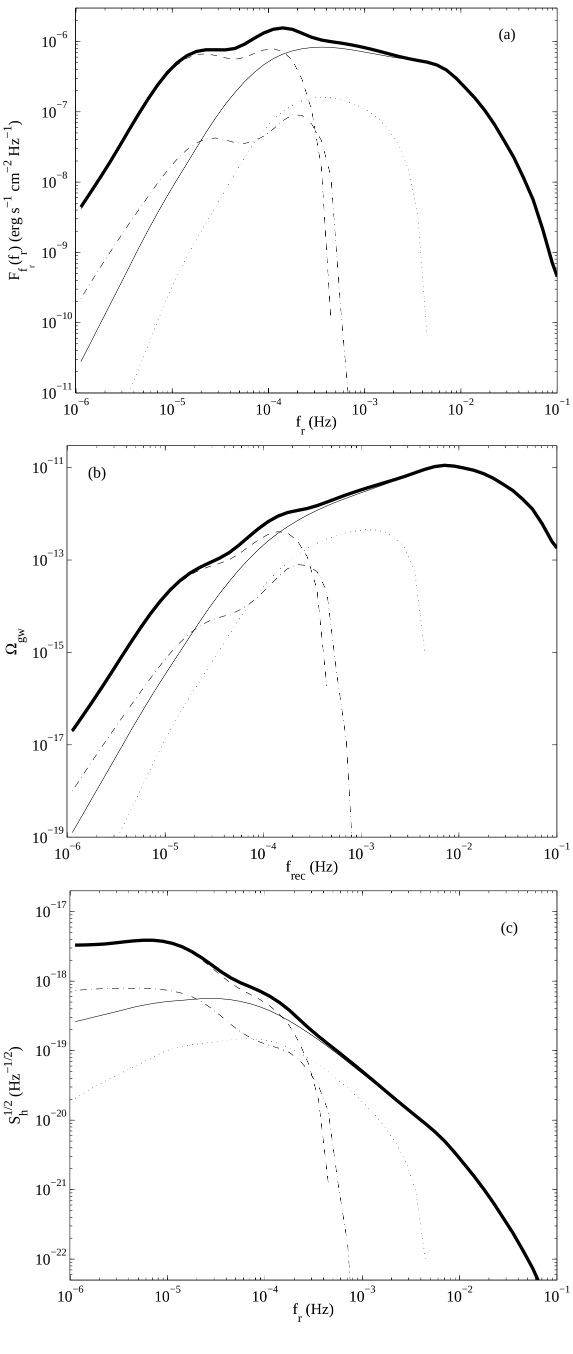

Figure 3.4: The GW background for our fiducial Model A, in terms of the three quantities described in section 3.6.3. Solid line: WD–WD pairs; dotted line: nHe–WD pairs; dashed line: MS–MS binaries, and dot-dash line: WD–MS binaries. The total signal (the sum of the four parts) is given by the thick solid line. Only $n = 2$ harmonics of the orbital frequency are plotted (see section 3.7.1.1).

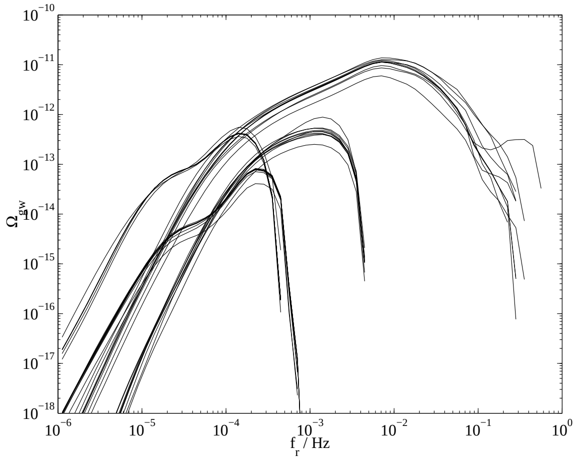

Figure 3.5: Comparison of spectral shapes for all Models. Curves are the same as in Fig. 3.4, but all quantities are plotted as solid lines. Only $n = 2$ harmonics of the orbital frequency are plotted (see section 3.7.1.1).

Model	$\Omega_{\rm gw}$(1 mHz)			$\langle\mathcal{M}\rangle$	%	N_0
	Total	R+C	C+C		AM CVn	
A	3.57	1.35	2.22	0.45	13	1.17
A′	3.61	1.36	2.26	0.45	14	1.18
C	3.06	0.60	2.47	0.44	16	0.90
D	3.66	1.64	2.02	0.47	13	1.20
E	4.21	1.35	2.86	0.47	10	1.57
F	1.94	0.72	1.22	0.41	13	0.75
H	4.10	1.53	2.58	0.43	25	1.17
J	3.62	1.38	2.24	0.45	13	1.17
K	4.29	2.09	2.20	0.48	13	1.29
L	3.80	1.53	2.27	0.45	12	1.25
M	2.80	0.66	2.14	0.44	15	0.92
N	3.43	1.36	2.07	0.46	13	1.13
O	3.89	1.31	2.57	0.46	13	1.27
P	5.46	1.00	4.46	0.55	16	1.20
Q	3.73	1.43	2.30	0.44	13	1.32
R	3.83	1.48	2.35	0.44	12	1.28

Table 3.2: Summary table of results from all models described in the text. $\Omega_{\rm gw}$(1 mHz) is in units of 10^{-12}. R+C refers to the RLOF+CEE route to the DD stage, while C+C refers to the CEE+CEE route (see section 3.3; note that for Model E, we hold the RLOF+CEE contribution fixed from Model A). The flux-weighted mean chirp mass $\langle\mathcal{M}\rangle$ contributing at $f_{\rm r} \sim 1$ mHz is in units of $M_\odot$, the next column lists the percentage contribution to $\Omega_{\rm gw}$ at 1 mHz from interacting binaries and the last column gives the inspiral remnant density N_0 today, in units of 10^6 Mpc^{-3}. Models B and G were rejected for reasons noted in Table 3.1.

Pairing	% over all time	% locally
He–He	12.4	29.5
He–CO	23.0	25.3
He–ONe	0.6	0.6
CO–CO	42.2	33.2
CO–ONe	8.1	4.4
ONe–ONe	1.0	0.2
(of which AM CVn)	3.6	4.7
nHe–WD	12.7	6.9
(of which AM CVn)	9.7	2.0
Total	100	100
(of which AM CVn)	13.3	6.7

Table 3.3: Percentage contribution to $\Omega_{\rm gw}$ at 1 mHz from different DD pairs, for both contribution to total integrated background and contribution to background coming from the local universe, $z = 0$. All for fiducial Model A. MS–MS and WD–MS binaries contribute negligibly at this frequency. ("nHe" denotes a naked nondegenerate helium star.) All contributions, including the AM CVn values, are given as fractions of the total flux at 1 mHz.

is slightly steeper than predicted, but this is due to the spectrum seen being the sum of spectra from populations with different chirp masses, as well as different merger and maximum injection frequencies (see Fig. 3.8), whose individual slopes in the spiral-in regime are closer to the predicted 2/3. Agreement with our simple predictions is therefore good and we feel that we understand well the origins of the spectrum.

3.7.1 Contributors

The breakdown of contributions to the background received at 1 mHz for our fiducial Model A is given in Table 3.3. In this section we identify the dominant source types, and those types whose contribution is negligible, then attempt to characterise the emitting population in terms of a mean chirp mass and inspiral remnant density.

3.7.1.1 Eccentric harmonics

As described in section 3.2, systems with eccentric orbits emit gravitational waves at all harmonics $n\nu$ of the orbital frequency, not just the $n = 2$ harmonic as for circular orbits.

The only *close* binaries we expect to be eccentric are unevolved MS–MS binaries in which tidal forces have not yet circularised the orbit. Almost every close evolved (e.g. WD–MS, WD–WD)

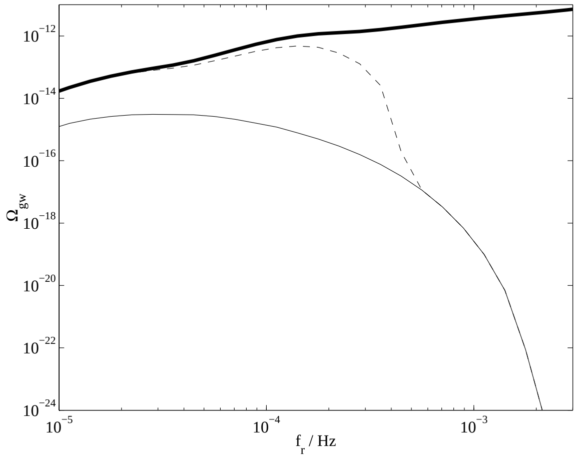

Figure 3.6: The GW background from harmonics with $n \geq 3$ from MS–MS pairs (thin solid line), plotted along with the total MS–MS pair contribution (dashed line), and the total background from all sources (thick solid line), demonstrating that the harmonic contribution is negligible. All for Model A.

system will have at some point experienced a Roche lobe–filling phase, which will likely have circularised the system, through tidal circularisation and/or common envelope evolution. Figure 3.6 shows the contribution from harmonics with $n \geq 3$ to the MS–MS GW spectrum for Model A (which has a thermal initial eccentricity distribution). Clearly the $n \geq 3$ harmonics contribute $\lesssim 10$ per cent of the MS–MS spectrum at frequencies $f_{\mathrm{r}} \lesssim 0.5$ mHz, and although they dominate the MS–MS spectrum above this frequency, these signals are buried deep below the other contributors at $f_{\mathrm{r}} > 0.5$ mHz (see Fig. 3.4). Hereafter we safely neglect the $n \neq 2$ contributions to L_{gw}, in the interests of computing time, though we do not neglect eccentric orbits in computing stellar evolution sequences.

3.7.1.2 Interacting binaries

Interacting binaries (those in which either a WD or non-degenerate naked helium (nHe) star is transferring mass on to a WD) contribute 13 per cent of the GW background at 1 mHz in Model A. Since at this frequency the majority of nHe star companions fill their Roche lobes, most of the nHe–WD background comes from interacting systems. At 10 mHz, 26 per cent of the GW signal comes from interacting binaries, all of these necessarily WD-donor systems. The GW spectrum due to interacting binaries is compared with the total signal in Fig. 3.7.

The percentage contribution from interacting systems is fairly constant across models, except for Model H, in which an accreting CO WD is permitted to accumulate 0.3 $M_\odot$ of helium before detonation, as opposed to the 0.15 $M_\odot$ in our fiducial model. This increase in survival rate boosts the interacting binary signal at 1 mHz by a factor of 2. For the other models, the interacting WD–WD signal is boosted when the WD–WD pairs formed typically have larger mass ratios, so that more systems can commence stable transfer upon Roche contact, e.g. Model C.

The spectral shapes from interacting systems are governed by the mass-radius relation of the Roche lobe–filling star, and so do not share the spectral slopes displayed by the detached binaries. The overall contribution from interacting pairs is sufficiently small, however, that the total spectral shape is little affected by their presence. This is in line with results for the Galaxy found by Hils & Bender (2000) and Nelemans et al. (2001c).

We can predict the spectral shape due to interacting WD–WD binaries using some simple scaling relations (in the notation of section 3.4): for a Roche lobe–filling WD of mass M_d, we have $M_\mathrm{d}^{-1/3} \propto R_\mathrm{d} = R_L \propto a M_\mathrm{d}^{1/3} \propto M_\mathrm{d}^{1/3} f^{-2/3}$, using Kepler's law (for conservative mass transfer). If we then assume that the mass of the donor WD is much less than that of the accretor, then the system chirp mass $\mathcal{M} \propto M_\mathrm{d}^{3/5}$, so that $f \propto \mathcal{M}^{3/5}$ and the system gravitational wave luminosity $L_\mathrm{gw} \propto f^{10/3} \mathcal{M}^{10/3} \propto f^{16/3}$.

For sources sweeping (backwards) through frequency space, we have $N(f) \propto 1/\dot{f} \propto \mathcal{M}^{-5/3} f^{-11/3} \propto f^{-14/3}$.

Putting these together, we then have, for the emitted flux in the logarithmic frequency interval around f, $\Omega_\mathrm{gw}(f) \propto fF(f) = fL_\mathrm{gw}N(f) \propto f^{5/3}$. From Fig. 3.7, we measure the spectral slope between 0.4 and 6 mHz to be $\sim$1.7, in good agreement with this calculation. Interacting WD–WD sources are not present below this frequency range because evolution to these frequencies requires more than a Hubble time. Above $\sim$ 6 mHz, the spectral shape depends on the fraction of sources of high enough mass to radiate at a given frequency; this number drops rapidly with increasing frequency. Note that, within the 0.4–6 mHz range, since the spectrum $\Omega_\mathrm{gw} \propto f^{5/3}$ for interacting

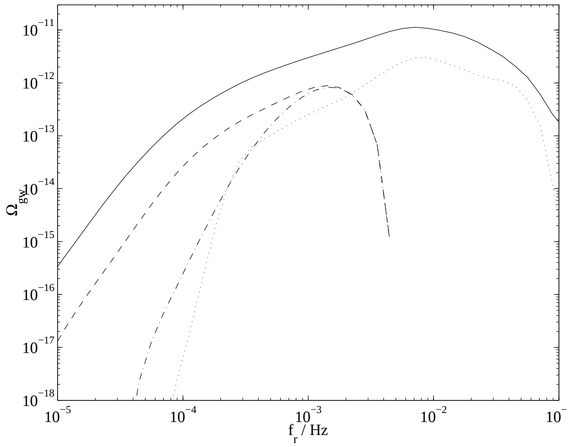

Figure 3.7: The spectrum due to interacting binaries, for our fiducial Model A. The solid line shows the total WD–WD binary contribution, while the dotted line gives the spectrum from interacting binary WD–WD systems. The dashed line is the total nHe–WD spectrum, of which the dash-dot line gives the interacting nHe–WD contribution.

WD–WD binaries rises relative to $\Omega_{\rm gw} \propto f^{2/3}$ for inspiralling detached binaries, interacting binaries are more important contributors at high frequencies than at low.

3.7.1.3 WD types, chirp mass, and merger rates

The dominant component of the background at frequencies 0.1–10 mHz comes from the inspiral of WD–WD systems. From Table 3.3, we see that approximately half of this background comes from CO–CO pairs, descended primarily from higher mass progenitors than the majority of He–He systems. The dominance of these systems is a result of both the shorter time delay between star formation and DD birth for more massive MS stars, and the larger chirp masses for CO–CO systems, since the flux in the inspiral part of the spectrum scales as $\Omega_{\rm gw} \propto f_{\rm r}^{2/3}\mathcal{M}^{5/3}$ (see section 3.4). These two factors outweigh the fact that, from the IMF, many more potential progenitors of He WDs are born than those that always produce CO or ONe WDs after envelope loss.

Figure 3.9 shows however that, as more low-mass MS stars evolve to the DD stage, the relative contribution to the GW luminosity density from pairs involving He WDs is rising, and will eventually dominate. The percentage contribution to the local ($z = 0$) WD–WD GW emission at 1 mHz from pairs including at least one He WD is 55 per cent, whereas their contribution to the integrated

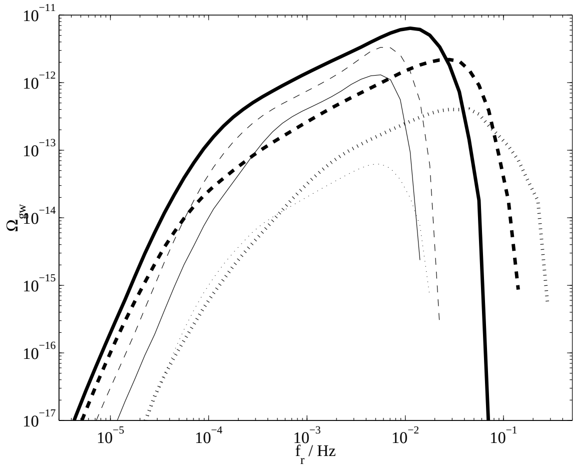

Figure 3.8: The background received from different WD–WD pairings, for Model A. From top to bottom at 1 mHz: thick solid line: CO–CO, thin dashed line: He–CO, thin solid line: He–He, thick dashed line: CO–ONe, thick dotted line: ONe–ONe, thin dotted line: He–ONe.

cosmological background received today is only 36 per cent.

A useful way to look at this is through the chirp mass distribution. Shown in Fig. 3.10 is the contribution to Ω_{gw} at 1 mHz as a function of system chirp mass (defined in section 3.4) for Model A, giving a flux-weighted mean chirp mass of 0.45 $\text{M}_{\odot}$. As increasingly lower mass systems evolve off the main sequence and become close DD pairs, this mean chirp mass is decreasing with time, as shown in Fig. 3.12. The chirp mass distribution depends on GW frequency (Fig. 3.11), most notably shifting towards higher masses at frequencies above which lower-mass WD–WD pairs will have merged. The mean chirp mass is somewhat higher below the critical spiral-in frequency, since for $f_{\text{e}} < 2\nu_{\text{crit}}$, we have $\Omega_{\text{gw}} \propto \mathcal{M}^{10/3}$, and above $2\nu_{\text{crit}}$, $\Omega_{\text{gw}} \propto \mathcal{M}^{5/3}$ (see section 3.4).

Phinney (2002) derived a simple expression for the GW background in terms of the chirp mass $\mathcal{M}$, assumed constant across all sources, and the current space density N_0 of remnant spiralled-in sources (with a weak dependence on cosmology and star formation history). We can assess the usefulness of this formula as a predictor of the background flux by using the results of our population synthesis calculations to see whether the computed fluxes can indeed be described by these two parameters only.

To calculate the remnant density N_0, we first calculate the source spiral-in rate as a function of

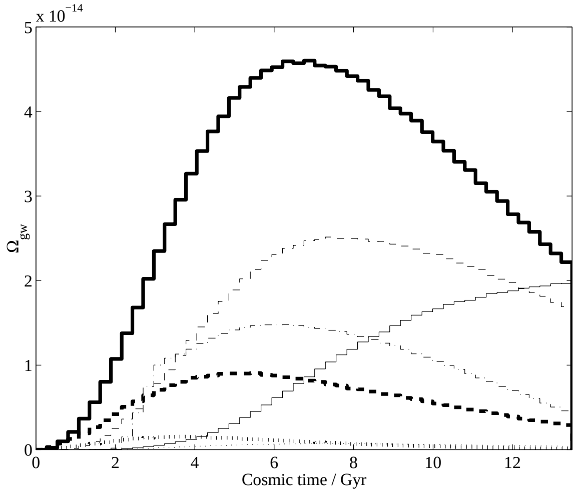

Figure 3.9: The contribution to $\Omega_{\rm gw}(1\ {\rm mHz})$ received today as emitted from each shell of cosmic time, $\Delta T = T_0/50$, from each source type, for Model A. Linestyles are as in Fig. 3.8, with the addition of the thin dash-dot line for nHe–WD pairs.

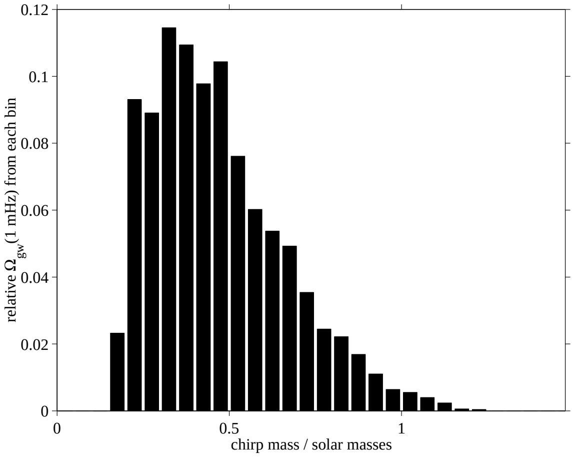

Figure 3.10: Relative contribution to $\Omega_{\rm gw}$ at 1 mHz as a function of chirp mass, for Model A, giving a mean flux-weighted chirp mass $\langle \mathcal{M} \rangle = 0.45\ {\rm M}_\odot$.

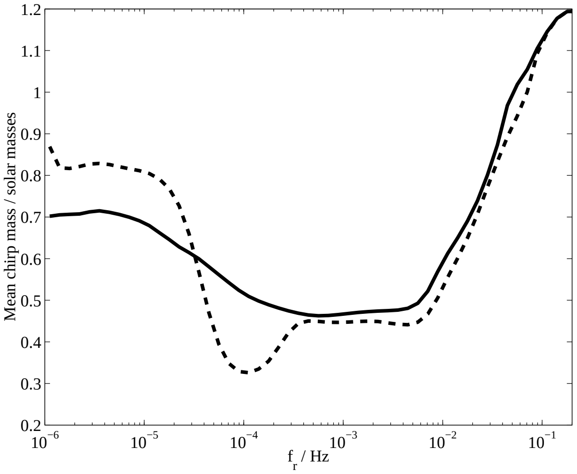

Figure 3.11: Flux-weighted chirp mass contributing to the GW background received, as a function of frequency, for Model A. Solid line: detached WD–WD pairs only, dashed line: all source types. The dip seen in this curve around 0.1 mHz is due to low-mass main sequence stars.

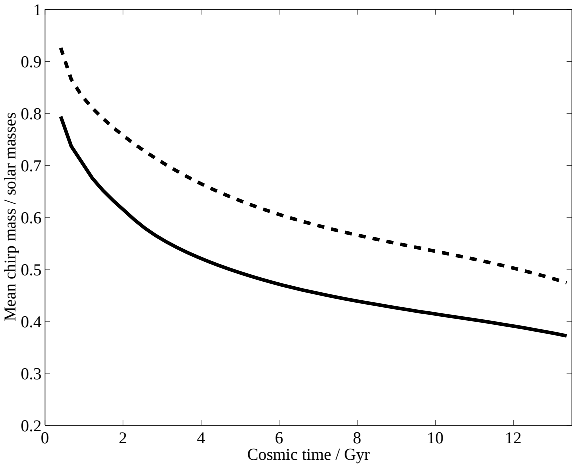

Figure 3.12: The flux-weighted mean chirp mass contributing to emission received today at 1–3 mHz (solid line) and 3–10 mHz (dashed line), from each shell of cosmic time. All for Model A.

cosmic time. The rate of occurrence of Roche lobe contact between WD–WD pairs (we shall call this the spiral-in rate) is different from the rate of WD–WD mergers, since for some subset of systems (those with mass ratios $q < 0.628$) stable mass transfer will commence upon overflow, and an AM CVn system will form. We keep track of both of these rates here.

For f greater than both $2\nu_{\rm max}$ and $2\nu_{\rm crit}$, i.e. in the part of the spiral-in regime above which sources are born (see section 3.4), then for a quasi-constant spiral-in rate $\dot{N}$ over the timestep $T_0/50$, the continuity equation (Eq. 3.6) simplifies to

$$\dot{N} = \sum_i \dot{\nu} N_i, \tag{3.26}$$

summed over all sources i at any given frequency satisfying the above requirement. For each source, $\dot{\nu}$ is given by Eq. 3.7. We perform this sum at each step in cosmic time, using systems with orbital frequencies in the range $0.8 < \nu < 1.6$ mHz, which is above the maximum injection frequency for the majority of sources, and below those frequencies at which the lowest mass WDs are coming into contact. We note that the inspiral time from $\nu \sim 0.5$ mHz is less than $T_0/50 = 0.27$ Gyr for all $\mathcal{M} \gtrsim 0.05$ M$_\odot$, so that at each timestep we are accurately representing the spiral-in rate at that time. The only exceptions are very low chirp mass systems, which we neglect here anyway, since these will be interacting binaries, which are spiralling *out*. We also neglect all nHe–WD pairs, since the evolution of these systems is not governed exclusively by gravitational radiation, but also via radial evolution of the nHe star, and also because Roche lobe contact occurs for these systems within our frequency range.

The spiral-in and merger rates obtained from Model A are plotted in Fig. 3.13. The present-day remnant density N_0 needed for the formula of Phinney (2002) is the time integral of the spiral-in rate, since this gives the total number of sources that have contributed to the background. From our calculated rate, we obtain $N_0 = 1.17 \times 10^6$ Mpc^{-3}.

Phinney (2002) deals only with the GW emission from non-interacting WD–WD systems, and so we should compare its predictions with only the non-interacting component of our computed signals, in addition to using a characteristic chirp mass $\mathcal{M}'$ for just those systems. For Model A, our flux-weighted mean chirp mass for detached WD–WD pairs is $\langle\mathcal{M}'\rangle = 0.47\ \mathrm{M}_\odot$ at 1 mHz. Eq. 16 of Phinney (2002), converting to $h_{100} = 0.7$, and omitting the $\left\langle(1+z)^{-1/3}\right\rangle$ scaling factor in the interests of simplicity, becomes

$$\Omega_{\mathrm{gw}} = 1.1 \times 10^{-17} \left(\frac{\mathcal{M}'}{\mathrm{M}_\odot}\right)^{5/3} \left(\frac{N_0}{\mathrm{Mpc}^{-3}}\right) \left(\frac{f_r}{1\ \mathrm{mHz}}\right)^{2/3}. \tag{3.27}$$

Using $\langle\mathcal{M}'\rangle$ and N_0 for Model A in the above, we find $\Omega_{\mathrm{gw}}(1\ \mathrm{mHz}) = 3.7 \times 10^{-12}$. We compare this with the computed value for detached WD–WD pairs, $\Omega_{\mathrm{gw}}(1\ \mathrm{mHz}) = 3.0 \times 10^{-12}$, and note that these agree to within 25 per cent. If we perform this same calculation for the other models, we find that Eq. 3.27 overestimates the computed background by a similar fraction.

The *variation* between models is thus well-fitted by the formula. The relative fluxes are reproduced by Eq. 3.27 to within 5 per cent for all models except D and E, whose fluxes relative to Model A are overestimated by 7 and 16 per cent respectively. The dominant scaling is due to variations in N_0, since in most cases $\langle\mathcal{M}'\rangle$ varies little between models. For the cases in which $\langle\mathcal{M}'\rangle$ does significantly change (D, E, F, K and P), the omission of the chirp mass scaling in Eq. 3.27 can improve (D, E) or worsen (F, K, P) the agreement with the results of our detailed calculations. This is perhaps as expected, since our flux-weighted chirp mass is in fact not the same average as that required in the generalisation of Phinney (2002) to accommodate a range of chirp masses. Such a value would also incorporate the redshift-scaling omitted in the above. We note, however, that neither N_0 nor either definition of $\mathcal{M}'$ is a directly observable quantity, requiring as they do integrations over cosmic time, and so are not easily determined from observations.

The computed spectral shape is not precisely $\Omega_{\mathrm{gw}} \propto f_{\mathrm{r}}^{2/3}$ (see Fig. 3.4), so we do not expect an exact reproduction of the spectrum using this formula. However, we conclude that with a knowledge of N_0 and $\mathcal{M}'$, we can quickly predict the detached WD–WD background amplitude and to some extent its variation if these values change. We note however that a full population synthesis calculation enables the inclusion of interacting systems, as well as the extraction of detailed spectral

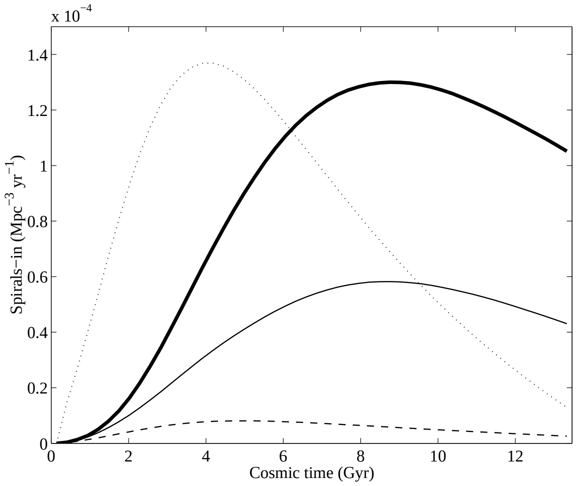

Figure 3.13: The rate of WD–WD spiral-in as a function of cosmic time. The thick solid line gives the total spiral-in rate, while the thin solid line shows the merger rate, that is the inspiralling sources that will merge, and not commence stable mass transfer (i.e. become AM CVn binaries) upon Roche lobe overflow. The thin dashed line gives the rate of merger of WD–WD pairs with combined mass > 1.4 M$_\odot$. For reference, the cosmic star formation rate, multiplied by $1/(1000\mathrm{M}_\odot)$, is plotted as the thin dotted line. All for Model A.

shapes and source property distributions, which are not available in a quick “manual” calculation.

3.7.2 Progenitors

Here we outline the relative contributions from the two main pathways to the DD stage, and we assess the impact upon each of these routes of varying the population synthesis model.

Figure 3.14 shows the contribution to $\Omega_{\rm gw}$ at 1 mHz as a function of the initial mass of the primary, for Model A. The descendants of primaries with ZAMS masses in the range 2–4 M$_\odot$ contribute 50 per cent of the signal, the flux-weighted mean progenitor primary mass being 3.7 M$_\odot$. Most of the sources in this range are the progenitors of CO WDs, since for $M \gtrsim 2$ M$_\odot$, a CO WD will be produced via a helium star upon envelope loss on the RGB, and a CO WD will be produced directly if the envelope is lost on the AGB. At 10 mHz, the mean progenitor mass rises to 4.7 M$_\odot$, since the (necessarily more massive) WD–WD pairs contributing there are descended from only the more massive ZAMS systems. The equivalent secondary mass distribution is not plotted here, but is always peaked towards initial mass ratios of unity.

Of perhaps more interest is the distribution in initial orbital semimajor axis (Fig. 3.15, for Model

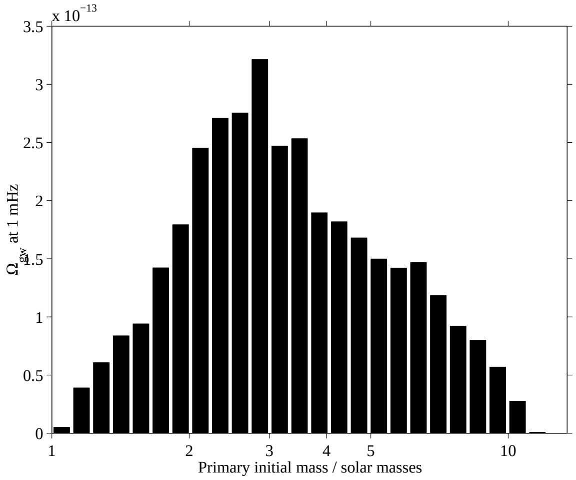

Figure 3.14: Contribution to $\Omega_{\rm gw}(1\ {\rm mHz})$ received, as a function of ZAMS mass of the primary, for our fiducial Model A.

A), which has a clear bimodal form, the peak at $a \sim 5a_{\rm min}$ corresponding to DDs which formed via RLOF+CEE, and the peak at $a \sim 50a_{\rm min}$ corresponding to the CEE+CEE route. We can therefore approximately determine the relative contributions from these two routes by dividing this distribution between the two peaks (at $a \sim 10a_{\rm min}$ for most models); the result of this division for each model is shown in Table 3.2. We note that the location of the CEE+CEE peak at $a \sim 50a_{\rm min}$ and the typical masses of the dominant progenitor stars mean that for this route, the dominant pathway involves primary overflow on the AGB, followed by secondary overflow on its RGB.

For Model A, $\Omega_{\rm gw}(1\ {\rm mHz}) = 1.4 \times 10^{-12}$ (~ 38 per cent of the total) comes from sources that evolved via the RLOF+CEE pathway. Since the WD–WD pairs from this route are generally more massive than CEE+CEE pairs, the percentage contribution at 10 mHz from this route rises to 44 per cent.

In general, we shall find that it is the RLOF+CEE contribution that is affected more by varying the population synthesis model. Although it can be affected significantly by varying the *form* of the common envelope prescription (models E and P), the CEE+CEE signal is quite robust to changes in the common envelope efficiency, since if systems originating at one separation happen to coalesce in a common envelope phase, using a given model, there exists a shell of sources at greater a to take their place as the closest WD–WD systems at birth, out to a maximum of $a \sim 10^3a_{\rm min}$ at which

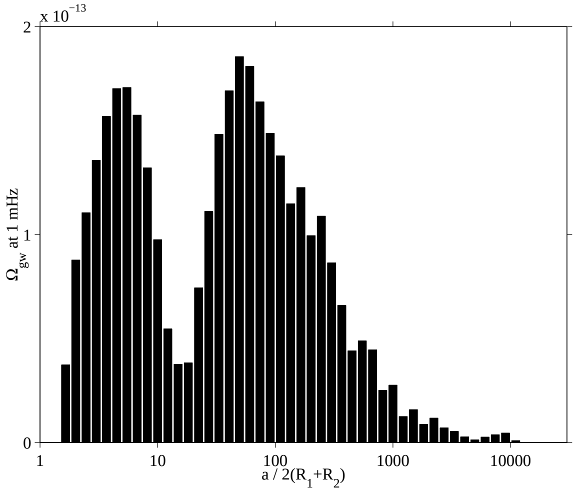

Figure 3.15: Contribution to $\Omega_{\rm gw}$(1 mHz) received, as a function of initial progenitor semimajor axis, expressed as a ratio of the initial semimajor axis to the minimum separation permitted in the code. For Model A.

Roche lobe overflow no longer occurs on the RGB or AGB. Webbink & Han (1998) describe this effect in terms of shifting the "window" in initial parameter space from which the closest DD systems are descended. The weak dependence of results upon the common envelope efficiency parameter is also seen in population synthesis calculations for other types of binary, e.g. Kalogera & Webbink (1998) for LMXBs.

Returning to the DD case, the RLOF+CEE pathway has no similar resource, occurring only in the rather narrow range of initial separation in which RLOF commences in the Hertzsprung gap. If we destroy more of these sources in the ensuing CEE phase, we lose more of the contributions from the RLOF+CEE route.

Decreasing α (Model C) has this kind of deleterious effect upon the RLOF+CEE pathway, but slightly increases the signal from CEE+CEE sources, since the systems that survive to the close DD stage were on average more widely spaced than for $\alpha = 3.0$, so that the giant stars were physically larger, i.e. more evolved, on average upon Roche lobe overflow, so gave rise to more massive WDs (also with more widely differing masses). This corresponds to moving the second peak in Fig. 3.15 to higher a. The lower mean chirp mass is largely attributable to the increased number of low-chirp mass interacting binaries present at this frequency, since the typical WD–WD mass ratio is larger, as described in section 3.7.1.2. Increasing the efficiency parameter (Model D) has the opposite effect upon each route. If, on the other hand, we use the common envelope formalism of Nelemans et al. (2000) (Model E), it becomes less simple to disentangle the two routes, since now they overlap somewhat in initial $a-$space, but since we know that this modification ought not to affect the RLOF+CEE contribution, we hold this fixed from Model A. The new CEE+CEE value turns out to be significantly enhanced, since a wider range of initial separations has been opened up to double common envelope survival. The nearer (by design) equality of WD pair masses leads to a decrease in the number of WD–WD AM CVn systems produced, and hence a smaller contribution from interacting systems than for Model A.

The envelope ionisation energy becomes a significant part of the energy balance in AGB stars, and so its inclusion is important in common envelope phases that commence at large orbital separations. Increasing the fraction of this energy included in the envelope binding energy (Model N) therefore decreases the number of wide binaries able to shrink enough to form close WD–WD pairs. Omitting it entirely (Model O), thus increasing the envelope binding energy, has the opposite effect.

Model P shows the greatest departure from Model A in terms of GW flux and mean chirp mass. The progenitor mass distribution for Model P is peaked towards higher mass (6–8 $M_{\odot}$) stars than for other models. These differences can be traced to the outcome of common envelope phases on the Hertzsprung Gap (HG). The BSE fitting formula returns values of λ substantially smaller than 0.5 for most HG stars, corresponding to a high degree of central concentration. Therefore using $\lambda = 0.5$ in Model P results in much less shrinkage in these situations.

High mass stars expand in radius by a large factor in their HGs, so that the final Roche contact (for both pathways) is often a common envelope phase involving a HG star. The survival rate from this CE phase is boosted by the fixed lambda as described above. The resulting GW flux is therefore also greatly boosted for these higher mass stars, whose descendent WDs are sufficiently massive that relatively few are required to dominate the background GW flux. Given however that small values of lambda are robust for HG stars (they are also seen in the calculations of Dewi & Tauris 2000), we choose not to consider this prescription as a reasonable uncertainty on the background.

The lower chirp mass $\langle\mathcal{M}\rangle$ seen for Model H is due to the inclusion of an increased number of interacting sources at 1 mHz, compared with Model A; the value $\langle\mathcal{M}'\rangle$ appropriate for just detached WD–WD pairs for this model (used in the previous section) is the same as for Model A.

Starting all systems with circular orbits (Model K) boosts the RLOF+CEE pathway, because fewer systems given initially tight orbits are lost due to immediate collision at periastron. Since systems descended from the RLOF+CEE route are generally higher-mass, the mean chirp mass for Model K is higher than for Model A. CEE+CEE route systems are little affected; the high-a peak in Fig. 3.15 is simply narrowed in a-space, since orbital separations are no longer altered by tidal circularisation before Roche contact.

Aside from orbital circularisation, the main role of tides in the evolution to the DD stage is in orbital shrinkage before Roche contact, due to spin-up of the giant star. Neglecting tidal effects is thus similar to increasing the common envelope efficiency parameter, i.e. the progenitors of close DDs from the CEE+CEE route have smaller initial orbital separations, and the DDs produced have smaller chirp masses on average. If on the other hand tidal effects were much stronger than in the BSE code, then we would expect little impact upon this route, since giant-star corotation is already typically achieved before Roche lobe overflow with the tides in BSE.

The CEE+CEE route is as expected largely unaffected when we perturb dynamically stable mass transfer on the Hertzsprung gap. Much as one might expect, the RLOF+CEE route is enhanced when one enhances the transfer on the Hertzsprung gap (Model L), so that more mass is transferred to the companion, and more systems avoid a common envelope phase during the first phase of mass transfer (which tends to lead to merger). The orbit is also widened to a greater extent during transfer, meaning that more systems will survive the common envelope phase when the secondary evolves. Making the transfer semiconservative (Model M) has an opposing effect; the orbit is widened less during stable overflow, meaning that more systems are destroyed in the ensuing common envelope phase.

The steeper Scalo IMF (Model F), normalised to the local space density of low-mass stars, produces fewer intermediate (and high) mass stars than the KTG IMF, and so fewer of the dominant

Model	% DD	$\langle q \rangle$	$\Omega_{\rm gw}(1\ {\rm mHz})$	$\langle \mathcal{M} \rangle$	N_0
Optimistic	26	0.75	5.99	0.46	1.85
A	18	0.62	3.57	0.45	1.17
Pessimistic	14	0.66	0.95	0.40	0.32

Table 3.4: Summary of the properties of the optimistic and pessimistic models, along with the fiducial Model A

progenitors in Fig. 3.14 are produced. More of the compact binaries are then descended from lower-mass progenitors than for Model A, giving rise to their lower mean chirp mass. If we had instead normalised to the local core-collapse supernova rate, as in Schneider et al. (2001), we would instead have ended up with a correspondingly *higher* background from Model F.

Altering the shape of the cosmic star formation history (Model J) has little impact upon the background, since most of the sources contributing have MS evolution times of less than a few Gyr (see Fig. 3.14). This is a strong argument in favour of using an integral constraint (such as IR luminosity density), and not a present-day constraint (such as local core-collapse supernova rate), since normalising according to the supernova rate introduces a strong dependence on the shape of the cosmic star formation history curve, through the difference in amplitude between the local rate and the rate at the peak of star formation, which can easily skew the overall normalisation.

Finally, models Q and R lead to larger gravitational wave backgrounds than Model A, mainly because lower metallicity stars tend to leave the main sequence earlier, and thus a greater fraction of the stellar mass in the universe today is present in the form of remnants. The difference in received flux is, however, slight, on the order of 10 per cent. We conclude that keeping detailed track of abundance variations is not essential to calculation at the present level of accuracy.

3.8 Outlook

Based on the above indications of which effects boost the GW background and which reduces it, we construct two models in an attempt to put upper and lower limits on the background we predict. Our use of the terms "optimistic" and "pessimistic" assumes that this background constitutes a signal for the reader; if it constitutes a noise, the nomenclature should be reversed.

Optimistic model: This has the properties of Model A, except for: the Nelemans et al. (2000) common envelope formalism, initially circular orbits, enhanced mass transfer on the HG, edge lit detonations only after accretion of 0.3 $\mathrm{M}_{\odot}$ and no ionisation energy in envelope binding energies used for common envelope phases. Note that some of these individually boosting effects do not make a double-boost in combination; for example the no spiral-in common envelope prescription tends to lead to DD mass ratios closer to unity, which means that fewer systems undergo stable mass transfer upon contact, and so the enhancement brought by the higher ELD limit is less effective in increasing the amplitude of the background. We also include the estimated error on our overall normalisation (see section 3.6.4), by using a cosmic star formation rate everywhere 30 per cent higher than our fiducial one.

Pessimistic model: The pessimistic model contains the elements found in the previous section to decrease the amplitude of the GW background. The properties of this model are thus the same as Model A, except for: $\alpha = 2.0$, Roche lobe overflow is semiconservative on the HG, the Scalo initial mass function is used and 100 per cent of the ionisation energy is included in the envelope binding energies used in common envelope phases. In addition, we use a star formation rate everywhere 30 per cent lower than our fiducial one, in our cosmological integral.

These prescriptions were used to create Galactic DD populations, which were found to compare reasonably with observations. Then the cosmological integrals were carried out for each. The results of this are summarised in Table 3.4, and the optimistic, fiducial and pessimistic total background spectra are plotted in Fig. 3.16 along with the LISA sensitivity curve, and the Galactic WD–WD background taken from Nelemans et al. (2001c). We plot both the "unresolved" ("average") background curve from their paper, which is for DD pairs only, with the resolved sources removed, and an extrapolated "total" background. In this we have added back in the resolved close binaries and made an approximation to the MS–MS contribution at lower frequencies, in an attempt to represent the Galactic signal over the full frequency range plotted.

Plotted in Fig. 3.17(a) is the number of systems per 1/(3 yr) frequency resolution element contributing to the GW background as received today. We see from this that at frequencies $f_{\mathrm{r}} \lesssim 50$

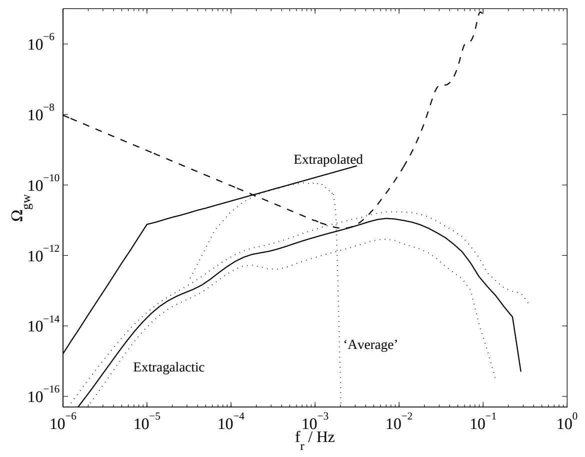

Figure 3.16: Optimistic (upper dotted), fiducial (Model A, lower solid line) and pessimistic (lower dotted) extragalactic backgrounds plotted against the LISA (dashed) single-arm Michelson combination sensitivity curve (see http://www.srl.caltech.edu/∼shane/sensitivity/). The 'unresolved' Galactic close WD–WD spectrum from Nelemans et al. (2001c) is plotted (with signals from binaries resolved by LISA removed), as well as an extrapolated total, in which resolved binaries are restored, as well as an approximation to the Galactic MS–MS signal at low frequencies.

mHz, there will be too many individual WD–WD sources contributing in each resolution element for this background to be completely resolved and subtracted source by source by missions with plausible lifetimes. However, from Fig. 3.17(b), we see that much of the flux comes from relatively nearby sources, and the WD–WD numbers drop rapidly above 50 mHz (leaving the lower background from rare neutron stars and black holes, not considered in this paper). Thus it may be possible for future missions more sensitive than LISA to subtract this background at high frequencies.

3.9 Comparison with previous work

Hils et al. (1990) and Kosenko & Postnov (1998) each made an order of magnitude estimate of the ratio of the extragalactic to Galactic GW flux from DDs. In order to facilitate comparison, and to compare like with like as far as possible, we divide our calculated extragalactic flux at 1 mHz by the most recently calculated value (Nelemans et al., 2001c) for the Galactic flux at the same frequency (which is a factor ∼ 3 smaller than that found by Webbink & Han 1998). The correct curve from Fig. 3.16 to use for this comparison is our "extrapolated" curve. We find $\Omega_{\rm extragal}/\Omega_{\rm gal} = 2.0$ per cent at 1 mHz for Model A, with a range of 0.5–3.4 per cent between optimistic and pessimistic

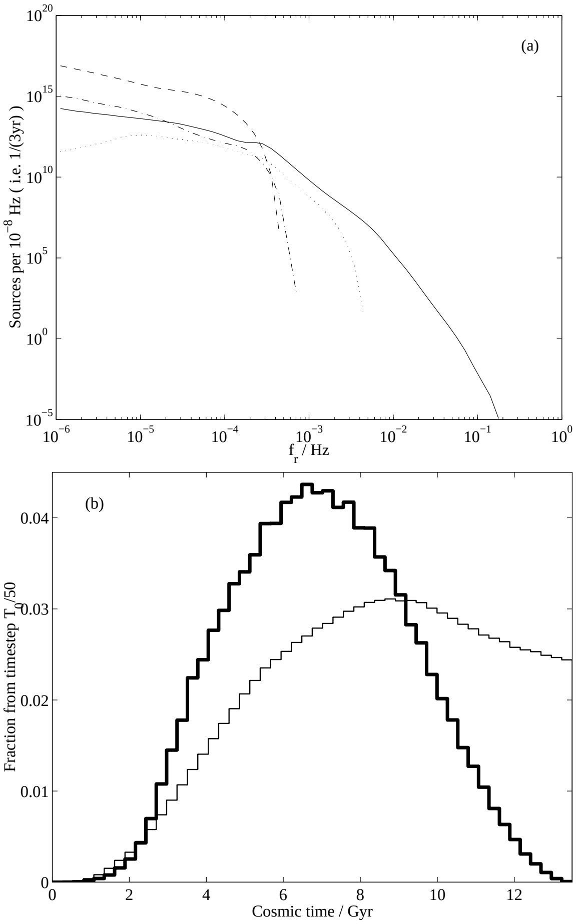

Figure 3.17: (a) The number of systems per 10^{-8} Hz contributing to the cosmological GW background as received today. Linestyles denote the evolutionary classes as in Fig. 3.4. (b) Thin line: the fractional contribution at 10 mHz to the GW background as a function of cosmic time (from shells of width $T_0/50$). Thick line: the same, but in terms of the number of sources contributing to the flux received from each cosmological time-shell.

models.

Hils et al. (1990) predicted a factor ~ 1.6 per cent (for an Einstein-de Sitter universe with no cosmological evolution of galactic GW luminosity). This estimate is in good agreement with our value.

Kosenko & Postnov (1998), on the other hand, predicted that, for a cosmology of the type used in this paper, the extragalactic background should be of order 10 per cent of the Galactic one, when one takes into account the evolution of star formation rate with redshift. This result is in clear disagreement with our findings, but this can be explained by noting that their ratio is artificially raised by a number of factors: first, the fiducial scalings of $\Omega_{\rm b}$, $\langle r \rangle$ and h_{100} in their eq. 13 are higher than their true values, boosting the extragalactic signal. Second, the same star formation rate as a function of redshift was used for different cosmologies, which leads to an artificial boost to the lambda-cosmology extragalactic flux (see e.g. Somerville, Primack & Faber, 2001). Lastly, the cosmic star formation rate adopted was not normalised to any integral constraint, but merely to the current star formation rate. All of these factors lead to their calculation yielding a misleadingly high extragalactic contribution to the GW background.

Schneider et al. (2001) made a more direct calculation of the background. At 1 mHz, their derived background level (for $h_{100} = 0.7$) is $\Omega_{\rm gw} = 1.2 \times 10^{-11}$, with no quoted uncertainty on this value. This lies a factor of two outside of our predicted range for the background. The discrepancy can be understood mainly in terms of their different method of normalisation: they normalised to the local core collapse supernova rate, and used the steep Scalo IMF, meaning that more low- and intermediate-mass stars were born in their simulations than measured by Cole et al. (2001). As explained in Section 3.6.4, we believe that normalising to an integral constraint on the birth of low-mass stars is a more robust method. Schneider et al. (2001) also used a binary fraction of 100 per cent, cf. our 50 per cent.

The *shape* of the spectrum in Schneider et al. (2001), however, we cannot explain. The spiral-in part of the spectrum ($f_{\rm r} \gtrsim 10^{-4}$ Hz) has the form expected from section 3.4, but the static regime instead displays a prominent 'bump' at frequencies ($f_{\rm r} \sim 3 \times 10^{-5}$ Hz) just below the transition to the spiral-in regime, the amplitude of which decays rapidly towards lower frequencies. No such feature is seen in our calculated spectra. This type of feature is difficult to explain in terms of the arguments in section 3.4, unless the vast majority of WD–WD pairs are born precisely into this 'bump', which seems unlikely, since the same feature is seen for all types of compact object pair (e.g. NS–NS, NS–BH), despite their very different formation routes.

3.10 Conclusions

We predict that the background of gravitational waves from extragalactic binary stars is

1. Dominated by double main sequence binaries for $f_{\rm r} < 10^{-4}$ Hz.

2. Dominated by double white dwarf binaries for $10^{-4} < f_{\rm r} < 10^{-1}$Hz.

Concentrating on the spectrum around 1 mHz:

1. The fraction of critical density in gravitational waves received in the logarithmic frequency interval around 1 mHz lies in the range $1 \times 10^{-12} < \Omega_{\rm gw} < 6 \times 10^{-12}$, with the most likely value in the range $3 - 4 \times 10^{-12}$.

2. The flux-weighted mean chirp mass of the contributing binaries is $\langle \mathcal{M} \rangle = 0.45$ M$_\odot$.

3. Half of the background comes from binaries whose more massive (primary) star had a mass in the range 2–4 M$_\odot$ (and $\sim$70 per cent from primaries originally less massive than 4 M$_\odot$). The estimate of the background is thus more robust to uncertainties in the IMF and mass cuts if normalised to the present density of starlight than if normalised to core-collapse supernova rates.

4. $\sim$ 60 per cent of the GW signal is from binaries with initial semi-major axes in the range of 30–1000 stellar diameters, in which the Roche contact of both primary and secondary stars led to unstable transfer and a common envelope. The background level produced by these systems is quite stable against uncertainties in the efficiency of the common envelope phase,

though the signal can be changed somewhat through use of a non-standard common envelope prescription.

5. ~ 40 per cent of the GW flux comes from systems descended from binaries with initial semi-major axes of about 5 stellar diameters, in which the first Roche contact occurred in the Hertzsprung gap, with stable overflow, but the second Roche contact led to unstable transfer and a common envelope. The background level produced by these systems is sensitive to uncertainties in common envelope and mass transfer physics.

6. interacting systems (AM CVn binaries) contribute only about 10 per cent of the energy density in gravitational waves.

The above holds true for $0.5 \lesssim f_{\rm r}(\rm mHz) \lesssim 5$. Above this range, as the lower-mass WD–WD pairs reach contact and drop out of the spectrum due to mergers, the properties change (values at 10 mHz in the parentheses that follow): the contribution from interacting binaries increases (26 per cent), the RLOF+CEE route contribution (44 per cent) and the mean primary progenitor mass increase (4.7 $\rm M_{\odot}$) and the mean chirp mass is higher (0.56 $\rm M_{\odot}$).

We find that at all frequencies, our derived spectral shape can be understood in terms of simple arguments, and that this shape is essentially independent of the population synthesis model used.

Acknowledgements

This research has been supported in part by NASA grant NAG5-10707. We thank J. Hurley for supplying the original BSE code, and for his advice and insights, and we also thank G. Nelemans and B. Hansen for useful discussions. We are also grateful to the referee, V. Kalogera, for her valuable comments.

Bibliography

Apreda R., Maggiore M., Nicolis A., Riotto, A. 2002, Nucl. Phys. B 631, 342

Belczynski K. Kalogera V., Bulik T. 2002, ApJ 572, 407

Cole S. et al. (2dFGRS Team), 2001, MNRAS, 326, 255

Cornish N., Larson S. L. 2002, gr-qc/0206017

Dewi J. D. M., Tauris T. M. 2000, A&A, 360, 1043

Duquennoy A., Mayor M., 1991, A&A, 248, 485

Evans C. R., Iben I, Smarr L. 1987 ApJ 323, 129

Finn L. S., Thorne K. S. 2000, Phys. Rev. D 62, 124021

Giampieri G., Polnarev A. G. 1997, MNRAS 291, 149

Han Z., Tout C. A., Eggleton P. P., 2000, MNRAS, 319, 215

Hansen B. M. S., 1999, ApJ, 520, 680

Hils D., 1998, in Folkner W. M., ed, AIP Conf. Proc. Vol. 456, Second International LISA Symposium on the Detection and Observation of Gravitational Waves in Space, p68

Hils D., Bender P. L., 1995, ApJ, 445, L7

Hils D., Bender P. L., 2000, ApJ, 537, 334

Hils D., Bender P. L., Webbink R. F., 1990, ApJ, 360, 75 (errata ApJ, 369, 271)

Hogan C. J., 2000, Phys Rev D, 62, 121302

Hogan C. J., Bender P.L., 2001 Phys. Rev. D, 64, 062002

Hogg D. W., 2000, astro-ph/9905116

Hughes S. A., 2001, Class. Quant. Grav. 18, 4067

Hurley J. R., 2000, PhD Thesis, Univ. Cambridge

Hurley J. R., Shara M. M., 2003, preprint (astro-ph/0302119)

Hurley J. R., Tout C. A., Pols O. R., 2002, MNRAS, 329, 897

Iben I., Livio M., 1993, PASP, 105, 1373

Israel G. L. et al., 2002, A&A, 386, L13

Jeffries R. D., 1997, MNRAS, 288, 585

Kalogera V., Webbink R. F., 1998, ApJ, 493, 351

Kamionkowski M., Kosowsky A., Turner M. S., 1994, Phys Rev D 49, 2837

Karl C., Napiwotzki R., Heber U., Lisker T., Nelemans G., Christlieb N., Reimers D., 2002, in de Martino D., Kalytis R., Silvotti R., Solheim J. E., eds, Proc. XIII Workshop on White Dwarfs, Kluwer, in press (astro-ph/0210004)

Kennicutt R. C., 1983, ApJ, 272, 54

Kosenko D. I., Postnov K. A., 1998, A&A, 336, 786

Kosenko D. I., Postnov K. A., 2000, A&A, 355, 1209

Kosowsky A., Mack A., Kahniashvili T., 2002, Phys Rev D 66, 024030

Kroupa P., Tout C. A., Gilmore G., 1993, MNRAS, 262, 545

Lipunov V. M., Postnov K. A., Prokhorov M. E., 1987, A&A, 176, L1

Lipunov V. M., Nazin S. N., Panchenko I. E., Postnov K. A., Prokhorov M. E., 1995, A&A, 298, 677

Livio M., Soker N., 1988, ApJ, 329, 764

Madau P., Ferguson H. C., Dickinson M. E., Giavalisco M., Steidel C. C., Fruchter A., 1996, MNRAS, 283, 1388

Marsh T. R., 2000, NewAR 44, 119

Marsh T. R., Dhillon V. S., Duck S. R., 1995, MNRAS, 275, 828

Maxted P. F. L., Marsh T. R., 1999, MNRAS, 307, 122

Maxted P. F. L., Marsh T. R., Moran C. K. J., 2002, MNRAS, 332, 745

Napiwotzki R., et al., 2002, in de Martino D., Kalytis R., Silvotti R., Solheim J. E., eds, Proc. XIII Workshop on White Dwarfs, Kluwer, in press (astro-ph/0210155)

Nelemans G., Verbunt F., Yungelson L. R., Portegies Zwart S. F., 2000, A&A, 360, 1011

Nelemans G., Portegies Zwart S. F., Verbunt F. Yungelson L. R., 2001a, A&A, 368, 939

Nelemans G., Yungelson L. R., Portegies Zwart S. F., Verbunt F., 2001b, A&A, 365, 491

Nelemans G., Yungelson L. R., Portegies Zwart S. F., 2001c, A&A, 375, 890

Peacock J. A., 1999, Cosmological Physics, Cambridge Univ. Press, Cambridge

Peters P. C., Mathews J., 1963, Phys. Rev., 131, 435

Phinney E. S., 2002, MNRAS, in press, astro-ph/0108028

Salpeter E. E., 1955, ApJ, 121, 61

Savonije G., de Kool M., van den Heuvel E. P. J., 1986, A&A, 155, 51

Scalo J., Fund. Cosm. Phys. 11, 1

Schneider R., Ferrari V., Matarrese S., Portegies Zwart S. F., 2001, MNRAS, 324, 797

Somerville R. S., Primack J. R., Faber S. M., 2001, MNRAS, 320, 504

Steidel C. C., Adelberger K. L., Giavalisco M., Dickinson M., Pettini M., 1999, ApJ, 519, 1

Taam R. E., Sandquist E. L., ARAA, 2000, 38, 113

Tauris T. M., Dewi J. D. M. 2001 A&A, 369, 170

Tinto M., Armstrong J. W., Estabrook F. B., 2001, Phys Rev D, 63, 021101

Tout C. A., Aarseth S. J., Pols O. R., Eggleton P. P., 1997, MNRAS, 291, 732

Turner M. S., 1997, Phys. Rev. D, 55, 435

Webbink R. F., 1984, ApJ, 277, 355

Webbink R. F., Han Z., 1998, in Folkner W. N., ed, AIP Conf. Proc. Vol. 456, Second International LISA Symposium on the Detection and Observation of Gravitational Waves in Space, p61

Chapter 4

Wave Damping by MHD Turbulence and its Effect upon Cosmic Ray Propagation in the ISM

Originally appeared as

Alison J. Farmer & Peter Goldreich,

“Wave damping by magnetohydrodynamic turbulence and its effect on cosmic ray propagation in the interstellar medium,”

Astrophysical Journal, vol. 604, pp 671–674

Reproduced with permission.

Abstract

Cosmic rays scatter off magnetic irregularities (Alfvén waves) with which they are resonant, that is waves of wavelength comparable to their gyroradii. These waves may be generated either by the cosmic rays themselves, if they stream faster than the Alfvén speed, or by sources of MHD turbulence. Waves excited by streaming cosmic rays are ideally shaped for scattering, whereas the scattering efficiency of MHD turbulence is severely diminished by its anisotropy. We show that MHD turbulence has an indirect effect on cosmic ray propagation by acting as a damping mechanism for cosmic ray generated waves. The hot ("coronal") phase of the interstellar medium is the best candidate location for cosmic ray confinement by scattering from self-generated waves. We relate the streaming velocity of cosmic rays to the rate of turbulent dissipation in this medium, for the case in which turbulent damping is the dominant damping mechanism. We conclude that cosmic rays with up to 10^2 GeV could not stream much faster than the Alfvén speed, but that 10^6 GeV cosmic rays would stream unimpeded by self-generated waves unless the coronal gas were remarkably turbulence-free.

4.1 Introduction

Cosmic ray (CR) scattering by resonant Alfvén waves has been proposed to be essential to their acceleration by shocks (e.g., Bell, 1978) and to their confinement within the Galaxy (e.g., Kulsrud & Pearce, 1969).

Much of the interstellar medium (ISM) is thought to be turbulent, providing a ready source of Alfvén waves. However, MHD turbulence has the property that, as energy cascades from large to small scales, power concentrates in modes with increasingly transverse wavevectors, i.e., perpendicular to the background magnetic field direction (Goldreich & Sridhar, 1995, 1997). Cosmic rays scatter best off waves that have little transverse variation, so CR scattering by MHD turbulence is necessarily extremely weak, leading to very long CR mean free paths (e.g., Chandran, 2000a; Yan & Lazarian, 2002).

If cosmic rays stream faster than the Alfvén speed, they can amplify waves (naturally of the correct shape for scattering) through the resonant streaming instability (see Wentzel, 1974). As the waves amplify, the scattering strength increases and the streaming velocity is reduced. For this process of self-confinement to operate, the excitation rate of the waves by streaming cosmic rays must exceed the sum of all rates of wave damping.

Wave damping depends upon the properties of the medium in which the CRs propagate. Important mechanisms include ion-neutral collisions in regions of partial ionization, and non-linear Landau damping in the collisionless limit. In this paper we introduce another mechanism, wave damping by background MHD turbulence. As cosmic ray–generated waves propagate along magnetic field lines, they are distorted in collisions with oppositely directed turbulent wavepackets. As a result, the wave energy cascades to smaller scales and is ultimately dissipated. This process, which is best viewed geometrically, is described in § 4.2.2. MHD turbulence thus becomes an impediment to the scattering of cosmic rays, as opposed to just an ineffective scatterer of them. This mechanism was mentioned briefly by Cho, Lazarian & Vishniac (2003), Lazarian, Cho & Yan (2002) and Yan & Lazarian (2002).

The paper is arranged as follows. Relevant properties of the MHD cascade are described in

§4.2.1, followed by an explanation of the turbulent damping rate in §4.2.2. In §4.3 we describe the competition between growth and damping of waves due to CR streaming. We apply these ideas to the problem of Galactic CR self-confinement in §4.4, and use this to place limits on the cascade rate of the turbulence in the coronal gas, assuming that the observed streaming velocities are due to self-confinement in this medium. In §4.4.1 we compare with other work in this area and in §4.5 we conclude.

4.2 The MHD cascade as a damping mechanism

4.2.1 Relevant properties of the cascade

The strong incompressible MHD cascade proposed by Goldreich & Sridhar (1995, 1997) has the property that as the cascade proceeds to smaller scales, power becomes increasingly concentrated in waves with wavevectors almost perpendicular to the local mean magnetic field. We envisage a situation in which turbulence is excited isotropically at an MHD outer scale $L_{\rm mhd}$ with RMS velocity fluctuations $v \sim v_{\rm A}$ and magnetic field fluctuations $\delta B \sim B_0$, where B_0 is the magnitude of the background magnetic field.[1] Well inside the cascade, the variations parallel to the magnetic field are much more gradual than those perpendicular to it, i.e., $v(\boldsymbol{\lambda}) \simeq v_{\lambda_\perp} \gg v_{\lambda_\parallel}$ if $\lambda_\perp \sim \lambda_\parallel$.[2] Equivalent relations hold for magnetic field fluctuations. For fluctuations of a given amplitude, therefore, the correlation length (defined so that $v_{\lambda_\perp} \sim v_{\Lambda_\parallel}$) parallel to magnetic field lines, $\Lambda_\parallel$, is much greater than that perpendicular to them, $\lambda_\perp$. Turbulent eddies are highly elongated parallel to magnetic field lines.

Strong MHD turbulence is characterized by "critical balance." In other words, a wavepacket shears at a rate which is comparable to its frequency $\omega = v_{\rm A} k_\parallel \sim v_{\rm A}/\Lambda_\parallel$ and is also of order $v_{\lambda_\perp}/\lambda_\perp$. Thus

$$\frac{v_{\lambda_\perp}}{\lambda_\perp} \sim \frac{v_{\rm A}}{\lambda_\parallel}\,. \tag{4.1}$$

[1] Turbulence can also be injected at smaller velocities on smaller scales, in which case $L_{\rm mhd}$ should be considered an extrapolation beyond the actual outer scale of the cascade.

[2] Throughout this paper, "perpendicular" and "parallel" wavelengths refer respectively to the inverse of the wavevector components perpendicular and parallel to the background magnetic field direction.

Application of the Kolmogorov argument for the constancy of the energy cascade rate ϵ per unit mass yields

$$\epsilon \sim \frac{v^2}{t_{\mathrm{cascade}}} \sim \frac{v_{\lambda_\perp}^3}{\lambda_\perp} \sim \frac{v_{\mathrm{A}}^3}{L_{\mathrm{mhd}}}, \tag{4.2}$$

from which we obtain the fluctuation amplitude on perpendicular scale $\lambda_\perp$,

$$v_{\lambda_\perp} \sim v_{\mathrm{A}} \left(\frac{\lambda_\perp}{L_{\mathrm{mhd}}}\right)^{1/3} \sim (\epsilon\lambda_\perp)^{1/3}. \tag{4.3}$$

An analogous relation holds for magnetic field perturbations. Well inside the cascade, $v \ll v_{\mathrm{A}}$ and $\delta B \ll B_0$. We combine equations (4.1) and (4.3) to obtain the eddy shape:

$$\Lambda_\parallel(\lambda_\perp) \sim L_{\mathrm{mhd}}^{1/3}\lambda_\perp^{2/3} > \lambda_\perp \,. \tag{4.4}$$

4.2.2 The turbulent damping rate

The energy cascade from large to small scales in MHD turbulence is due to distortions produced in collisions between oppositely directed Alfvén wavepackets. This is best visualized geometrically as being due to the shearing of wavepackets as they travel along wandering magnetic field lines. A good description is given in Lithwick & Goldreich (2001).

Consider the fate of a wavepacket with initial perpendicular and parallel wavelengths $\lambda_\perp$ and $\lambda_\parallel$. It suffers an order unity shear after traveling over a distance along which the fields lines that guide it spread by order $\lambda_\perp$. By then the energy it carries has cascaded to smaller scales, ultimately to be dissipated as heat at the inner scale. This process occurs not only for waves that are part of the turbulent cascade, but also for any other Alfvén waves in the medium. As these waves travel along the field lines, they are distorted in collisions with oppositely directed turbulent wavepackets.

On a perpendicular scale $\lambda_\perp$, the field lines spread by order unity over a parallel distance $\Lambda_\parallel$, where $\Lambda_\parallel(\lambda_\perp)$ is a property of the background turbulence and is given by equation (4.4). Therefore any wavepacket of perpendicular scale $\lambda_\perp$ cascades once it travels this distance. Because of the nature of the MHD cascade, this corresponds to many wave periods for a wave with $\lambda_\parallel \lesssim \lambda_\perp \ll \Lambda_\parallel$

(but to one wave period for waves shaped like those in the turbulent cascade, as described by critical balance). The damping rate is a function of $\lambda_\perp$:

$$\Gamma_{\rm turb} \sim \frac{1}{t_{\rm cascade}(\lambda_\perp)} \sim \frac{v_{\lambda_\perp}}{\lambda_\perp} \sim \frac{v_{\rm A}}{L_{\rm mhd}^{1/3}\lambda_\perp^{2/3}} \sim \frac{\epsilon^{1/3}}{\lambda_\perp^{2/3}}. \tag{4.5}$$

This damping rate applies to any wave with perpendicular wavelength $\lambda_\perp$ propagating in a background of strong MHD turbulence, so long as $L_{\rm mhd} \gg \lambda_\perp \gg l_{\rm dissipation}$. The appropriate value of $\lambda_\perp$ to use for CR-generated waves will be considered in §4.3.2.

4.3 Competition between growth and damping

4.3.1 Resonant scattering of cosmic rays

As cosmic rays stream along magnetic field lines, they are scattered in pitch angle by magnetic irregularities (Alfvén waves, of appropriate shape; see below), and thus exchange momentum (and energy) with particular waves. If cosmic rays stream faster than the Alfvén speed, they can excite Alfvén waves traveling in the same direction. Provided the excitation rate exceeds the total damping rate due to other processes, the waves amplify exponentially. Initial perturbations too weak to significantly scatter CRs can strengthen until the scattering reduces the CR streaming velocity. Even thermal fluctuations could provide seed waves in the absence of other sources. The reduction of the streaming velocity by cosmic ray–amplified waves is known as self-confinement. Next we describe which random fluctuations are selectively amplified by cosmic ray protons with energy γ GeV.

4.3.1.1 Parallel lengthscale

Cosmic rays spiraling along a mean magnetic field $\mathbf{B_0} = B_0\hat{\mathbf{z}}$ scatter in pitch angle off Alfvén waves with which they are parallel-resonant, i.e., waves for which

$$k_\parallel = \frac{1}{\mu r_L}, \tag{4.6}$$

where μ is the cosine of the particle's pitch angle, and r_L is its gyroradius (Kulsrud & Pearce, 1969; Wentzel, 1974). On the timescale of the CR's passage, the wave is almost static since the CR is relativistic and $v_{\mathrm{A}} \ll c$. Thus the wave's time dependence is neglected in the above resonance condition. When resonance holds, the cosmic ray experiences a steady direction-changing force.

4.3.1.2 Perpendicular lengthscale

A cosmic ray is most efficiently scattered by parallel-propagating waves, $\lambda_\perp \gg \lambda_\| \sim r_L$, because in these the direction changing force maintains a steady direction in one gyroperiod. Moving through waves that have significant perpendicular components, $\lambda_\perp \ll \lambda_\|$, the cosmic ray traverses many perpendicular wavelengths, leading to oscillations of the direction-changing force and inefficient scattering. This explains why cosmic rays are weakly scattered by MHD turbulence (see, e.g., Chandran, 2000a; Yan & Lazarian, 2002), and also why the waves in the turbulent cascade damp faster than cosmic rays can excite them.

The closer to parallel waves propagate, the faster streaming cosmic rays can excite them. The growth rate for waves that are parallel-resonant and reasonably close to parallel-propagating ($\lambda_\| \lesssim \lambda_\perp$) is given by (see Kulsrud & Pearce, 1969)

$$\Gamma_{\mathrm{cr}}(k_\|) \sim \Omega_0 \frac{n_{\mathrm{cr}}(>\gamma)}{n_{\mathrm{i}}} \left(\frac{v_{\mathrm{stream}}}{v_{\mathrm{A}}} - 1 \right), \tag{4.7}$$

where v_{stream} is the net streaming velocity of the cosmic rays measured in the rest frame of the ISM, $\Omega_0 = eB_0/mc$ is the CR cyclotron frequency in the mean field, n_{i} is the ion number density in the ISM, and $n_{\mathrm{cr}}(>\gamma)$ is the number density of cosmic rays with gyroradius $r_L > \gamma mc^2/eB_0 = 1/k_\|$, i.e., those particles which can, for the appropriate value of μ, be resonant with waves of parallel wavevector $k_\|$. Because the cosmic ray energy spectrum is steep, the energies of most resonant particles are close to the lowest energy that permits resonance with the wave. Therefore we associate $k_\| \sim 1/r_L(\gamma)$ and $n_{\mathrm{cr}}(>\gamma) \simeq \gamma n_{\mathrm{cr}}(\gamma)$.[3]

[3] Particles with close to 90-degree pitch angles ($\mu \ll 1$) are scattered mainly by mirror interactions (Felice & Kulsrud, 2001).

4.3.2 Growth and damping

Growth rates are highest, and damping rates lowest, for the most closely parallel-propagating waves, that is, those waves with largest $\lambda_\perp$. Therefore, we consider the limiting case of the most parallel-propagating wave that can be excited. This most-parallel wave sets the minimum streaming velocity required for the instability to operate. The limit to parallel propagation is set by the turbulent background magnetic field: the largest wave aspect ratio possible is fixed by the straightness of the field lines. In the presence of MHD turbulence, the field direction depends on position. The change in direction across scale $\lambda_\perp$ is set by turbulent field fluctuations on this scale. It is not meaningful to talk about waves propagating at an angle less than $\delta B(\lambda_\perp)/B_0$ away from parallel, because the field direction changes by this much across the wavepacket. We can therefore only have waves with

$$\frac{\lambda_\parallel}{\lambda_\perp} > \frac{\delta B(\lambda_\perp)}{B_0} \sim \left(\frac{\lambda_\perp}{L_{\rm mhd}}\right)^{1/3} \sim \left(\frac{\epsilon r_L}{v_{\rm A}^3}\right)^{1/4}, \tag{4.8}$$

where we have used $\lambda_\parallel \sim r_L$, the resonance condition.

To obtain the damping rate of the most closely parallel-propagating wave, we substitute equation (4.8) into equation (4.5), which yields

$$\Gamma_{\rm turb,min} \sim \left(\frac{\epsilon}{r_L v_{\rm A}}\right)^{1/2}; \tag{4.9}$$

all other waves damp faster than this one, for given r_L.

We can view the damping as being due to the introduction of perpendicular wavevector components to the CR-generated wave. This is how the background turbulence cascades, and the CR-generated wave is being integrated into the cascade. We can decompose the modified wave into components with almost-perpendicular and almost-parallel wavevectors. The perpendicular part is not excited and is more strongly damped, but the almost-parallel-propagating component continues to be amplified by resonant cosmic rays.

For the instability to operate, we require the maximum possible growth rate (eq. [4.7]) to be larger than the minimum damping rate (eq. [4.9]):

$$\Gamma_{\rm cr}[F_{\rm A}(\gamma)] > \Gamma_{\rm turb,min}(\gamma). \tag{4.10}$$

where $F_{\rm A}(\gamma) \sim (v_{\rm stream} - v_{\rm A})n_{\rm cr}(> \gamma)$ is the cosmic ray flux measured in the frame moving with the waves. Equation (4.10) can be written in the form $F_{\rm A}(\gamma) > F_{\rm crit}(\gamma)$. If $F_{\rm A}$ is less than $F_{\rm crit}$ then wave amplification does not occur and the cosmic rays are not significantly scattered. Equivalently, the resonant streaming instability cannot reduce $F_{\rm A}$ below $F_{\rm crit}$.

If the instability is to confine cosmic rays to regions of shock acceleration, or to the Galaxy (which we discuss in the next section), then the level of background turbulence must be low enough to permit the growth of resonant waves.

4.4 Application to cosmic ray self-confinement in the ISM

Cosmic rays are preferentially produced in the denser regions of the Galaxy, and they escape from its edges. Two lines of evidence imply that they do not stream freely out of the Galaxy: the CR flux in the solar neighborhood is observed to be isotropic to within ~0.1% at energies less than $\sim 10^6$ GeV, and the abundance of the unstable nucleus ^{10}Be produced by spallation establishes that CRs are confined within the Galaxy for $\sim 10^7$ years (Schlickeiser, 2002).

Scattering by Alfvén waves has been viewed as the leading mechanism for confinement. Both waves associated with background MHD turbulence and those resonantly excited by cosmic rays have been considered in this regard. Prior to the recognition that MHD turbulence is anisotropic, the former were generally favored. Now self-confinement appears to be the more viable option.

The most promising location for the operation of the streaming instability is the hot ISM (HISM), a.k.a. the coronal gas (Cesarsky & Kulsrud, 1981; Felice & Kulsrud, 2001). The abundances of cosmic ray nuclei produced by spallation suggest that cosmic rays spend about two-thirds of their time in this medium (see, e.g., Schlickeiser, 2002). Ion-neutral damping of waves is ineffective in the HISM. The coronal gas is hot ($T \sim 10^6$ K) and tenuous ($n_{\rm i} \sim 10^{-3}$ cm^{-3}), with an Alfvén velocity, assuming $B_0 \sim 3\ \mu$G, of $v_{\rm A} \sim 2 \times 10^7$ cm s^{-1}. The gyroradius of a relativistic proton in this field, $r_L \sim 10^{12}\gamma$ cm, lies within the inertial range of the MHD cascade.

Assuming the cosmic ray density in the HISM to be similar to that near the Sun,[4] $n_{\rm cr}(>\gamma) \simeq 2 \times 10^{-10}\gamma^{-1.6}$ cm^{-3} (Wentzel, 1974), we can calculate the velocity above which the streaming instability in the HISM would turn on, assuming our turbulent damping to be the dominant damping mechanism. To accomplish this, we substitute equations (4.7) and (4.9) into inequality (4.10), treating it as an equality. We find

$$\begin{aligned} v_{\rm stream} &\sim v_{\rm A}\left[1 + \frac{n_{\rm i}}{n_{\rm cr}(\gamma)}\frac{\omega_0}{\Omega_0}\left(\frac{L_{\rm mhd}}{r_L}\right)^{1/2}\right] \\ &\sim v_{\rm A}\left[1 + \left(\frac{\epsilon}{700\ \text{erg s}^{-1}\ \text{g}^{-1}}\right)^{1/2}\gamma^{1.1}\right], \end{aligned} \tag{4.11}$$

where $\omega_0 = v_{\rm A}/L_{\rm mhd}$ is the turbulent decay rate on the outer scale.

The mean rate at which turbulent dissipation heats the coronal gas is unlikely to exceed its radiative cooling rate, $\epsilon \sim 0.06$ erg g^{-1} s^{-1} for solar abundances (Binney & Tremaine, 1987, , p580). Unfortunately, we do not know whether the heating is continuous or episodic, and what fraction is due to shocks as opposed to turbulence.[5]

Roughly one supernova explosion occurs per century in the Galaxy, or on average one per square 100 pc of the disk every 1×10^6 yr. Turbulence injected with $v \sim v_{\rm A}$ on scales $L \sim 100$ pc decays in a time $L/v_{\rm A} \sim 5 \times 10^5$ yr, so it might be replenished before decaying. However, supernovae occur predominantly in the Galactic plane and it is uncertain how effective they are in stirring the coronal gas, which has a large vertical scale height. Suppose that each supernova releases 10^{51} erg of mechanical energy that is ultimately dissipated by turbulence. This amounts to a dissipation rate of 3×10^{41} erg s^{-1}, which, if evenly distributed by volume throughout a disk of radius 10 kpc and thickness 1 kpc, would provide a mean heating rate of $\bar{\epsilon} \sim 25$ erg s^{-1} g^{-1} in the HISM. This value is much greater than our estimate of the radiative cooling rate.

[4] If $n_{\rm cr}$ is lower than it is near the Sun, then confinement will begin to be problematic at lower energies, and vice versa

[5] It seems plausible that shocks, especially if they intersect, would efficiently excite turbulence.

The cosmic ray anisotropy measured locally is $\lesssim 0.1\%$ for $\gamma \lesssim 10^6$ (Schlickeiser, 2002), i.e. up to the "knee" in the CR energy spectrum. The Alfvén velocity in the HISM is of the same order as the local streaming velocity: $v_\mathrm{A}/c \simeq 0.1\%$. Substituting into equation (4.11) the value of ϵ obtained by balancing the radiative cooling of the hot gas with heating due to steady state turbulent dissipation, we obtain

$$v_\mathrm{stream} \sim v_\mathrm{A}(1 + 9 \times 10^{-3}\gamma^{1.1}). \tag{4.12}$$

Equation (4.12) suggests that self-confinement in the HISM might account for the small observed cosmic ray anisotropy up to $\gamma \lesssim 10^2$, but not much beyond. To limit the streaming velocity of protons with $\gamma \sim 10^6$ to $\sim v_\mathrm{A}$ would require the turbulent dissipation rate to be astonishingly low, $\epsilon \lesssim 4 \times 10^{-11}$ erg s^{-1} g^{-1}.

4.4.1 Comparison with previous work

That background MHD turbulence might be an impediment to the self-confinement of cosmic rays was mentioned briefly in Lazarian et al. (2002), Yan & Lazarian (2002) and Cho et al. (2003).

Kulsrud (1978) proposed non-linear Landau damping as the dominant damping mechanism for cosmic ray-generated waves in the HISM. Wave damping occurs when plasma ions "surf" on beat waves produced by the superposition of CR-generated waves. The damping rate for this process,[6] for similar HISM parameters as adopted in this paper, gives $v_\mathrm{stream} \simeq v_\mathrm{A}(1 + 0.05\gamma^{0.85})$ (Cesarsky & Kulsrud, 1981). This predicted streaming velocity is not very different from that obtained in equation (4.12). Both damping mechanisms are too strong to permit self-confinement to reduce the streaming velocity of high energy cosmic rays to the locally observed levels.

Chandran (2000b) proposes that magnetic mirror interactions in dense molecular clouds may provide confinement of high energy cosmic rays. The present paper provides further support for the idea that a confinement mechanism other than scattering by Alfvén waves is dominant for high energy cosmic rays.

[6] We use the unsaturated damping rate, as justified in Felice & Kulsrud (2001).

4.5 Conclusions

A background of anisotropic MHD turbulence acts as a linear damping mechanism for MHD waves excited by the streaming of cosmic rays. Low energy cosmic rays are numerous enough to excite Alfvén waves in the HISM when streaming at velocities compatible with observational limits on their anisotropy. However, high energy Galactic cosmic rays could only be self-confined to stream this slowly if turbulent dissipation in the HISM accounted for only a tiny fraction of its heat input.

Acknowledgements

This research was supported in part by NSF grant AST 00-98301. The authors thank Russell Kulsrud for raising the issue discussed in this paper and for illuminating discussions.

Bibliography

Bell, A. R. 1978, MNRAS, 182, 147

Binney, J. & Tremaine, S. 1987, Galactic Dynamics, Princeton

Cesarsky, C. J. 1980, ARA&A, 18, 289

Cesarsky, C. J., & Kulsrud, R. M. 1981, in IAU Symp. 94, Origin of Cosmic Rays, ed. G. Setti, G. Spada, & A. W. Wolfendale (Dordrecht: Reidel), 251

Chandran, B. D. G. 2000a, Phys Rev Lett, 85, 22

Chandran, B. D. G. 2000b, ApJ, 529, 513

Cho, J., Lazarian, A., Vishniac, E. T. 2003, Lect Notes Phys, 614, 56 (astro-ph/0211031)

Felice, G. M., & Kulsrud, R. M., 2001, ApJ, 553, 198

Goldreich, P., Sridhar, S. 1995, ApJ, 438, 763

Goldreich, P., Sridhar, S. 1997, ApJ, 485, 680

Lazarian, A, Cho, J., & Yan, H., 2002, to be published in Recent Research Developments in Astrophysics (Research Signpost) (astro-ph/0211031)

Lithwick, Y., & Goldreich, P., 2001, ApJ, 562, 279

Kulsrud, R. M. 1978, in Astronomical Papers dedicated to Bengt Strømgren, ed. A. Reiz & T. Anderson (Copenhagen: Copenhagen Univ. Obs.), 317

Kulsrud, R. M., & Pearce, W. P. 1969, ApJ, 156, 445

Schlickeiser, R. 2002, Cosmic Ray Astrophysics, Springer-Verlag

Wentzel, D. G. 1974, ARA&A, 12, 71

Yan, H. & Lazarian, A. 2002, Phys Rev Lett, 89, 281102

Chapter 5

Ghosts of Saturn: Studies for the Ring Spokes

Abstract

We conduct a thorough tour of the Saturn system as relevant to the formation of ring spokes, the transient radial dust lanes originally observed by *Voyager* and believed to be electrodynamic in origin. We present a modern review of spoke properties, and investigate in detail the environment in which they form, through a series of physical models. We discuss the formation and charging of dust grains, the currents and the electric fields in the rings, and the significant implications of the discovery by *Cassini* of a ring ionosphere. We hope that these new observations, coupled with the familiarity we have with the environment, will lead us to a new understanding of the mysterious spokes.

5.1 Introduction

In 1980, at a distance of 1/3 AU from Saturn, *Voyager 1* started to observe dark, radial features in Saturn's B ring. These transient dust lanes quickly became known as "spokes." Spoke composition and behavior has been well documented, but the origin of spokes is still unknown. Every proposed theory has serious deficiencies, so cannot account for the observations.

Long recognized as a mystery, spokes are back in the spotlight for two reasons. First, paradoxically, because they have disappeared: *HST* observations tracked the fading and eventual disappearance of the phenomenon between 1996 and 1998. The second reason is the arrival of the *Cassini* spacecraft at Saturn. The exquisite measurements *Cassini* is making of the planet, rings and magnetosphere have the potential to reveal clues that have been missing thus far. The ring system, particularly in the light of new discoveries by *Cassini*, is a gold mine of physics.

In this chapter, we make a series of studies of the environment in which spokes form. We proceed as follows. In §5.2, we describe the observed properties of spokes, and detail the immediate conclusions that can be drawn about their formation and evolution. To have a chance of understanding the spokes, it is important to be familiar with the environment in which they form. We present a summary of relevant properties of the planet and rings in §5.3. The interaction of the ring plane with solar photons, the major difference between day and night on the rings, is discussed in §5.4. We predict the properties of the dust above the rings in §5.5, since spokes must ultimately be made of charged dust. Dust charging is explored in §5.6. In §5.7, we analyze the fascinating new results from *Cassini* on the presence of a molecular oxygen atmosphere and ionosphere over the rings. §5.8 is devoted to the electrodynamics of the ionosphere-magnetosphere-ring system. We calculate the currents expected to flow between the rings and the planet from a variety of sources, and we place a limit on the thickness of the ring atmosphere, using the observation that the ring ionosphere corotates with Saturn. The planet's ionosphere is the subject of §5.9, in which we estimate its electrical conductivity. A summary of the spoke environment is presented in §5.10. We then move on to discuss a new theory that has some potential, as well as previous theories of spoke formation, as well as a new theory that has some potential, in §5.11 and §5.12. In §5.13 we conclude.

5.2 Observations of the spokes: the clues

The nature of the "spokes" in Saturn's rings remains a matter of speculation 25 years after their discovery by the *Voyager* spacecraft. We summarize their properties here; for more details see reviews by Mendis et al. (1984), Cuzzi et al. (1984) and references therein.

1. **Spokes are transient radial albedo features superposed on Saturn's rings.**

2. **Spokes are composed of dust with a narrow size distribution centered at 0.5 micron.** Spokes are darker than surrounding ring regions in backscattered light, but viewed from the unilluminated side of the rings they appear brighter than their surroundings. This suggests that spokes are composed of particles of size comparable to visible wavelengths. Modeling the scattering as a function of wavelength, McGhee et al. (2005) found the peak size

$a_{\text{eff}} = 0.57 \pm 0.05 \mu$m, with a narrow size distribution: less than a decade in a. The rest of the main rings is remarkably dust-free (Doyle et al., 1989), consistent with zero and with an upper limit certainly less than 1% in optical depth.

3. **Spokes are about 10% darker than non-spoke regions in backscattered light** (Smith et al. 1981, 1982). The normal optical depth in dust required for this contrast is of order $\tau \sim 0.01$ (Doyle et al., 1989). The mass of dust in a spoke is then $m_{\text{spoke}} \sim 10^{11}$ g. From edge-on ring observations, the dust layer must be < 80 km in height.

4. **Spokes are only seen near corotation**, which is where the Keplerian angular velocity of the ring particles matches the planet's rotational angular velocity. Corotation occurs in the outer B ring. Spokes are seen on each side of corotation, and some extend across it.

5. **Individual spokes extend over about 10% of the ring radius,** measuring ~ 10,000 km in radial length and ~ 2,000 km in azimuthal width. Spokes come in a variety of morphologies. Many are wedge-shaped, with the apex at corotation. Those spokes that are most nearly radial tend to be narrower. The inner edge of the spoke zone (at $1.72 R_S$) is particularly sharp.

6. **Spokes are seen preferentially on the morning ansa.** Spokes have been seen more frequently on the morning ansa, and spokes seen there tend to be of higher contrast than those seen elsewhere (Smith et al., 1981).

7. **Spokes fade as they are distorted by differential rotation.** As spokes are followed from morning toward evening ansa, their edges usually move with Keplerian velocities. Spoke evolution is illustrated in Figure 5.1. The edges of some narrow spokes show deviations in the direction of corotation (Grun et al., 1983). Combined with differential rotation, this would lead to the broadening of the spoke into a wedge shape. Tracing the morning ansa spokes back to where they would have been radial shows that they formed at night, but no correlation is seen with either evening or morning terminator. The contrast of a given spoke decreases with age. Some spokes have been tracked all the way around an orbit, but those seen in the morning are thought to be new spokes imprinted on an old spoke region. This is because differential

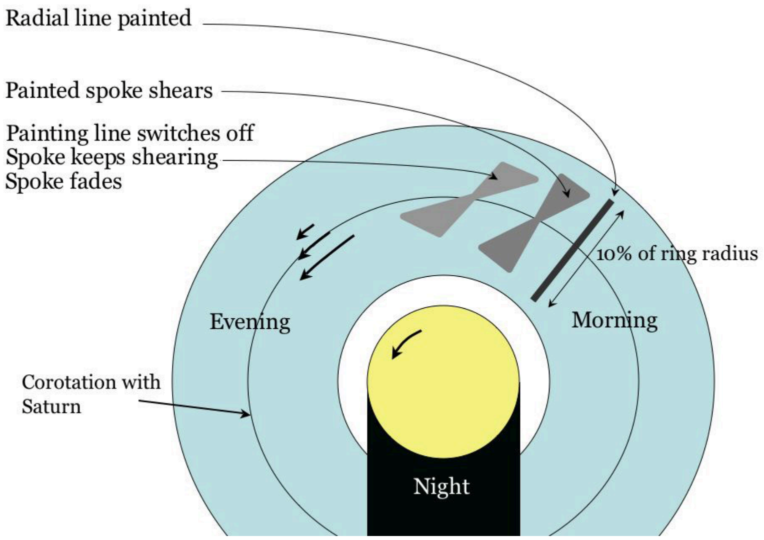

Figure 5.1: A sketch of the evolution of a spoke from the morning to the evening ansa. Spoke length has been exaggerated for clarity. The "painting line" may continue to paint for about an hour, leading to the wedge shape of older spokes.

rotation would shear the spoke by 75° over an orbit, which is much larger than the inclination of any spoke observed.

8. **Spokes may form rapidly.** A few observations have been interpreted as showing the birth of individual spokes within 5 minutes along their entire lengths (Grun et al., 1983). This timescale implies a propagation velocity of at least 20 km s^{-1}. Spokes increase in contrast for their first 20 minutes, after which they start to fade. The radial morphology of spokes suggests that, barring conspiracies, they are formed more rapidly than the differential rotation would shear the spoke by of order its width. This timescale is $\sim$ 1 hour.

9. **Spokes are only present when the rings are close to edge-on to the Sun.** Using *HST*, McGhee et al. (2005) showed that spoke contrast and frequency as seen from Earth decreases as the rings open up to the Sun. No spokes have been seen since the opening angle was $\sim -15.5°$ in 1998. It was unclear whether this was because spokes were no longer present, or whether the viewing angle from Earth (necessarily almost the same as the ring plane opening angle) was not conducive to spoke viewing. Upon arrival at Saturn, *Cassini*'s failure to detect spokes

(Porco et al., 2005a) indicated that spokes are currently truly absent.

10. **Spoke activity is correlated with a magnetic field anomaly.** Porco and Danielson (1982) quantified spoke activity in terms of the number and contrast of spokes seen as a function of time, using data from *Voyagers 1* and *2*. When Fourier transformed, the spoke activity displays a period similar to that of the rotation period of Saturn. Spoke activity was strongest when a particular magnetic longitude was aligned with the morning ansa. This longitude is the location of a suspected anomaly in the otherwise highly symmetric magnetic field of Saturn: when this part of the field faces the Sun, the Saturn Kilometric Radiation — auroral radio emission — is strongest. The connection of these observations with spoke formation is unclear, since the magnetic field lines on which the auroral emission is produced intersect the equatorial plane far outside the rings.

5.2.1 Immediate inferences

The observations suggest that spoke formation involves the sudden lifting of a radial lane of dust grains from the surface of the rings, most often in the morning or just before dawn. The subsequent fading and distortion of spokes is compatible with the elevated dust grains moving on Keplerian orbits that intersect the ring plane half an orbital period (i.e., about 5 hours) later, at which point the dust grains reaffix to the rings.

The low velocity dispersion of ring particles implies that no purely dynamical signal could travel along the length of a spoke fast enough for a feature to be radial. Electrodynamical effects can be much faster. The correlation of spoke activity with the magnetic field, as well as the appearance of spokes only near corotation, indicates an electromagnetic origin. The narrow size spectrum of dust may be due to selection of grains of a given charge to mass ratio.

There is too much dust in a spoke for it to have been swept up from non-spoke regions. The mass of dust is also too large to originate from any reasonably sized impactor. The dust must instead be lifted suddenly from the surface of the rings. The ring surface has the reflectance signature of a "fine grain frost" (e.g., Nicholson et al., 2004), suggestive of a dusty regolith that could be lifted.

5.3 Saturnian system and rings

5.3.1 Saturn

Orbiting the Sun at 9.5 AU, Saturn is the second largest planet in the solar system, with a mass of $M_S = 6 \times 10^{29}$ g, 100 times the mass of Earth and 1/3000 times the mass of the Sun. Its orbital period is 29.5 years, its orbital eccentricity is 0.06, and it rotates rapidly on its axis with a period of 10 h 39 min, i.e., $\Omega_S \simeq 1.6 \times 10^{-4}$ s^{-1}. Its radius $R_S \simeq$ 60,330 km. The spin axis is inclined at 26.7° to its orbital plane, meaning that there are seasons on Saturn.

Saturn's magnetic field is axisymmetric (aligned along spin axis) to within the limits of detection (about 1%); however the field is not dipolar; the magnetic equator is offset $\sim 0.1 R_S$ to the north of the geographic equator of the planet. The magnetic dipole moment is $m \sim 4 \times 10^{28}$ G cm^3, so the surface field $B(R_s) \sim 0.2$ G. Balancing the pressure of the magnetic field with that of the impinging solar wind places the magnetopause at $r \sim 12 R_S$. Saturn's rings, therefore, are shielded from the solar wind and we do not consider its influence.

Saturn had been visited three times before the arrival of *Cassini* on June 30 2004. *Pioneer 11* flew by on September 1, 1979, followed by *Voyager 1* on November 12, 1980, and *Voyager 2* on August 25, 1981.

5.3.2 Rings

Saturn's rings extend from about 0.1 R_S outside its surface to at least 7 R_S. The "main rings" comprise the A, B, and C rings, between 1.24 and 2.27 R_S from the center of the planet. The highest optical depths are found in the B ring, where $\tau \sim 1.8$.

The rings are composed of icy boulders of sizes 1 cm – 10 m in Keplerian orbits in the equatorial plane of Saturn. The rings and their distances from Saturn are listed in Table 5.1. Also listed in the table is the corotation radius at 1.86 R_S. This, as the location of the spokes, is an important place in this study. The radius of corotation is the distance at which the Keplerian angular velocity matches the rotational angular velocity of Saturn. Some relevant ring properties near corotation are

Ring or Gap	Width (km)	Edges (in R_S from center of Saturn)
D		1.11–1.24
C		1.24–1.45
Maxwell Gap	253	1.45 (center)
B		1.52–1.95
Corotation		1.86
Cassini Division	4540	1.99 (center)
A		2.02–2.27
Encke Gap	328	2.14 (center)
Keeler Gap	31	2.26 (center)
F	50	2.33
G		2.8 (center)
E		3–8

Table 5.1: The rings of Saturn and their gaps, adapted from Esposito et al. (1984)

given in Table 5.2.

The main rings are remarkably free of dust (except in spoke regions) (Doyle et al., 1989). Due to dissipative collisions, ring particles are thought to be arranged in a layer no more than $h \sim 10$ m thick, corresponding to a velocity dispersion ~ 1 mm s^{-1}.

The surface mass density of the rings is $\Sigma \sim 10^2$ g cm^{-2}, measured from observations of density waves. The vertical gravity field corresponding to this density is $g \sim 4 \times 10^{-5}$ cm s^{-2}, while the tidal gravity field from Saturn is $g \sim \Omega_K^2 z$, where z is measured from the center of the ring plane. Around corotation, $\Omega_K \sim 1.6 \times 10^{-4}$ s^{-1}, and so above $z \sim 1.6 \times 10^3$ cm, the tidal field dominates. Since the ring thickness is of order $h \sim 10^3$ cm, the gravity field of the rings essentially does not dominate anywhere above the surface of the rings. The vertical gravitational acceleration above the ring surface will not fall below $g_{\min} \sim 4 \times 10^{-5}$ cm s^{-2}.

The nominal equilibrium temperature of a surface of albedo A inclined at angle θ to incoming radiation at a distance D from the Sun is[1]

$$T = T_\odot \left(\frac{R_\odot}{D} \right)^{1/2} [(1-A)\sin\theta]^{1/4}. \tag{5.1}$$

For Saturn's rings, $A \sim 0.5$, $D \sim 9.5$ AU, $\theta \sim 15°$, so $T \sim 76$ K. The measured temperature of the rings (unlit side) is currently $\sim$ 70–90 K (Spilker et al., 2004), depending on location.

[1] We neglect reflected sunlight and IR radiation emitted from Saturn

Property at corotation	Value
Distance from center of Saturn	1.86 R_S (outer B ring)
Magnetic field strength	0.03 G
Latitude of magnetic connection on planet	43°
Keplerian angular velocity	1.6×10^{-4} s^{-1}
Length of night	2 hours
Ring optical depth	1.8

Table 5.2: Properties at the corotation radius in the rings

The opening angle of the rings to the Sun varies over the orbital cycle of Saturn. The maximum ring plane opening angle is 26.7°. Every 15 years, the rings are edge-on to the Sun. This happened most recently in 1995, and before that in March 1980, shortly before *Voyager 1* flew by. Between 1980 and 1995, the northern side of the rings was illuminated. Currently Saturn is in northern winter, so the southern side of the rings is illuminated. Spokes have been seen during both northern summer and winter,[2] suggesting that the magnetic equator's offset to the north does not prevent spokes being formed on either side of the rings.

The spoke contrast decreased gradually after the ring plane crossing of 1995, and no spokes have been detected since 18 Oct 1998, when the ring plane opening angle (to the Sun) was -15.47°, and increasing (McGhee et al., 2005). We provide a possible understanding of this in §5.6.2. If spokes are only visible when the ring plane is within 15.47° of the ecliptic, then spokes should be visible for two intervals of 6 years each, during the Saturn orbital cycle, while the rings are close to edge-on. There are two periods of 9 years between times when the spokes are not seen, and we are currently in one of these. The spokes should reappear in October 2007, at which time *Cassini* should still be operational.

5.4 Photoelectron layer

The major difference between day and night on the rings is the presence of photoelectrons, knocked from the ring particles by solar photons during the day.

We start by considering the case of a vacuum outside the ring particles, and will discuss in §5.7.3

[2] Though only at the beginning and end of these seasons

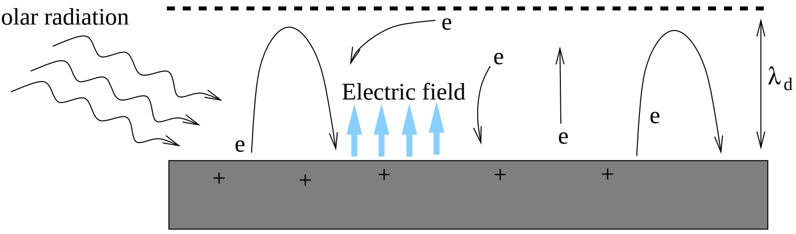

Figure 5.2: The Debye sheath with vacuum outside (i.e., when no ring plasma is present)

the effects of magnetospheric plasma.

When solar photons strike the rings, electrons are ejected with mean energy $\epsilon \sim 1$ eV, leaving the rings charged positive. An ejected electron will travel a distance λ_d from the ring plane until the electric field E of the rings turns it around and returns it to the surface (see Figure 5.2):

$$eE\lambda_d \sim \epsilon. \tag{5.2}$$

The distance λ_d is known as the Debye length. Outside the layer of electrons of thickness λ_d, the electric field is (almost) zero.[3] Within the layer, the electric field is

$$E \sim en_e\lambda_d, \tag{5.3}$$

where n_e is the number density of electrons. The ring surface charge density is equal and opposite to that in the Debye sheath, and the voltage drop across the layer is $\epsilon/e \sim 1$ V. To obtain the steady state electron density above the rings, we balance their photoejection and reabsorption rates

$$yS_\gamma \sin\theta \sim n_e v_e, \tag{5.4}$$

where $S_\gamma \sim 10^8$ cm^{-2} s^{-1} is the solar photoionizing flux at Saturn and $y \sim 0.1$ is the efficiency of ejection (Achterberg et al., 1983). The angle θ is the opening angle of the rings to the Sun.

[3] The electric field is not zero above height λ_d because the electrons have a distribution of energies, as opposed to the monoenergetic distribution assumed here. Some energetic electrons make it further from the ring than λ_d before turning back.

Putting eqs. 5.2, 5.3 and 5.4 together, we obtain

$$\lambda_d \sim \frac{3\times 10^3}{\sqrt{\sin\theta}}\ \mathrm{cm},\ \ \mathrm{n_e} \sim 1\sin\theta\ \mathrm{cm}^{-3},\ \ \mathrm{E} \sim 2\times 10^{-6}\sqrt{\sin\theta}\ \mathrm{statvolt\ cm}^{-1}. \tag{5.5}$$

An electron takes $t_b \sim \lambda_d/v_e \sim 10^{-4}$ s to return to the ring after being ejected, and this is the timescale on which the layer responds to changes in irradiation. As night falls, the Debye layer collapses back on to the ring, and no more photoejections occur until the morning.

5.4.1 Distribution of surface charge

In the above, we calculated the average electric field at the surface of the rings. However, in reality the charge is not evenly distributed across this insulating surface, both because of Poisson variations in the arrival times of photons (a small effect) and because of the redistribution of electrons over the surface.

Electrons ejected from the surface begin to gyrate in the magnetic field, with gyroradius $r_L \sim v_e/\Omega_g \sim 1\times 10^2$ cm. An electron performs about 10 gyrations before falling to the ring plane. Initial directions and gyrophases are random, so the electron returns to the ring within a radius r_L from its ejection point. On scales larger than r_L, the ring surface charge density will be uniform, because all of the electrons that are ejected from a region of area $\gtrsim r_L^2$ are returned to it within a time t_b.[4] On scales smaller than r_L, the charge on a patch of size a can build up until either electrons are deflected from hitting it ($Z = -Z_{\max}$), or electrons are deflected into it ($Z = +Z_{\max}$).

To determine $Z_{\max}$, consider an electron raining down with kinetic energy ϵ towards a patch of size a on the side. Within a distance $z \sim a$, the field from the patch will dominate, and will deflect the electron by a distance

$$x \sim \frac{e^2 Z(a)}{a^2 m_e}\left(\frac{a}{v}\right)^2. \tag{5.6}$$

[4] Though there can be fluctuations on timescales smaller than this, at the level of the Poisson fluctuations in photon arrival times.

This gives a deflection of size a, i.e. the electron is deflected in or out of the patch, when

$$Z \sim Z_{\rm max} \sim m_e v^2 a/e^2 \sim \epsilon a/e^2 \sim 690 a_\mu, \tag{5.7}$$

where a_μ is the patch size in microns. The charge on a particular region performs a biased random walk. In this case, the rms charge on a patch is $Z_{\rm max}^{1/2} \sim 30 a_\mu^{1/2}$. The rms surface charge density then scales as $\sigma_{\rm rms} \propto a^{-3/2}$. The electric field at height z is determined by the surface charge density in the patch of size $\sim z$ beneath it. The ratio of the rms electric field to the mean electric field is then

$$\frac{E_{\rm rms}(z)}{E_{\rm mean}} \sim \left(\frac{z}{1\ {\rm cm}}\right)^{-3/2} \sin\theta^{-1/2}. \tag{5.8}$$

Close to the rings, the electric field can thus be much larger than, and of opposite sign to $E_{\rm mean}$. On scales $a \gtrsim 1(\sin\theta)^{-1/3}$ cm, the rms field variations are smaller than the mean field, and the field will always point away from the rings. The charge on a patch smaller than this will change sign on a timescale $Z_{\rm rms}^2/(a^2 y S_\gamma \sin\theta)$.

5.5 Dust around the rings

5.5.1 Dust from impacts

A spacecraft traveling through the solar system suffers collisions with micrometeoroids, i.e. dust particles of sizes $\sim$ 1–100 micron. Over the duration of the *Pioneer* mission, 80 such impacts were measured; this is still the standard data for estimating the population of micrometeoroids in the outer solar system. The nominal rate of impact on the rings cited by Cuzzi et al. (2002) is $\dot{m} = 5 \times 10^{-17}$ g cm^{-2} s^{-1} (unfocused, one-sided), though this number is uncertain to at least an order of magnitude. Impact velocities must be above a minimum of 20 km s^{-1} (the escape velocity from the ring region) but are expected to be 30–40 km s^{-1}.

We can estimate how much dust these impacts knock off the rings. Experimental investigations of impacts into ice and silicate-ice mixtures have measured this (Koschny & Grün 2001a,b), and these data are the basis for our estimates.

When an impact knocks a dust grain off the ring, the grain is inserted on an inclined Keplerian orbit around Saturn (neglecting any electromagnetic forces). A body on such an orbit crosses the ring plane twice per orbital period, so that after around 5 hours the grain will strike the ring plane, and probably stick to its icy surface.

Overlapping dust clouds: First we investigate whether we expect dust clouds to be lifted all over the rings at once. If we take the micrometeoroid flux to be composed of 10-micron particles, then the arrival rate is 5×10^{-6} cm^{-2} s^{-1}. The minimum velocity of ice particles leaving an impact site was found by Onose (1996) to ~ 10 cm s^{-1}. The extent of an impact cloud will then be at least 10 cm s^{-1} × 5 hours ~ 2 km. This is similar to the scale height of a cloud, $h \sim v/\Omega_K \sim 0.6$ km. Each patch of size 1 km^2 receives 2×10^5 such impacts over the lifetime of a cloud, meaning that the clouds overlap and so we can treat the whole surface of the rings as being impacted at the mean rate.

Mean dust density above the rings: We adopt a crater yield of 1000 (Koschny and Grün, 2001b), so an impactor liberates 1000 times its own mass in dust grains from the surface.[5] The size spectrum of dust produced is $dM/da \propto a^0$ (Koschny and Grün, 2001a) so that the column number density of particles present above the rings is

$$N \sim \frac{10^3 \dot{m}}{\Omega_K a^3 \rho} \frac{a}{a_{\max}}, \tag{5.9}$$

in a logarithmic range around a. The largest liberated body has mass about 1% of the total ejecta mass (Koschny and Grün, 2001a), which for a 10 micron impactor is about 20 micron.

Thus we expect at all times an optical depth of

$$\tau(a) \sim Na^2 \sim 10^{-7} \tag{5.10}$$

[5] At this rate, the entire mass of the rings will be pulverized in ~ 70 million years.

in dust grains in each size range, independent of size. This is compatible with the non-detection of dust grains in non-spoke regions of the B ring. The mass in micron-sized dust above the whole ring system is then $m_{\rm dust} \sim 10^9$ g, just 1 % of the mass of a single spoke.

The number density of dust grains as a function of size is

$$n_g \sim \frac{N}{h} \sim \frac{2 \times 10^{-4}}{a_\mu^2} \text{ cm}^{-3}. \tag{5.11}$$

5.6 Dust charging

Once dust grains are liberated from the rings via impacts, they are exposed to solar UV photons and to plasma above the rings. Dusty plasma physics, the study of this type of situation, has become an experimental and theoretical industry, sparked by the 1980 discovery of the spokes (e.g. Shukla and Mamun, 2002). Here we will cover the basics of dust charging as relevant to the ring environment, extending the treatment of Achterberg et al. (1983).

We assume in the following that grains are uncharged when they are liberated.[6] If their motions are dominated by gravity, then a grain has 5 hours in which to charge and respond to magnetic fields before it reaffixes to the ring surface. The charging process is illustrated in Figure 5.3.

Grain charging proceeds differently according to the plasma environment and the local density of other dust grains. We begin by treating the case of dust in a vacuum, then move on to charging within the Debye sheath and dust levitation, and finally we discuss dust charging in the magnetospheric plasma.

Analogous to the "patches" on the rings discussed in §5.4.1, dust grains can charge either positive or negative up to the point at which they cannot accept any more electrons, or cannot lose any more, given the photon energies available. This sets an upper bound of $\pm Z_{\rm max} = a\epsilon/e^2$ to the charge on a grain of size a, corresponding to the potential energy of an electron at the surface of the grain being equal to its kinetic energy.

[6] This may be a very bad assumption, since impacts can probably impart large charges to dust grains.

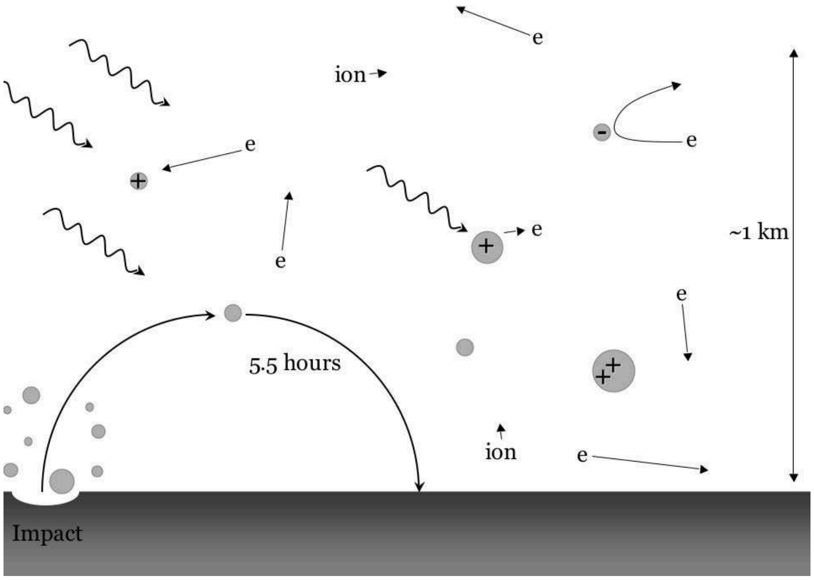

Figure 5.3: Charging of dust grains in a plasma layer above the rings, in the presence of a photon field.

5.6.1 Dust charging in vacuum

An isolated grain in an ionizing photon field charges positive to

$$Z(t) = \min(Z_{\max}, yS_\gamma a^2 t), \tag{5.12}$$

where t is the length of time the grain spends in the photon field.

For a dust grain number density n_g, the density of ejected electrons $n_e = n_g Z$ is bounded by the value at which the flux of electrons on to a grain matches the photoejection rate.[7]

$$4n_e a^2 v_e = yS_\gamma a^2 \tag{5.13}$$

The electron density cannot rise above this equilibrium level. The mean charge on a dust grain is then $\bar{Z} = n_e/n_g$, which is positive but may be small. The charge on an individual grain will fluctuate; it performs a biased random walk due to the gain and loss of electrons. Similar to the

[7] Although we are neglecting factors of order π throughout this work, we retain the factor of 4 here to account for the larger cross-section to electron impacts, which can come from any direction, relative to that for photons, which are coming from one direction only. This turns out to be important (see §5.6.2).

discussion of the ring "patches," we find

$$Z_{\rm rms}(t) = \min\left(\sqrt{yS_\gamma a^2 t}, Z_{\rm max}^{1/2}\right). \tag{5.14}$$

Under the solar flux at the rings, grains smaller than $a = 0.02$ micron do not reach $Z_{\rm rms} = Z_{\rm max}^{1/2}$ in 5 hours. Grains smaller than $a = 0.001$ micron have $Z_{\rm max} < 1$, and charging for these grains must be considered probabilistically (not all grains will have charges).

5.6.2 Dust charging within the Debye sheath

If the dust is within the Debye sheath, then provided $\tau_{\rm dust} \lesssim \sin\theta$, the electron density is determined by the rings, and is set by the photoejection rate, which depends on the solar inclination angle:

$$n_e = \frac{yS_\gamma \sin\theta}{v_e}. \tag{5.15}$$

A dust grain receives electrons at a rate $4a^2 n_e v_e = 4a^2 yS_\gamma \sin\theta$, but loses them at a rate $a^2 yS_\gamma$, because being roughly spherical it receives the direct solar flux. For a neutral grain these rates are unmatched. To reach a steady state the grain becomes charged, so that the rates of loss and gain become equal. For $4\sin\theta > 1$, i.e. $\theta > 15°$, electron collisions dominate over photoejection and the grains are negatively charged in steady state. For lower ring opening angles, the grains will charge positive. The rings, meanwhile, are always charged positive. Therefore for $\theta < 15°$, grains are electrostatically repelled from the ring surface, and for $\theta > 15°$ they are attracted to it. This was noticed by Nitter et al. (1998). The changeover in physical behavior occurs at the angle at which the spokes disappeared: spokes have not been seen when $\theta \gtrsim 15°$ (McGhee et al., 2005). Since the spokes are probably made of levitated dust, it is possible that this piece of physics plays a part in their appearance. The presence of plasma above the ring plane may destroy this effect (see §5.7.3).

5.6.3 Levitation of dust

This brings us to the subject of dust levitation. A dust particle can be levitated if its electrostatic repulsion from the rings is stronger than the gravitational attraction. Since the spoke particles are thought to be made of a narrow size distribution of charged dust, sorting by size must be important, and the levitation condition has a strong dependence on grain mass. Dust levitation has been studied theoretically (e.g., Nitter et al., 1998) and experimentally, and observed on the moon.

We illustrate this phenomenon for the case in which the solar inclination angle is small enough that the grains and ring are charged positive. We assume that the dust is charged to $Z = Z_{\rm max}$. The electric field from the rings is of order $E \sim 1 \times 10^{-6}$ statvolt cm^{-1}. At the ring surface, $g \sim 4 \times 10^{-5}$ cm s^{-2}, and so the ratio of electrostatic to gravitational forces on a grain of size a is

$$\frac{F_E}{F_G} \sim \frac{4 \times 10^{-13} a_\mu}{4 \times 10^{-17} a\mu^3} \sim \frac{8 \times 10^3}{a_\mu^2}. \tag{5.16}$$

For grains smaller than $a \sim 90\mu$, $F_E > F_G$ and the grain can levitate. Because $F_G \propto z^2$, and F_E decreases with height above the ring plane, there will be an equilibrium height at which the forces balance. Making the approximation $E(z) =$ constant inside the Debye sheath and zero outside it, grains of $a < 14\mu$ will be levitated to the top of the sheath.

This picture is complicated by a number of factors. First, the electric field and electron density change gradually with z, so the rate of grain charging is a function of z. This can lead to damped oscillations of grains around their equilibrium heights in the Debye sheath (Nitter et al., 1998), so that a grain ejected from the ring plane need not end up back there.

Sticking: A more serious complication is the sticking forces that have to be overcome in lifting a dust grain from the ring surface. Two micron-sized ice spheres adhere via van der Waals forces, and require $F_{\rm stick} \sim 10^{-2}$ dyne to separate them (Blum, 2004). This is to be compared with the repulsive force on a charged 1 micron grain at the ring surface, $F_E \sim 4 \times 10^{-13}$ dyne. When we include the concentration of surface charges predicted in §5.4.1, this increases to 4×10^{-7} dyne, still much less than $F_{\rm stick}$. Another way to achieve high surface electric fields is through "fairy castles"

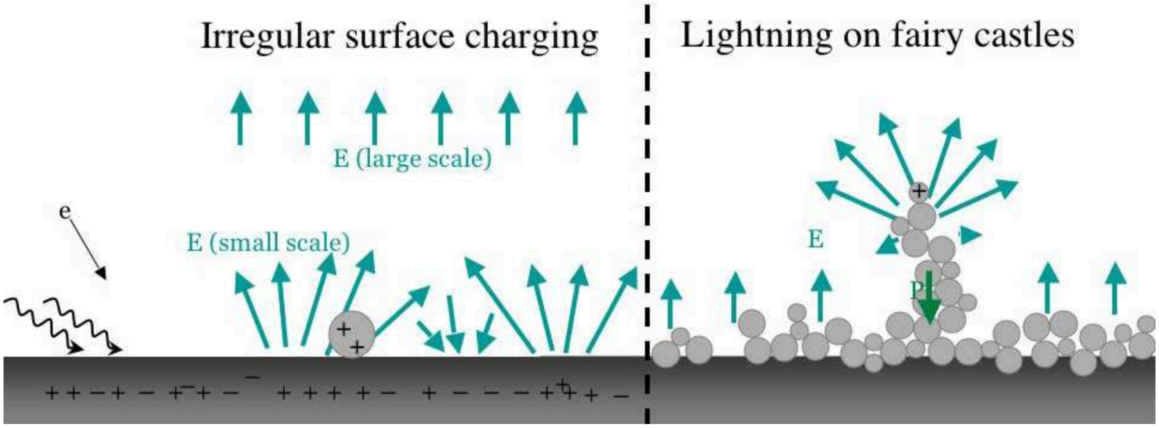

Figure 5.4: Possible concentrations of electric field at the surface of the rings

on the ring surface, where grains form underdense irregular shapes after impacts. Electric fields are concentrated at the "spires" of these castles (see Figure 5.4).

Sticking may be overcome by constant "stirring" of the surface by micrometeoroid impacts, but it remains difficult to understand how a large number of grains could be lifted from the surface with purely electrostatic forces, given the predicted fields.

Maximum charge levitated: There is a limit to the amount of charge that can be levitated by the rings. Because the levitated charges are within a Debye length of the rings, their own electric field is not shielded from the ring plane. If the surface charge density of levitated particles exceeds that on the rings, then the direction of the electric field in the Debye sheath is reversed, and some of the charges fall back to the ring plane. Assuming that grains are charged to $Z = Z_{\rm max}$, and that $\theta \sim 15°$, this limit corresponds to an optical depth

$$\tau_{\rm max} \sim \frac{1 \times 10^{-8}}{a_\mu} \tag{5.17}$$

in grains in a logarithmic range around a.

5.6.4 Dust charging in a plasma

Dust charging in plasma works on the same principle as charging in the Debye sheath (§5.6.2). If the dust is in a plasma without photoejection, it charges negative up to a maximum of $Z \simeq -Z_{\rm max}$, so long as this does not use up all of the electrons in the plasma cloud. If there is also a photoelectron flux, then the sign of charge on the grains is determined by whether electron loss or gain is faster on a neutral grain. In a high electron density environment, a dust grain charges negative until the rates of photoejection and electron absorption balance, again limited by $Z = -Z_{\rm max}$.

5.7 The ring atmosphere and ionosphere

One of the most intriguing discoveries made by *Cassini* at Saturn is the oxygen ionosphere near the ring plane, inferred to be formed by photoionization of a ring atmosphere of molecular oxygen. A ring ionosphere has the potential to significantly affect the electrodynamics of the ring system, and its discovery may be the key to understanding the spoke phenomenon. In the following we piece together the observations to place limits on how much oxygen can be present over the rings, and we use these results to develop a model of the atmosphere/ionosphere. We find the intriguing possibility that the atmosphere closest to corotation is the coldest, and so there could be instabilities associated with carrying field-aligned currents in the rarefied high-altitude atmosphere at this location.

Oxygen ions were detected over the ring plane during the Saturn Orbital Insertion phase (see Figure 5.5). Oxygen atmospheres are observed around other icy bodies; Ganymede and Europa are both inferred to have oxygen atmospheres (e.g., Hall et al., 1998; Yung & McElroy, 1977). There is evidence for the presence of molecular oxygen trapped in the surface lattice of Ganymede (Spencer et al., 1995).

When irradiated by UV, ice is decomposed into an array of hydrogen- and oxygen-bearing species, which can be liberated from the surface of the rings due to finite temperature or impacts (Johnson et al., 2004). The important difference with molecular oxygen (and H_2) is that these homopolar diatomic molecules have very low polarity, and so when they strike the icy surface, they tend to

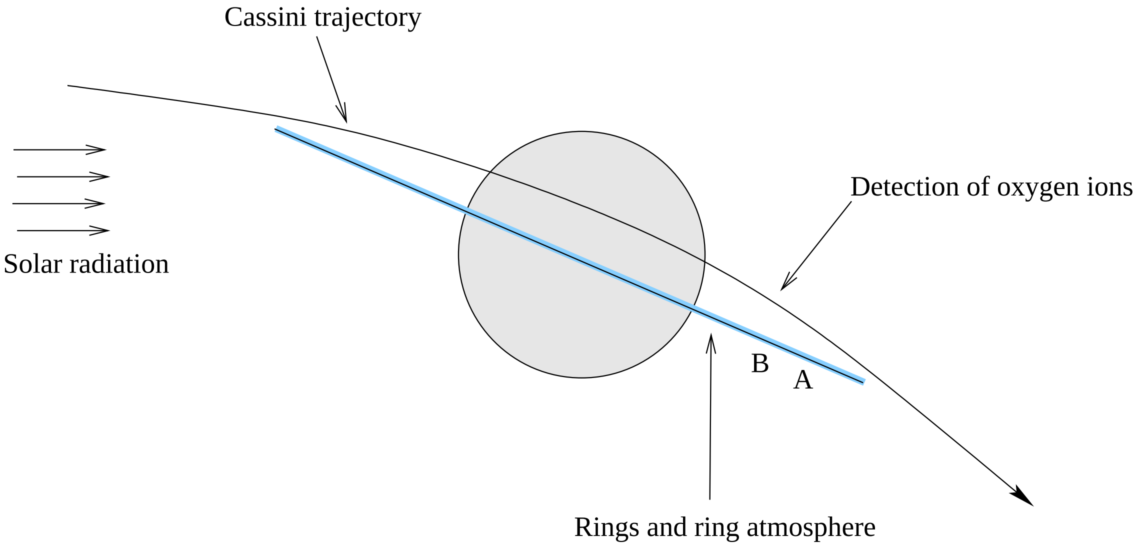

Figure 5.5: A sketch of the trajectory of *Cassini* during the Saturn Orbital Insertion phase. The observations of oxygen ions were made many thermal scale heights above the rings.

bounce and not stick. Hydroxyl species, as well as atomic oxygen and hydrogen, tend to adhere to the surface. In this way, the O_2 and H_2 produced remain as an atmosphere. The hydrogen is thought to be absent because it is light and can escape the rings easily, leaving a molecular oxygen atmosphere in which photoionization can form oxygen ions.

5.7.0.1 The observations

The Cassini Plasma Sensor (CAPS) instrument detected both O^+ and O_2^+ over the A and B rings (Young et al., 2005). Over most of the B and A rings, O_2^+ dominates over O^+ by a factor ~ 4. Between 1.85 and 1.89 R_S (right around corotation), this ratio is reversed. Hydrogen ions (atomic or molecular) comprised less than 3% of the ions detected. The altitude of the spacecraft over the rings varied from 18,000 km over the mid-B ring at the start of the observations, to 10,000 km over the inner A ring. Particle energies were consistent with a temperature of 0.75 eV relative to corotation, i.e., pickup velocities corresponded to kinetic energies of 0.75 eV (2 km s^{-1} for a molecular oxygen ion).

The Ion Neutral Mass Spectrometer (INMS) confirmed the detection of the oxygen ionosphere, with detections of both O^+ and O_2^+ over the A ring at an altitude of 7,000 km (Waite et al., 2005).

Their results are consistent with corotating ions with gyrospeeds of 5 km s^{-1} or less. Inferred densities are between 0.1 and 1 cm^{-3} for O_2^+ and H^+, and from 0.03 to 0.3 cm^{-3} for O^+.[8] INMS also measures neutral particle densities. However, while Cassini was over the main rings at 7,000 km altitude, the neutral particle signal was dominated by noise associated with opening the shutter of the instrument, and only an upper limit of 10^5 cm^{-3} could be measured.

A third instrument, the Radio and Plasma Wave Sensor (RPWS) measured electron densities over the rings. From the presence of auroral hiss emission, a low-energy electron beam was inferred to be coming off the rings near the corotation radius (Gurnett et al., 2005). A deep minimum in electron density was seen at the same place, $1.7 \pm 0.1 R_S$, where $n_e \sim 3 \times 10^{-2}$ cm^{-3}. The electron density is higher, $n_e \sim 1 - 10$ cm^{-3}, over the A ring.

Using these observations, we model the properties of the oxygen atmosphere and ionosphere.

5.7.1 Conservative interpretation

The photoproduction rate of molecular oxygen in the icy surface of the rings is $S \sim 10^6$ cm^{-2} s^{-1} (Johnson et al., 2004).[9] We assume that all the molecular oxygen produced is liberated from the ring surface (see vapor pressure, below for justification), and that once liberated it does not stick to the surface again. We also assume that the atmosphere is collisionally thin, so that oxygen molecules collide more frequently with the ring than with other molecules, i.e.,

$$\tau \sim N_{\mathrm{O}_2} \sigma_n < 1, \qquad (5.18)$$

where N_{O_2} is the column density of molecular oxygen, and $\sigma_n \sim 10^{-15}$ cm^2 is the neutral collision cross section.

The oxygen atmosphere will be at the same temperature as the rings, 70–90 K. The lowest atmospheric scale height is then $h \sim v_{\mathrm{th}}/\Omega_S \sim 1000$ km near corotation. The atmosphere has a

[8] It is not clear why INMS detected so much hydrogen when CAPS detected so little.

[9] This number may be poorly constrained and will be investigated further outside of this thesis.

Gaussian density profile,

$$n = n_0 \exp[-(z/1000\text{km})^2]. \tag{5.19}$$

The lowest scale height is comparable to the azimuthal extent of a spoke, but whether this is significant is not known.

Atmospheric oxygen molecules are subject to photoionization, photodissociation, and charge exchange. We assume that when the resulting species strike the rings, they are instantly reformed into oxygen molecules, which are again liberated. Thus the only way to lose molecular oxygen is through its diffusion off the rings. Each time a molecule bounces, it makes a 1000 km horizontal jump in a random direction due to its thermal velocity. If the molecule bounces off the rings, it is lost from the atmosphere. To random walk $L \sim 10^{10}$ cm, 10^4 bounces are required, taking a time $t_{\text{loss}} \sim 2 \times 10^8$ s.

Equating production and loss rates, we obtain[10]

$$n_{\text{O}_2} \sim \frac{S t_{\text{loss}}}{h} \sim 2 \times 10^6 \text{ cm}^{-3}, \tag{5.20}$$

We also estimate the ion density, by equating the rate of photoionization of O_2 with the loss rate of ions, assuming that ions stick to the rings when they strike them. Thus

$$n_i \sim S_\gamma n_{\text{O}_2} \sigma_{\text{pi}} t_{\text{bounce}} \sim 4 \text{ cm}^{-3}, \tag{5.21}$$

where $\sigma_{\text{pi}} \sim 10^{-18}$ cm^2 is the photoionization cross section, and we use the same $S_\gamma \sim 10^8$ cm^{-2} s^{-1} as in §5.4. We next assess whether these predictions agree quantitatively with the number densities of ions observed. The problem is that the *in situ* observations were made at $z \sim 7,000$–$18,000$ km. The density of thermal ions at these altitudes is tiny, using the Gaussian profile (Eq. 5.19) and the low-altitude ion density (Eq. 5.21). Johnson et al. (2004) resolve this by conjecturing that the observed ions are non-thermal, as follows. Most photoionizations (75%) produce cold O_2^+ and an electron, and about 25% produce O, O$^+$ and e. The O$^+$ is "hot," with energy ~ 1 eV and hence a

[10] The collisional thickness of this atmosphere is $\tau \sim n_{\text{O}_2} h \sigma_n \sim 0.2$, justifying our collisionless bouncing assumption.

scale height $\sim 13,000$ km. This explains the presence of O^+ at high altitudes. Johnson then says that the observed O_2^+ must also be non-thermal, heated by collisions in the lower atmosphere.

Our interpretation of this heating is the following. After an O_2^+ ion forms in a Keplerian orbit, it is picked up by the magnetic field, about which it gyrates with a (horizontal) velocity v_p equal to the difference between corotational and Keplerian velocities. More than ∼800 km from corotation, $v_p > v_{\rm th}$. A collision that transfers some of this horizontal velocity into the vertical direction can thus "heat" an ion and increase its scale height.[11] Over the A ring, $v_p \sim 5$ km s^{-1} so the scale height of a heated O_2^+ ion is ∼30,000 km. The fraction of ions heated before they stick is

$$f \sim \frac{n_{\rm O_2}\sigma_n v_i}{\Omega_K} \sim \tau \frac{v_i}{v_{\rm th}} \sim 5 \tag{5.22}$$

over the A ring, i.e., every ion is heated. Thus we also expect to observe O_2^+ at high altitude. Where heating occurs, the density of ions near the ring plane is reduced from that given by Eq. 5.21. Near corotation $v_p < v_{\rm th}$ and so collisions do not result in heating. This explains why O^+ ions, which do not require heating to reach high z, were dominant when *Cassini* was over corotation.

Collisional heating also leads to increased neutral loss rates, because both neutral and ion species leave the collision "hot." A hot neutral has a longer bounce step and leaves the rings faster. Although every ion experiences a collision with a neutral, only a small fraction of neutrals collide with an ion in a given bounce. The ion-neutral collision rate is $\nu_c \sim 8 \times 10^{-11} \max[1, v_p/v_{\rm th}]$ s^{-1}, somewhat slower than the rate of loss by thermal bouncing calculated above. We do not consider it further.

Vapor pressure: The pressure of the conservative atmosphere is $p_0 \sim 2 \times 10^{-8}$ dyne cm^{-2}, which corresponds to 2×10^{-11} mm Hg. The vapor pressure of molecular oxygen at 50 K is 1 mm Hg.[12] This vapor pressure is strictly for oxygen over an oxygen layer, and over an ice layer the pressure may be lower because of increased attractive forces with polarized water molecules. If this is the case, the formation of a monolayer of molecular oxygen on the surface would return the vapor pressure to the quoted value. We do not expect any more than a monolayer of O_2 to be retained on the surface

[11] This process can also be thought of as Ohmic dissipation of the radial current flowing in the atmosphere.

[12] from the CRC Press Handbook of Chemistry and Physics

of the rings.

5.7.2 Liberal interpretation

Instead of calculating the atmospheric density based on a collisionally thin atmosphere, we ask how much oxygen could be above the rings without contradicting observations. A denser atmosphere/ionosphere may have more influence on the electrodynamics of the rings. We adopt a collisionally thick atmosphere, requiring $\tau > 1$, i.e., $N_{\mathrm{O}_2} > 1/\sigma_n \sim 1 \times 10^{15}$ cm^{-2}.

5.7.2.1 Neutral loss rate

In the collisional atmosphere, molecules diffuse away from their source at a rate

$$\text{Flux} \ \sim \lambda \mathrm{v}_{\mathrm{th}} \nabla \mathrm{n}_{\mathrm{O}_2}, \tag{5.23}$$

where λ is the mean free path. We assume that the number density changes over a distance of order the ring radius $R \sim 10^{10}$ cm. The rings as a whole then lose oxygen molecules at a rate

$$\frac{1}{t_{\mathrm{loss}}} \sim \frac{\lambda v_{\mathrm{th}} n_{\mathrm{O}_2}}{R} = \frac{v_{\mathrm{th}}}{\sigma_n R} \sim 2 \times 10^{27} \ \mathrm{s}^{-1}. \tag{5.24}$$

The loss rate is thus *independent* of the atmospheric density, for collisionally thick atmospheres. The production rate of molecular oxygen over the whole rings, $SR^2 \sim 10^{26}$ s^{-1}, is similar to the loss rate (within the uncertainties of our estimates). Therefore the atmospheric density could be unconstrained by the equality of production and loss rates.

Maximum neutral density As an upper limit to the atmospheric density, $n_{\mathrm{O}_2} < 10^5$ cm^{-3} was measured at 7,000 km above the rings (Waite et al., 2005). This corresponds to a ring plane density $n_{\mathrm{O}_2} < 4 \times 10^{13}$ cm^{-3} and a column density $N_{\mathrm{O}_2} \sim 4 \times 10^{21}$ cm^{-2}.[13] We place more stringent limits on the density in future sections.

[13] which would actually be thick to ionizing photons

5.7.2.2 Loss processes for ions in the collisional atmosphere

In a collisionally thick atmosphere, there are more possible ion loss processes. The available mechanisms depend on the ion species.

As mentioned in §5.7.1, 75% of ionizing photons on O_2 produce O_2^+, while the rest produce O and O^+ (Johnson et al., 2004). Thermalization of the ionized species is rapid in a collisional atmosphere.[14] Radiative recombination rates are always small, but molecular oxygen ions can be removed by dissociative recombination with an electron:

$$\mathrm{O}_2^+ + \mathrm{e} \rightarrow \mathrm{O} + \mathrm{O}. \tag{5.25}$$

The cross section for this process is

$$\sigma_{\mathrm{dr}} \sim \frac{r_0 e^2}{kT_e} \sim \frac{10^{-13}}{(T_e/80\mathrm{K})}\ \mathrm{cm}^2, \tag{5.26}$$

since in order to recombine, the electron must physically collide with the ion (of size $\sim r_0$), to which it is electrostatically attracted. This route is unavailable to O^+ ions, which can however be lost through charge exchange with molecular oxygen: the reaction

$$\mathrm{O}^+ + \mathrm{O}_2 \rightarrow \mathrm{O} + \mathrm{O}_2^+ \tag{5.27}$$

is exothermic, with a cross section of order $\sigma_n \sim 10^{-15}$ cm^2. The resulting molecular oxygen ions are lost to dissociative recombination.

Both species of ion are lost when they strike the ring plane. Only those in the lowest mean free path of the atmosphere can directly collide with the rings. We perform a simplified calculation of the diffusion of ions in the lowest scale height of the atmosphere, assuming a uniform (mostly neutral) atmosphere of density n_{O_2}, height h and mean free path $\lambda = 1/(n_{\mathrm{O}_2}\sigma_n)$. The ion velocity is

[14] This means that scale heights are the same for all species. However, ions produced or heated in the topmost mean free path can still reach large altitudes without suffering further collisions, and so the considerations of observed hot ions are similar to those for the conservative atmosphere.

$v_i = \max[v_{\rm th}, v_p]$, which is valid in the limit in which ions have time to acquire their pickup velocities between collisions, $\nu_c \ll \Omega_g$. The use of this limit is justified at the end of the section. The flux of ions through the neutral atmosphere is

$$F(z) = -\frac{\lambda v_i}{3}\frac{dn_i}{dz} = -D\frac{dn_i}{dz}, \tag{5.28}$$

where D is the diffusion constant. The ion production rate is $\nu n_{\rm O_2}$, and so

$$\frac{dF}{dz} = -D\frac{d^2 n_i}{dz^2} = \nu n_{\rm O_2}. \tag{5.29}$$

We use boundary conditions $F(h) = 0$, because no ions enter from the top of the atmosphere, and

$$F(0) = -\frac{v_i n_i(\lambda)}{3}, \tag{5.30}$$

because only ions in the last mean free path hit the rings. Solving Eq. 5.28 subject to these boundary conditions, we obtain

$$n_i(z) = \frac{3\nu n_{\rm O_2}}{\lambda v_i}\left(hz - \frac{z^2}{2} + \frac{\lambda^2}{2}\right). \tag{5.31}$$

The number density of ions in the lowest scale height of the atmosphere is (Eq. 5.31)

$$n_i \sim \frac{\nu n_{\rm O_2} h^2}{\lambda v_i} \sim \frac{5\tau^2}{\max[1, v_p/v_{\rm th}]}. \tag{5.32}$$

This corresponds to balancing a diffusive loss rate $\sim n_i \Omega_K/\tau$ with a production rate $\nu n_{\rm O_2}$.

Coulomb collisions suffered by ions do not contribute to the diffusion time calculated above. An ambipolar electric field keeps the electrons and ions diffusing through the atmosphere at the same speed. With no net velocity between charged species, Coulomb collisions will have no effect.

Loss rates as a function of density To determine the dominant loss process, we compare the loss rates for each species above. For O_2^+ ions, loss via dissociative recombination becomes important when

$$\sigma_{\mathrm{dr}} n_e v_e \gtrsim \frac{\Omega_K}{\tau}. \tag{5.33}$$

Substituting Eq. 5.32 for the ion density in the diffusive limit, we obtain

$$N_{\mathrm{O}_2} \gtrsim 1 \times 10^{16} \left(\frac{v_i}{5\ \mathrm{km\ s}^{-1}} \right)^{1/3}\ \mathrm{cm}^{-2},\ \text{i.e.,}\ \tau \gtrsim 10. \tag{5.34}$$

Above this critical neutral column density the molecular oxygen ion density is set by the balance of dissociative recombination with ionization:

$$n_i \sim \left(\frac{\nu n_{\mathrm{O}_2}}{\sigma_{\mathrm{dr}} v_e} \right)^{1/2}. \tag{5.35}$$

At lower column densities, the O_2^+ density is given by Eq. 5.32.

For the O^+ ions, loss via charge exchange becomes important when

$$\sigma n_{\mathrm{O}_2} v_i \gtrsim \frac{\Omega_K}{\tau}, \tag{5.36}$$

which is where

$$N_{\mathrm{O}_2} \gtrsim 1 \times 10^{17} \left(\frac{v_i}{5\ \mathrm{km\ s}^{-1}} \right)^{1/3}\ \mathrm{cm}^{-2},\ \text{i.e.,}\ \tau \gtrsim 10^2. \tag{5.37}$$

In §5.8.2.2 we use the observed corotation of the upper ionosphere to limit the atmospheric column density to

$$N_{\mathrm{O}_2} \lesssim 1 \times 10^{16}\ \mathrm{cm}^{-2} \quad \text{liberal atmosphere,} \tag{5.38}$$

which is within the range at which the O^+ ions are still lost by diffusion on to the ring plane, and at the upper end of the range at which the O_2^+ ions can be treated this way. From now on we adopt the diffusive loss rates, though investigations are ongoing.

Heating the atmosphere: In the collisionally thick atmosphere, molecules heated by collisions with pickup ions will share their kinetic energy with other molecules in the atmosphere before striking the ring plane and returning to the ring temperature. The energy dissipation can equivalently be thought of as Joule heating from the current system. The atmosphere loses its excess energy on a timescale

$$t_{\rm diff} \sim \frac{h^2}{\lambda v_n} \sim \frac{\tau}{\Omega_K}. \tag{5.39}$$

At equilibrium, the gain and loss rates of energy must balance:

$$\frac{\Omega_K}{N_{\rm O_2}\sigma}(\epsilon_n - kT) \sim n_i \sigma_n v_i \epsilon_i, \tag{5.40}$$

which is valid so long as $\epsilon_i \gg \epsilon_n$. The equilibrium neutral velocity ($\epsilon_n = m_n v_n^2/2$) is

$$v_n \sim \left(\frac{n_i^{3/2}\sigma_n^{3/2}}{\Omega_K}\right)^{1/2} v_i^{3/2} + v_{\rm th} \sim \left(\frac{N_{\rm O_2}^3 \sigma_n^3 \nu}{\Omega_K}\right)^{1/2} v_i + v_{\rm th}, \tag{5.41}$$

using Eq. 5.32. The approximation $\epsilon_i \gg \epsilon_n$ ceases to be valid for $N_{\rm O_2} \gtrsim 1 \times 10^{17}$ cm^{-2}. For larger column densities the atmosphere is heated to the temperature of the pickup ions, and its scale height is equal to $\max[v_i, v_{\rm th}]/\Omega_K$. For smaller column densities

$$h \sim h_0 + \frac{v_n}{\Omega_K} \sim h_0 + \left(\frac{\nu \sigma_n^3 N_{\rm O_2}^3}{\Omega_K^3}\right)^{1/2} v_i. \tag{5.42}$$

Our liberal atmosphere is in this latter category. The heating is significant, i.e., the atmosphere puffs up, where $h - h_0 \gtrsim h_0 \sim 1000$ km. For the liberal atmosphere, this occurs where

$$v_i \gtrsim 6 \text{ km s}^{-1}, \tag{5.43}$$

corresponding to a distance $\Delta r \sim 25{,}000$ km from corotation. This is interestingly similar to the extent of a spoke.

Corotation was measured over the A ring where the pickup speed is $v_i \sim 5$ km s^{-1}, and so the heating of neutrals is not significant there. However, an atmosphere just 10 times thicker would be at the pickup ion temperature, so the scale height would be 30,000 km.

Increased atmospheric temperatures may lead to increased loss rates, and so, farther than 25,000 km from corotation, neutral column densities may be lower. Neglecting this possible variation in N_{O_2} for now, the ion column density scales with ion velocity as

$$N_i \sim n_i h \propto \quad v_i^{-1} v_i \propto v_i^0 \qquad \text{heated} \tag{5.44}$$

$$v_i^{-1} v_i^0 \propto v_i^{-1} \qquad \text{unheated.} \tag{5.45}$$

Ion-neutral collision frequencies in the atmosphere (per ion) scale as

$$\nu_c \sim n_n \sigma v_i \propto \quad v_i^{-1} v_i \propto v_i^0 \qquad \text{heated} \tag{5.46}$$

$$v_i^0 v_i \propto v_i \qquad \text{unheated.} \tag{5.47}$$

The collision frequency reaches a maximum at the transition between unheated and heated regions of the atmosphere, and remains constant outside this point. For our liberal atmosphere, this maximum collision rate is $\nu_c \sim 0.1$ s^{-1}. The ion gyrofrequency is $\Omega_g \sim 9$ s^{-1}. We are therefore always justified in adopting the low collision limit $\nu_c \ll \Omega_g$.[15]

The transport of angular momentum through the atmosphere via collisions will be the subject of a future study.

5.7.3 Debye sheath in presence of plasma

The presence of plasma changes the charging and discharging rates of the ring plane. In particular, it can cause the rings to become negatively charged. We assume in the following that the plasma electrons have the same temperature as those ejected from the ring plane.

No photoemission case: When a plasma is present, the faster-moving electrons strike the ring surface more frequently than the ions until the surface becomes sufficiently negative to equalize the rates. The surface potential will be of order $eV \sim -kT_e$. The Debye length is set by the plasma

[15] Note that this is a *different* collisionless limit from that determining whether molecules collide more often with each other or with the rings.

density:

$$\lambda_d \sim \left(\frac{\epsilon}{e^2 n_i}\right)^{1/2}. \tag{5.48}$$

With photoemission: In steady state, the rates of charge gain and loss on the rings must balance (using here fluxes in a collisionless atmosphere)

$$v_i n_i + y S_\gamma \sin\theta = n_e v_e. \tag{5.49}$$

To determine whether this requires the rings to be positively or negatively charged, we compare the terms when evaluated for a neutral ring plane. The critical plasma density is[16]

$$n_{\rm crit} \sim y S_\gamma \sin\theta / v_e. \tag{5.50}$$

If $n_i > n_{\rm crit}$ then the electron flux exceeds the photon flux on to a neutral ring plane, so the rings must be negatively charged to satisfy Eq. 5.49. If $n < n_{\rm crit}$, then the rings must be positively charged.

In our conservative oxygen atmosphere, electrons would rain down on a neutral ring plane at a rate $n_e v_e \sim 2 \times 10^8$ cm^{-2} s^{-1}, which is greater than the rate at which they are ejected, $y S_\gamma \sin\theta \sim 3 \times 10^6$ cm^{-2} s^{-1}. The ring plane is therefore negatively charged. The Debye sheath thickness will be set by Eq. 5.48. The rate of plasma loss to the ring plane is set by $n_i v_i \sim 4 \times 10^4$ cm^{-2} s^{-1}.[17] We assumed this implicitly in §5.7.1.

Ion build-up: If the plasma density is less than $n_{\rm crit}$ then the ring plane will be charged positive, so the ions are repelled from the rings: they are levitated. Because the ions are not lost from the plasma, and the electrons are returned to the plasma by photoejection as soon as they hit the ring plane, the plasma density can build up to $n_{\rm crit}$. Provided there is a neutral atmosphere and a photon flux, the plasma density will be at least $n_{\rm crit}$.

[16] The ion flux $n_i v_e$ is usually small compared with the other fluxes on to the neutral ring plane.

[17] So long as the atmosphere is not very collisionally thick to ions.

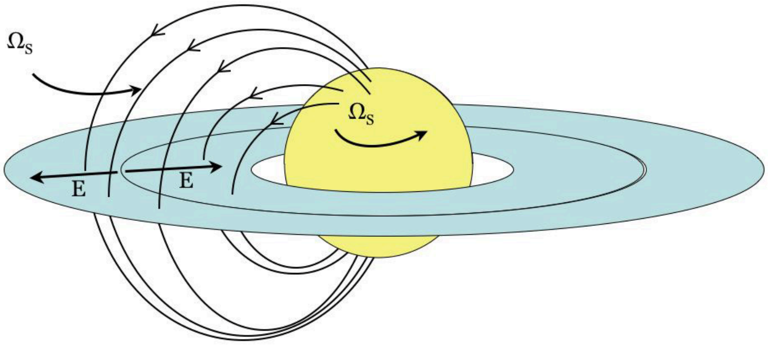

Figure 5.6: The electrodynamics of the Saturn system: the plasma in the magnetosphere corotates with the highly conducting ionosphere of Saturn, resulting in radial electric fields in the rings, pointing away from corotation.

5.8 Electrodynamics and currents

As discussed in §5.2, the spokes are thought to be the product of electric forces, so it is important to understand the electrodynamics of the ring system. In addition, a preference for corotation has only electrodynamical meaning. Radial currents are particularly of interest, because spokes are radial. Azimuthal currents are also interesting because they produce radial magnetic field perturbations. In this section we estimate the magnitude of the currents and electric fields expected near corotation. We consider only azimuthally symmetric conductivities at this stage.

Qualitative discussion: Saturn's magnetic field threads both its ionosphere and the rings. Modeling the field as a dipole, a field line from colatitude λ on the planet intersects the ring plane at a distance $R_S/\sin^2\lambda$ from the center of the planet. The magnetic field lines also thread Saturn's tenuous magnetospheric plasma (see Figure 5.6).

Except at corotation, the rings and ionosphere rotate relative to each other. The radial electric fields will then be different in the two frames, and a current is driven around the circuit joining rings

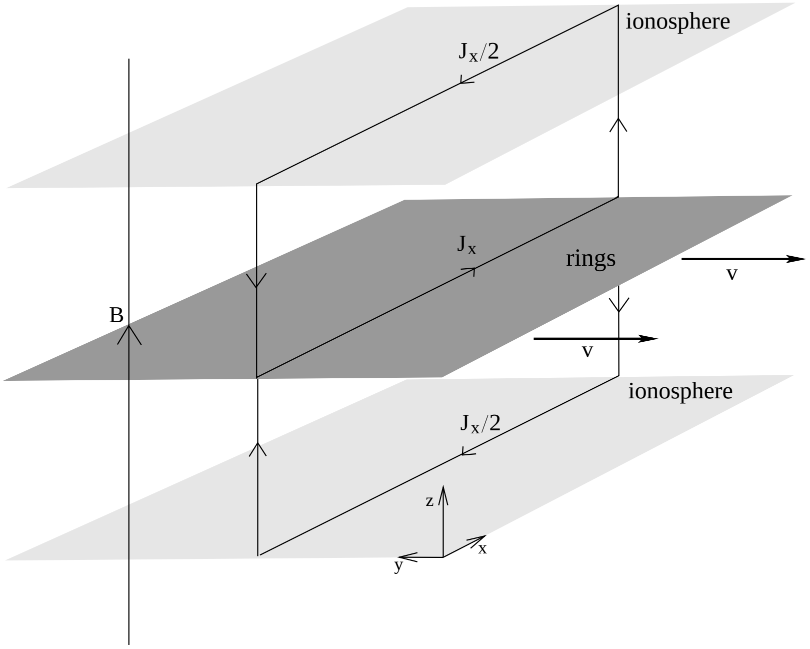

Figure 5.7: The currents in the model rings-ionosphere system

and ionosphere. We assume that the magnetospheric plasma is a perfect conductor along field lines (but see §5.8.3.3). Because the ionosphere is a better conductor than the rings, the driven current requires only a small electric field in the ionosphere frame, and a much larger one in the ring frame. The magnetospheric plasma is at rest in the frame with no electric field, which is close to the frame of the ionosphere, i.e., the magnetosphere "corotates" with Saturn.

At the corotation radius, there is no relative angular motion between the ionosphere and the rings, so the radial electric field is zero in both the ring and ionosphere frames. It is surprising that the spokes occur at such a "quiet" place in the rings. The velocity difference from corotation at the end of a spoke is ~ 2 km s^{-1}, and the potential drop along the 10,000 km length of a spoke is $V \sim 30$ kV. The electric field points away from corotation. Radial currents vanish at corotation. However, at corotation there are non-vanishing azimuthal currents, such as the grad–B drift current, and the current produced by the drift of the corotation charge density. There are also vertical field-aligned currents, coming from the non-vanishing divergence of the radial current at corotation.

Physical model: Modeling the ionosphere-magnetosphere-ring system as infinite plane parallel conducting plates joined by vertical magnetic field lines (see Figure 5.7), the electric fields in the two frames are related by

$$\mathbf{E}_r = \mathbf{E_i} + \frac{\mathbf{v} \times \mathbf{B}}{c}, \tag{5.51}$$

where $v = v_r - v_i = (\Omega_K - \Omega_S)r$. The relative velocity v is zero at corotation. We use local Cartesian coordinates in which y is the azimuthal direction, and x is radial, pointing outwards. The electric field is always perpendicular to the field lines, which are perfect conductors along their lengths. When E is measured in the rest frame of a conducting plate, the height-integrated current is

$$\mathbf{J} = \sigma \mathbf{E} + \Sigma \frac{\mathbf{B} \times \mathbf{E}}{B}, \tag{5.52}$$

where σ is the height-integrated direct (Pedersen) conductivity $(J \parallel B)$, and Σ is the height-integrated Hall conductivity $(J \perp B)$. Radial electric fields will thus produce both radial and azimuthal currents in the rings and ionosphere. The radial currents must close in loops, with field-aligned currents joining the plates. Azimuthal currents are simply ring currents, and close on themselves.[18]

The radial currents in the ionosphere and rings are

$$J_{x,\text{ionosphere}} = \sigma_i E_{x,i}, \quad J_{x,\text{rings}} = \sigma_r \left(E_{x,i} + \frac{vB}{c} \right). \tag{5.53}$$

Using Kirchoff's law, $J_{x,\text{ionosphere}} + J_{x,\text{rings}} = 0$, we obtain

$$E_{x,i} = \frac{-vB}{c} \frac{\sigma_r}{\sigma_i + \sigma_r}. \tag{5.54}$$

The frame in which there is no radial electric field is that in which particles in the magnetosphere between the plates are at rest. This frame will have a velocity v_m (measured in the ionosphere

[18] For azimuthally homogeneous conductivity only — see Chapter 6.

frame) such that $E_i + v_m \times B/c = 0$. This is

$$v_{m,x} = -\frac{cE_{x,i}}{B} = \frac{\sigma_r}{\sigma_i + \sigma_r} v. \tag{5.55}$$

Magnetic field perturbations: Currents bend magnetic field lines. When a magnetic field line passes through a height-integrated current sheet J, there is a "kink" in the field line, of size determined by Maxwell's equations:

$$\nabla \times B = \frac{4\pi j}{c}, \tag{5.56}$$

When j is in the $x, y-$plane, and Eq. 5.56 is integrated through the current sheet,

$$\Delta B_x = \frac{4\pi J_y}{c}, \;\; \Delta B_y = -\frac{4\pi J_x}{c}. \tag{5.57}$$

A field line bends by an angle $\theta \sim \Delta B/B$ upon passing through the current sheet.

5.8.1 Conductivities in a magnetic field

When an electric field is applied perpendicular to the magnetic field, a charged particle gyrating around B drifts at constant speed in the $E \times B$ direction. The particle's guiding center is not free to accelerate in the direction of the electric field. Only when the particle suffers a collision does it have the chance to move a little in the direction of E, before it is "picked up" again by the magnetic field. This is illustrated in Figure 5.8. A particle starting from rest will acquire a velocity in the electric field direction of $v_i = qEt/m$. Pickup happens when $F_B = F_E$, i.e., when $v_i = v_p \sim cE/B$, after which only $E \times B$ drift occurs. Before pickup, the particle has traveled a distance v_p/Ω_g, which corresponds to the gyroradius of the particle r_L once it is picked up.

When there are few collisions per gyroperiod, the particle spends most of its time drifting, and the mean velocity in the electric field (x) direction is $v_x \sim r_L/t_c$, where t_c is the time between collisions. In the opposite limit, $v_x \sim qEt_c/m$.

The $E \times B$ drifts may give rise to Hall currents. The velocity in this direction is $v_y \sim cE/B$ in

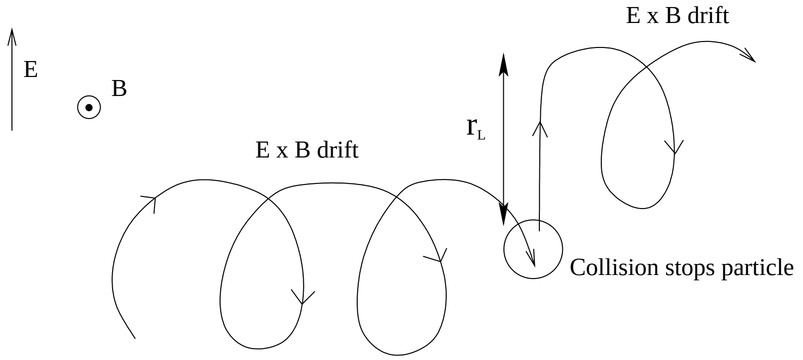

Figure 5.8: A positively charged particle in crossed electric and magnetic fields. The particle's guiding center can only accelerate in the direction of the electric field for $\sim$ one gyroperiod after a collision.

the low collision rate case, and $v_y \sim 0$ when $\Omega_g t_c \ll 1$.

Accelerations along magnetic field lines are not restricted by the magnetic field, and so we always have $v_z \sim qEt_c/m$. In the case of infrequent collisions, v_z can become large, so that there is often little resistance along field lines.

When the drift velocities are calculated in detail using the equations of motion (see Appendix), we obtain

$$\bar{v}_x = \frac{qE}{m}\frac{\nu_c}{\Omega_g^2 + \nu_c^2}, \tag{5.58}$$

$$\bar{v}_y = \frac{qE}{m}\frac{-\Omega_g}{\Omega_g^2 + \nu_c^2}, \tag{5.59}$$

where Ω_g is the gyrofrequency, and ν_c is the collision frequency $= 1/t_c$.

The currents are then given by sums over all the species of particles:

$$j_x = \sum_i n_i q_i v_{x,i}, \;\; j_y = \sum_i n_i q_i v_{y,i}, \tag{5.60}$$

and the height-integrated conductivities are

$$\sigma \sim \frac{j_x h}{E}, \;\; \Sigma \sim -\frac{j_y h}{E}. \tag{5.61}$$

Assuming the same temperature and collision cross sections for all species, the collision frequency scales with particle mass as $\nu_c \propto m^{-1/2}$, and the gyrofrequency as $\Omega_g \propto m^{-1}$. Direct conductivities depend on v_x, which at high collision rate scales as $\sigma_{\text{hicoll}} \propto m^{-1/2}$. At low collision rate, it scales instead as $\sigma_{\text{locoll}} \propto m^{1/2}$. Therefore direct currents tend to be dominated by small particles at high collision rates, and by large ones at low collision rates.

It is important to note that net Hall currents can only arise from the *difference* in drift velocity between species of opposite signs of charge. If all charges drift with equal velocity, there is no current.

5.8.2 Currents in the ring plane

We now calculate the currents from various sources in the ring plane, and the magnetic field line bends that they produce. Except where noted, these currents vanish at corotation. Numerical values are calculated for the end of a spoke at 10,000 km from corotation, where $v_p \sim 2$ km s^{-1}, assuming that the ionosphere is a much better conductor than the rings.

5.8.2.1 Electrons in Debye sheath

If there is no other plasma above the rings, the electrons in the Debye sheath may be an important source of current. The electron gyrofrequency is $\Omega_g \sim 5 \times 10^5$ s^{-1}, and the collision frequency with the rings is the Debye frequency, $\nu_c \sim 10^4$ s^{-1}. Each time an electron is ejected from the ring plane, it moves one gyroradius in the direction of the electric field before being picked up. Using Eq. 5.61, the height-integrated direct conductivity is $\sigma \sim 3 \times 10^3$ cm s^{-1}. At the spoke-end distance from corotation, this yields $\Delta B/B \sim 8 \times 10^{-12}$.

5.8.2.2 Ions in ring atmosphere

Conservative atmosphere: The mostly neutral atmosphere of the rings is collisionally coupled to the rings, meaning that it rotates at the Keplerian angular velocity. The ions in the ring atmosphere corotate with the magnetic field until they collide with a neutral in the atmosphere, then they are picked up again. A stopped ion feels an electric field v_pB/c.

The collision frequency is of order $\nu_c \sim n_{O_2}\sigma_n v_i$ for ions, corresponding to $\nu_c \sim 4 \times 10^{-4}$ s^{-1} at the end of a spoke.[19] For electrons, whose velocities are insignificantly increased by pickup, $\nu_c \sim 0.1$ s^{-1}. The ion gyrofrequency is $\Omega_g \sim 9$ s^{-1}.

Using Eq 5.61, we obtain height-integrated conductivities $\sigma_{\text{ions}} \sim 8 \times 10^6$ cm s^{-1} and $\sigma_e \sim 4 \times 10^4$ cm s^{-1} for ions and electrons, respectively. The bend in B due at the spoke ends due to this current (dominated by the ions) will be $\Delta B/B \sim 2 \times 10^{-8}$.

Liberal atmosphere: We can use the *Cassini* observation that the ions above the rings are corotating to place a constraint on the possible density of the ring ionosphere. If the ring height-integrated conductivity were larger than the ionosphere height-integrated conductivity, then the magnetosphere above the rings would rotate with the Keplerian velocity. This would have been observed by *Cassini*, which instead observed corotation above the A ring, at a place where the difference between corotational and Keplerian velocities is $v_p \sim 5$ km s^{-1}.

Using Eq. 5.61 in the low collision rate limit, where $\nu_c \propto \max[1, v_p/v_{\text{th}}]$, and neglecting the variation in B, N_{O_2} and Ω_K with distance from corotation, the height-integrated conductivity due to ions is

$$\sigma_{\text{ions}} \sim \frac{N_i e^2}{m_i}\frac{\nu_c}{\Omega_g^2} \sim 6 \times 10^6 \tau^3 \text{ cm s}^{-1}, \tag{5.62}$$

independent of distance from corotation. For constant neutral column density, the height-integrated conductivity does not depend on v_i, and so it does not depend on whether the atmosphere is heated.[20]

The ring atmosphere conductivity is less than that of the dayside ionosphere ($\sigma_{\text{day}} \sim 10^{12}$ cm s^{-1}) when

$$N_{\text{O}_2} \lesssim 6 \times 10^{16} \text{ cm}^{-2}. \tag{5.63}$$

This is a severe limit to the density of the ring atmosphere, corresponding to

$$n_{\text{O}_2} \lesssim 4 \times 10^8 \text{ cm}^{-3}, \;\; n_i \lesssim 7 \times 10^2 \text{ cm}^{-3} \tag{5.64}$$

[19] We neglect the possible separation in altitude of the ions and neutrals in calculating this collision rate; this may artificially increase the conductivity we calculate.

[20] assuming that the heating does not affect the Keplerian rotation of the atmosphere; this is not covered in this thesis but will be studied elsewhere

near the A ring, and

$$n_{\mathrm{O}_2} \lesssim 6 \times 10^8 \ \mathrm{cm}^{-3}, \ \ n_i \lesssim 2 \times 10^4 \ \mathrm{cm}^{-3} \tag{5.65}$$

near corotation. The ion mean free path near corotation is then $\lambda \sim 2 \times 10^6$ cm $\sim h/50$, so the diffusive atmosphere is a reasonable assumption.

The effect of Coulomb collisions is considered in the Appendix, where it is found that they are unimportant.

If we accept that the ionospheric conductivity of Saturn is 100 times smaller on the nightside than on the dayside (Kaiser et al., 1984) and §5.9, $\sigma_{\mathrm{night}} \sim 10^{10}$ cm s^{-1}, then we should take this as our limit on the conductivity of the ring ionosphere, since the *in situ* plasma observations were made over a part of the rings that was magnetically connected to Saturn's nightside ionosphere. In this case we obtain the more stringent limit

$$N_{\mathrm{O}_2} \lesssim 1 \times 10^{16} \ \mathrm{cm}^{-2} \quad \text{liberal atmosphere,} \tag{5.66}$$

corresponding to

$$n_{\mathrm{O}_2} \lesssim 7 \times 10^7 \ \mathrm{cm}^{-3}, \ \ \mathrm{n_i} \lesssim 1 \times 10^2 \ \mathrm{cm}^{-3} \tag{5.67}$$

near the A ring, and

$$n_{\mathrm{O}_2} \lesssim 1 \times 10^8 \ \mathrm{cm}^{-3}, \ \ \mathrm{n_i} \lesssim 3 \times 10^3 \ \mathrm{cm}^{-3} \tag{5.68}$$

near corotation. Thus near corotation the ion-neutral collision rate is $\nu_c \lesssim 3 \times 10^{-2}$ s^{-1}, and the mean free path is $\lambda \gtrsim 200$ km.

5.8.2.3 Dust grains

Currents carried by dust grains depend on the charging environment. For definiteness, we consider here the part of the dust layer outside the Debye sheath and in the conservative atmosphere with $n_e \sim 4$ cm^{-3}.[21] Collision with electrons is faster than photoejection by a rate of approximately 10,

[21] We neglect the possible reduction of this density due to collisional heating.

so we assume that the grains just absorb electrons over their 5 hour time of flight, up to a maximum of $Z = -Z_{\rm max}$.

Small grains will not reach $Z_{\rm max}$ in this time. The size at which a grain just becomes maximally charged is

$$a^2 n_e v_e t_b \simeq Z_{\rm max} = 690 a_\mu, \tag{5.69}$$

i.e., $a_\mu \sim 0.02$. Grains smaller than this acquire a charge $Z \sim 4 \times 10^4 a_\mu^2$ (which for grains smaller than $a_\mu \sim 5 \times 10^{-3}$ is a probabilistic quantity), and larger grains are charged to $Z \sim 690 a_\mu$.

Using the grain densities from micrometeoroid impacts in §5.5, the charge density on grains in a logarithmic range around a is

$$Zn_g \sim 7e \text{ cm}^{-3}, \;\; \mathrm{a}_\mu < 0.02 \tag{5.70}$$

$$Zn_g \sim \frac{0.1e}{a_\mu} \text{ cm}^{-3}, \;\; \mathrm{a}_\mu > 0.02. \tag{5.71}$$

We note that this corresponds to all of the electron density in the plasma being absorbed by small dust grains. This is not a problem because more plasma will diffuse in from outside the dust layer on a timescale less than 5 hours.

The gyrofrequency of a grain with $Z = Z_{\rm max}$ is

$$\Omega_g \sim \frac{Z_{\rm max} eB}{\rho a^3 c} \sim \frac{3 \times 10^{-7}}{a_\mu^2} \text{ s}^{-1}, \;\; \mathrm{a}_\mu > 0.02, \tag{5.72}$$

and for smaller grains, it is

$$\Omega_g \sim \frac{2 \times 10^{-5}}{a_\mu} \text{ s}^{-1}, \;\; 0.005 < \mathrm{a}_\mu < 0.02. \tag{5.73}$$

For grains that can have only one charge or zero ($a_\mu < 5 \times 10^{-3}$), the gyrofrequency for those that are charged is

$$\Omega_g \sim \frac{5 \times 10^{-10}}{a_\mu^3} \text{ s}^{-1}, \;\; \mathrm{a}_\mu < 0.005. \tag{5.74}$$

We assume for all grains that the collision frequency corresponds to collision with the rings every 5 hours, i.e. $\nu_c \sim 6 \times 10^{-5}$ s^{-1}. Grains larger than $a_\mu \sim 0.07$ have $\nu_c > \Omega_g$, and hardly feel the magnetic field on 5 hour timescales. Using the same approach as in previous sections, the grain conductivity is (for logarithmic size ranges around a_μ):

$$\begin{aligned} \sigma \quad &\sim \quad \frac{3 \times 10^4}{a_\mu^3} \text{ cm s}^{-1}, \text{ a}_\mu > 0.07, && (5.75) \\ & \quad\quad 1 \times 10^9 a_\mu \text{ cm s}^{-1}, \ 0.005 < \text{a}_\mu < 0.07, && (5.76) \\ & \quad\quad 3 \times 10^{13} a_\mu^2 \text{ cm s}^{-1}, \text{ a}_\mu < 0.005. && (5.77) \end{aligned}$$

The greatest contribution to the direct conductivity comes from grains at the boundary of the first two categories: these are the grains that are light enough to respond to the electric field but heavy enough that they are not much affected by being tied to field lines. Using $\sigma \sim 9 \times 10^7$ cm s^{-1} as our nominal dust direct conductivity, the magnetic field line bend at spoke-end is $\Delta B/B \sim 3 \times 10^{-7}$.

To calculate the Hall currents due to dust grains, we could go through a process similar to the above, using our formulae for Hall conductivity. However, it is better to note that the grains will be immersed in a net positively charged plasma. The plasma will be picked up essentially immediately. Grains that corotate will cancel the current due to the drift of the positively charged plasma that compensates their charge. Hall currents are only due to those dust grains that are not picked up by the magnetic field. Thus the Hall current is also dominated by the large grains (or the drift of the positive ions which balance their charge, depending on one's frame) with $a_\mu \gtrsim 0.07$. This results in a Hall conductivity

$$\Sigma \sim \frac{n_q evh}{E} \sim \frac{n_q ech}{B} \sim \frac{5 \times 10^6}{a_\mu} \text{ cm s}^{-1}. \tag{5.78}$$

The radial magnetic field perturbation at spoke-end produced by the resulting ring current (using $a_\mu = 0.07$) is $\Delta B/B \sim 2 \times 10^{-7}$. We conclude that currents due to the expected steady density of dust grains above the rings are small, and that they perturb the magnetic field little.

Particle	Ω_g (s^{-1})	thermal r_L (cm)	r_L at spoke end (cm)
Electron (1 eV)	5×10^5	1×10^2	1×10^2
Electron (80 K)	5×10^5	1×10^1	1×10^1
O_2^+ ion (80 K)	9	2×10^3	2×10^4
1 μ grain	3×10^{-7}	4×10^7	8×10^{11}
0.07 μ grain	6×10^{-5}	2×10^5	4×10^9
0.03 μ grain	3×10^{-4}	3×10^4	6×10^8

Table 5.3: Typical gyrofrequencies and gyroradii for particles at corotation and at the end of a spoke.

5.8.3 Nonvanishing currents at corotation

5.8.3.1 Ring current due to grad B drift

Once ions or dust grains are picked up, they start to drift due to the gradient of the magnetic field strength. Because charges of opposite signs drift in opposite directions, the grad B drift produces a ring current even where there is no Hall current.

The grad B drift velocity is

$$v_d \sim v_i \frac{\nabla B}{B} r_L. \tag{5.79}$$

The velocity v_i is the total velocity of the particle, i.e., original thermal velocity plus pickup velocity cE/B. Thus the resulting current does not decrease to zero at corotation. Similarly, the Larmor radius in the above is for v_i as defined above, i.e., it is the "hot" Larmor radius, if this is larger than the pickup Larmor radius.

The drift current is then

$$J \sim NZev_d \sim \frac{c}{BR} N\epsilon, \tag{5.80}$$

where $\epsilon = mv^2$ of the particles being considered. At corotation, where only thermal velocities contribute to the kinetic energy of gyration, the quantity $N\epsilon$ is largest for the electrons, for our collisionless conservative atmosphere. These give rise to a current that gives $\Delta B/B \sim 9 \times 10^{-10}$ at corotation.

Away from corotation, the gyrational velocities are augmented by the pickup velocity. This makes the biggest difference to dust grains, because their thermal velocities are small. At spoke-end, the grad B drift current for dust grains is

$$J \sim 1 \left(\frac{a_\mu}{0.03}\right) \text{ abamps cm}^{-1}, \tag{5.81}$$

for $a_\mu < 0.03$. The resulting field line bend is $\Delta B/B \sim 1 \times 10^{-8}$.

5.8.3.2 Ring current due to corotation charge

In order for the magnetosphere to corotate with Saturn, a polarization charge density is required (Goldreich and Julian, 1969);

$$\rho_c = \frac{\Omega_K \cdot B}{2\pi c}. \tag{5.82}$$

At corotation, this corresponds to an electron density of $n_e \sim 5 \times 10^{-8}$ cm^{-3}. Because this charge is corotating, there is an associated height-integrated azimuthal current $J \sim \rho_c R_S \Omega_K r$, whose magnitude is such that the radial perturbation to the magnetic field is $\Delta B/B \sim 5 \times 10^{-9}$.

5.8.3.3 Field-aligned currents

Because the rings are differentially rotating, the radial current driven by their motion through the electric field varies with radius. Because all currents must close, there are field-aligned currents joining the rings and ionosphere (see Figure 5.9). If the current carried requires that the current-carrying particles drift faster than their thermal speed, then instabilities (e.g. the two-stream instability) can occur, leading to an increase of resistance along the field lines. We have so far assumed that field lines are perfectly conducting along their length. If the charge carriers are calculated to travel faster than the speed of light, then there is definitely a problem.

The height-integrated current in the rings, $J_r = \sigma E$, vanishes at corotation, but $dJ_r/dr \neq 0$. Current conservation requires $\nabla \cdot \mathbf{J} = 0$. A z-current is therefore necessary:[22]

$$j_z = -\frac{1}{2}\nabla \cdot J_r = -\frac{1}{2r}\frac{\partial}{\partial r}(rJ_r) \sim \frac{J_r}{R_S}. \tag{5.83}$$

[22] The factor of two accounts for the current carried in both hemispheres.

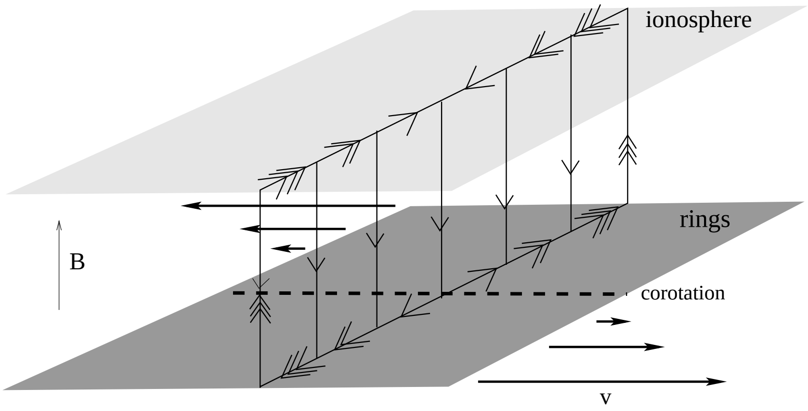

Figure 5.9: Field-aligned currents required when the rings are differentially rotating. Only one hemisphere of the current system is shown.

Property	Value
Temperature	10^3 K
Depth	100 km
Electron density (dayside)	10^5 cm^{-3}
Electron density (nightside)	10^3 cm^{-3}
Magnetic field	0.2 G
Electron gyrofrequency	3×10^6 s^{-1}
Proton gyrofrequency	2×10^3 s^{-1}
Neutral density	10^{11} cm^{-3}
Dominant species	H_2
Altitude	10^3 km

Table 5.4: Ionospheric properties from Atreya et al. (1984).

At corotation the field-aligned current flows on to the rings. This agrees with the *Cassini* observation of electrons streaming off the rings there (Gurnett et al., 2005).

Using the largest of the conductivities at corotation that we calculated in previous sections ($\sigma \sim \sigma_{\rm day}$), we obtain $j_z \sim 2 \times 10^{-4}$ abamps cm^{-2}. This current requires $n_e v_e \sim 5 \times 10^5$ cm^{-2} s^{-1}, which could be a problem at low magnetospheric densities. Using the observed electron density at corotation $n_e \sim 0.03\,{\rm cm}^{-3}$ from Gurnett et al. (2005), the required streaming velocity is $v_e \sim 2 \times 10^7$ cm s^{-1}, unlikely to be a problem. Currents must be carried all the way to the ionosphere, however, and at higher altitudes the electron density is almost certainly lower.

5.9 The ionosphere

Saturn's ionosphere is at an altitude at which the main neutral species is molecular hydrogen. Some of its properties are listed in Table 5.4. The ionospheric electron densities have been measured through propagation effects on coherent radio signals from Voyager at the terminator (e.g. Lindal et al., 1985). The day/night variations are calculated by monitoring the cutoff frequency of Saturn Electrostatic Discharges, thought to be radio wave emissions from lightning in the lower atmosphere. The plasma frequency in the ionosphere is inferred from these cutoffs (Kaiser et al., 1984).

The conductivity of the ionosphere is very important to the electrodynamics of Saturn's magnetosphere. Here we estimate the ionospheric conductivities (Hall and direct). For this, we need to know the collisional and gyrofrequencies for the charge-carrying species, protons and electrons. Since the electron and proton temperatures are the same, and the low collision rate regime applies, the protons are the dominant current-carriers, and we need consider only their properties.

5.9.1 Collision types

We consider the rates of each of the types of collision suffered by ionospheric protons:

Neutral collisions: Collisions with hydrogen molecules have cross sections of order $\sigma \sim 10^{-15}$ cm^2. At 1000 K, proton thermal velocities are of order $v_p \sim 3 \times 10^5$ cm s^{-1}. Then the collision frequency is

$$\nu_c \sim n_{\mathrm{neut}} \sigma v_p \sim 10^2 \ \mathrm{s}^{-1}. \tag{5.84}$$

Some of these collisions are charge-exchange reactions, though not many, because the ionosphere is composed of mainly molecular hydrogen, which has a higher ionization energy (15.4 eV) than atomic hydrogen (13.6 eV). Charge exchange is possible with hydrogen molecules in high vibrational states ($v > 4$) (Atreya and Waite, 1981), but collisions with such molecules are sufficiently infrequent to be irrelevant for calculating conductivities.

Coulomb collisions: Collisions with protons do not affect the drift of other protons. We need only consider collisions with electrons. The rate of isotropization of the momentum of a proton in the drift frame of the electrons is then

$$\nu_c \sim \frac{m_p}{m_e} n_e \sigma_c v_e \sim 3 \times 10^{-3} \text{ s}^{-1} \text{ by day,} \tag{5.85}$$

and perhaps 100 times slower at night. The rate of isotropization of an electron's momentum in the drift frame of the protons is

$$\nu_c \sim n_i \sigma_c v_e \sim 2 \times 10^{2} \text{ s}^{-1} \text{ by day.} \tag{5.86}$$

As in §5.7.2 for the ring ionosphere, this makes little difference to the conductivities due to either protons or electrons (see Appendix).

Comparison: Ion-neutral collisions are the dominant kind of collision suffered by protons, and the rate $\nu_c < \Omega_g$, meaning that the protons are in the low-collision regime. The electrons are even more collisionless ($\Omega_g/\nu_c \propto m^{-1/2}$), and basically perform an $E \times B$-drift. Using our equations 5.58 and 5.60, the direct conductivity is given by

$$\sigma_p = \frac{h n_p e^2}{m_p} \frac{\nu_c}{\Omega_g^2 + \nu_c^2}, \tag{5.87}$$

where h is the depth of the ionosphere. The Hall conductivity is obtained by subtracting the proton drift velocity (Eq. 5.59) from the electron drift velocity, to find the net drift rate of charge. This rate is

$$v_d \sim \frac{cE}{B} - \frac{eE}{m_p} \Omega_g \Omega_g^2 + \nu_c^2 \simeq \frac{cE}{B} \left(\frac{\nu_c}{\Omega_g} \right)^2 . \tag{5.88}$$

For our daytime values above, we obtain a height-integrated direct conductivity of $\sigma_{\text{day}} \sim 1 \times 10^{12}$ cm s^{-1} and a height-integrated Hall conductivity $\Sigma_{\text{day}} \sim 3 \times 10^{9}$ cm s^{-1}. Night-time conductivities are each about 100 times smaller, due to the reduced electron (and hence ion) density.

Our calculated direct conductivity agrees well with published values (Cheng & Waite 1988; Atreya et al. 1984). Compared with Atreya et al. (1984) values, our Hall conductivity is about an order of magnitude too low; this may be because a more detailed calculation would involve an integral over altitude in the atmosphere, and the Hall conductivity is higher at lower altitudes because the collision rate is higher.

As claimed in §5.8, the ionospheric height-integrated conductivity is higher than all of those found for the ring plane (except in §5.8.2.2 in which we deliberately set them equal as a limit), and so it was reasonable to assume in discussions that the magnetosphere corotates with the ionospheric angular velocity.

5.9.2 Ionospheric winds

The ionosphere is itself differentially rotating: the two ends of a field line joining north and south hemispheres can be in regions with different wind speeds. There are latitudinal bands with large wind speeds (Porco et al., 2005b) as high as $v_w \simeq 400$ m s^{-1}, and the ionosphere is frictionally coupled to the rest of the atmosphere. Radial currents have to flow to maintain the field lines as equipotentials. The calculation is analogous to those for the ionosphere-magnetosphere-rings system.

Assuming that the two hemispheres have the same direct conductivity, the radial current is

$$J_r = \sigma_{\rm day} v_w B/(2c) \sim 1 \times 10^5 \text{ abamps cm}^{-1}. \tag{5.89}$$

The resulting azimuthal field line bend is $\Delta B/B \sim 3 \times 10^{-4}$, the largest such bend we have found so far.

This current must close through the ionosphere, and the width of shell over which it flows is set by the sharpness of the edge of the windy bands. Scaling to an edge of $x = 5000$ km (taken from Porco et al., 2005b), the current density that must then be carried by the magnetosphere is

$$j_z \sim \frac{J_r}{x}, \tag{5.90}$$

giving a required flux

$$n_e v_e \sim \frac{4 \times 10^5}{(x/5000 \text{ km})} \text{ cm}^{-2} \text{ s}^{-1}. \tag{5.91}$$

The latitudes that are magnetically connected to the region of corotation on the rings (about 45 degrees) have typical zonal wind speeds about 4 times lower than this, at about 100 km s^{-1}.

5.10 Summary of environment: what have we learned and where do we go from here?

Through the above studies we have obtained an intimate familiarity with the electrodynamics of the Saturn ring system as relevant for making spokes. We know how much dust is present over the rings, how the dust charges, and how much dust can be lifted off the rings. The electrodynamical difference between day and night is understood, as is the effect of seasons on dust charging. We have explored the discovery by *Cassini* of oxygen ions over the ring, inferring the thickness of an atmosphere/ionosphere over the rings, and determining how it would manifest itself electrodynamically. The atmosphere by itself is of great scientific interest, regardless of the spoke connection. We have calculated various currents in the ring system, noting that many of these are smallest near corotation. We now understand the spoke-forming environment better than anyone has before. This piecewise approach to problem solving is highly educational. We know enough to discredit the leading spoke theory (see Chapter 6).

It is still difficult to see how spokes are formed. Formation probably requires large electric fields, or large currents to bombard the ring surface. These are generically absent near corotation. We must investigate perturbations to the equilibrium system that can produce large fields or currents.

Unstable magnetosonic waves in a conducting fluid could potentially produce large electric fields or magnetic perturbations. Before the discovery of the ring ionosphere, we rejected this possibility because there was no candidate fluid with sufficient conductivity to support persistent waves. Now, given the presence of the ionosphere/atmosphere, this must be reassessed. Corotation might be singled out in this case because of amplification in a resonant cavity, possibly delimited by qualitatively different atmospheric properties, such as the “puffiness” discussed in §5.7.2.

It is tempting to associate radial spokes with the paths of radial currents, which we know already to be present. Like currents attract, and so a current bunching instability might produce overdensities in radial lines. The problem with this is that currents are typically small, and are smallest near corotation so that such an instability would be unlikely to produce spokes as observed.

Another way to build up electric fields is through charge separation, i.e., electrostatically. We investigated this possibility in some depth, and a potential spoke formation theory based on charge separation through a 2-stream instability is described in the next section.

5.11 2-stream instability

A promising spoke-formation mechanism that we investigated derives from the well-known electrostatic 2-stream instability, which prevents fluids of charged particles from passing smoothly through each other. The relative motion comes from the difference between corotational and Keplerian velocities. In the simplest version of our theory, the fluids are composed of charged dust grains. Large dust grains ($a \gtrsim 0.03$ micron) are not picked up by the magnetic field during their 5 hour lifetimes, and so remain in Keplerian orbits. Lighter dust begins to corotate before reaffixing to the rings. The motion of these two fluids through each other is unstable, and leads to “bunching” of the charges and build-up of azimuthal electric fields (see Figure 5.10). If the wavevector k of the unstable mode also has a vertical component, then vertical electric fields will also be produced, lifting more charged grains off the rings and potentially making a spoke. We adopt here $k_y \sim k_z \sim 1$ km, the height of the dust layer.[23]

The truly attractive feature of this instability is that it only operates near corotation. Throughout this chapter we have seen that large electric fields are generally absent at corotation, making the 2-stream instability especially interesting. The instability is only operational when the difference in fluid velocities is less than some limit

$$v_p < \frac{\omega_p}{k_y}, \tag{5.92}$$

[23] The observed k_y of a spoke is much less than this.

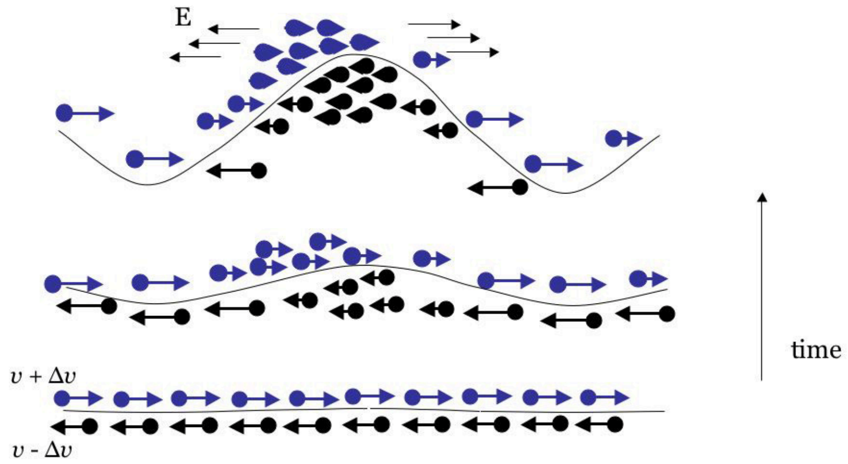

Figure 5.10: The progress of the 2-stream instability. Assume that the right- and left-moving fluids have the same sign of charge (and that there is a neutralizing background in the unperturbed state). Starting from the bottom of the figure: if a slight clumping occurs in the left-moving fluid, then the right-moving fluid is repelled from the clump. This causes the particles of the right-moving fluid to slow down, so they clump too. The left-moving fluid sees this clump, and clumps more, and so on. In this way, large density contrasts and electric fields are built up, so long as the fluids have time to respond as they slide over the clumps of charge, as discussed in the text.

where ω_p is the plasma frequency. Physically, the limit on v_p corresponds to the fact that the fluids cannnot respond to changes in potential that happen faster than the inverse of their plasma frequency. If the fluid moves too quickly through a pileup of charge, there is no unstable build-up of the pile.

In this simple example we consider the two fluids to comprise grains just light enough to corotate, and grains just too heavy to corotate, so that their plasma frequencies are similar. A fuller treatment, including fluids with different plasma frequencies, is given in the Appendix. Using the results of §5.8.2.3, we estimate

$$\omega_p \sim \left(\frac{4\pi n_g Z^2}{m} \right)^{1/2} \sim 1 \text{ s}^{-1}. \tag{5.93}$$

The critical velocity above which the instability does not operate is then $v_p \sim 1$ km s^{-1}, corresponding to a distance $\sim 4,000$ km from corotation. This is intriguingly similar to the extent of the spoke region.

The growth rate of the instability is of the same order as the plasma frequency, so electric fields amplify on a timescale ~ 1 s, easily fast enough to explain the observations. The instability as described above can produce an electric field ~ 100 times larger than the (vacuum) daytime Debye field at the rings. Combined with the above agreement on the size and position of the unstable region, this seems to be a plausible spoke formation method.

The problems, of course, lie in the details. The most serious of these is the omission in the above of the other species that will be corotating, in addition to the light dust grains. Specifically, plasma poses a problem, and the discovery of the oxygen ionosphere over the rings decreased our enthusiasm for this theory. Because plasma particles are light, they tend to short out vertical electric fields by sliding along field lines. The instability then saturates at very small electric fields. Other problems include explaining why spokes formed by the 2-stream process should be radial. Investigations are ongoing.

5.12 Previous attempts

Over the years, many potential spoke explanations have been advanced. These range from alignment of dust grains (e.g., Carbary et al., 1982), electron beams from the ionosphere (Hill & Mendis, 1981, 1982), meteoroid impacts and many more. The most widely accepted (and most sophisticated) of these theories is that by Goertz and Morfill (1983), in which spokes form under a moving plasma cloud created by an impact. In Chapter 6, we show that this theory is physically inconsistent, and that a proper treatment of the process does not lead to spokes. Another theory is that of Tagger et al. (1991) who propose that the spokes form at corotation because of the interaction of magnetosonic waves in the disk of the rings. However, this treatment treats the rings as a perfect conductor, and the magnetosphere as a vacuum, both of which are certainly false, and may very well lead to misleading results.[24]

There are problems with every theory that has been proposed. Many of the theories share the problem that spokes are not formed at corotation. It is very difficult to understand how such an apparently energetic process can happen at what one would expect to be electrically the quietest place in the rings. The spoke mystery remains an open problem.

5.13 Conclusions

In our study of spokes so far we have implemented ideas from many areas of physics to familiarize ourselves with the electrodynamical environment of Saturn's rings. In particular:

1. We have presented the first comprehensive review of spoke properties since the *Voyager* era, including recent observational surprises.

2. We have conducted a thorough tour of the Saturn system as relevant to spoke formation, including physical assessments of

    (a) dust production rates

[24] We rejected all wave theories before, on the grounds of the poor conductivity of the ring plane, but a new assessment is warranted in light of the discovery of the potentially highly conducting oxygen ionosphere/atmosphere

   (b) dust and ring charging by photons and plasma

   (c) the newly discovered oxygen atmosphere and ionosphere over the rings. This may hold the key to spoke formation.

   (d) ring currents and conductivities near corotation

3. We have discussed the feasibility of spoke formation via a 2-stream instability, and indicated that the presence of the ring atmosphere may prevent its operation

4. We have discredited the leading theory for spoke formation (see Chapter 6)

Data from *Cassini* continue to be released, and we are ready to understand their relevance as they arrive. Further study of the effects of the ring ionosphere are merited, and may reveal the origin of the spokes well before their reappearance in 2007.

Acknowledgements

We thank C. Porco, J. H. Waite, and D. Young for useful discussions on results from *Cassini*.

Bibliography

Achterberg, A., Blandford, R. D., Goldreich, P. 1983, unpublished manuscript.

Atreya, S. K., Donahue, T. M., Nagy, A. F., Waite, J. H., McConnell, J. C. 1984. Theory, measurements, and models of the upper atmosphere and ionosphere of Saturn. Saturn 239–277.

Atreya, S. K., Waite, J. H. 1981. Saturn ionosphere — theoretical interpretation. Nature 292, 682.

Blum, J. 2004 Grain growth and coagulation. ASP Conf Series 309, 369–391, Eds. Witt, A. N., Clayton, G. C., Draine, B. T.

Carbary, J. F., Bythrow, P. F., Mitchell, D. G. 1982. The spokes in Saturn's rings — A new approach. Geophysical Research Letters 9, 420–422.

Cheng, A. F., Waite, J. H. 1988. Corotation lag of Saturn's magnetosphere — Global ionospheric conductivities revisited. Journal of Geophysical Research 93, 4107–4109.

Cuzzi, J. N., Lissauer, J. J., Esposito, L. W., Holberg, J. B., Marouf, E. A., Tyler, G. L., Boishchot, A. 1984. Saturn's rings — Properties and processes. IAU Colloq. 75: Planetary Rings 73–199.

Cuzzi, J. N., and 10 colleagues 2002. Saturn's Rings: pre-*Cassini* status and mission goals. Space Science Reviews 104, 209–251.

Doyle, L. R., Dones, L., Cuzzi, J. N. 1989. Radiative transfer modeling of Saturn's outer B ring. Icarus 80, 104–135.

Esposito, L. W., Cuzzi, J. N., Holberg, J. B., Marouf, E. A., Tyler, G. L., Porco, C. C. 1984. Saturn's rings — Structure, dynamics, and particle properties. Saturn 463–545.

Goertz, C. K., Morfill, G. 1983. A model for the formation of spokes in Saturn's rings. Icarus 53, 219–229.

Goldreich, P., Julian, W. H. 1969. Pulsar electrodynamics. Astrophysical Journal 157, 869.

Grun, E., Morfill, G. E., Terrile, R. J., Johnson, T. V., Schwehm, G. 1983. The evolution of spokes in Saturn's B ring. Icarus 54, 227–252.

Gurnett, D. A., and 26 colleagues 2005. Radio and plasma wave observations at Saturn from *Cassini*'s approach and first orbit. Science 307, 1255–1259.

Hall, D. T., Feldman, P. D., McGrath, M. A., Strobel, D. F. 1998. The far-ultraviolet oxygen airglow of Europa and Ganymede. Astrophysical Journal 499, 475.

Hill, J. R., Mendis, D. A. 1981. On the braids and spokes in Saturn's ring system. Moon and Planets 24, 431–436.

Hill, J. R., Mendis, D. A. 1982. The dynamical evolution of the Saturnian ring spokes. Journal of Geophysical Research 87, 7413–7420.

Johnson, R. E., and 13 colleagues 2004. The production and redistribution of oxygen in Saturn's magnetosphere: CAPS *Cassini* data. AGU Fall Meeting Abstracts A6.

Kaiser, M. L., Desch, M. D., Connerney, J. E. P. 1984. Saturn's ionosphere — Inferred electron densities. Journal of Geophysical Research 89, 2371–2376.

Koschny, D., Grün, E. 2001. Impacts into ice-silicate mixtures: ejecta mass and size distributions. Icarus 154, 402–411.

Koschny, D., Grün, E. 2001. Impacts into ice-silicate mixtures: crater morphologies, volumes, depth-to-diameter ratios, and yield. Icarus 154, 391–401.

Lindal, G. F., Sweetnam, D. N., Eshleman, V. R. 1985. The atmosphere of Saturn — an analysis of the *Voyager* radio occultation measurements. Astronomical Journal 90, 1136–1146.

McGhee, C. A., French, R. G., Dones, L., Cuzzi, J. N., Salo, H. J., Danos, R. 2005. HST observations of spokes in Saturn's B ring. Icarus 173, 508–521.

Mendis, D. A., Hill, J. R., Ip, W.-H., Goertz, C. K., Gruen, E. 1984. Electrodynamic processes in the ring system of Saturn. Saturn 546–589.

Nicholson, P. D., Brown, R. H., Clark., R. N., Cruikshank, D. P., Showalter, M. R., Sicardy, B. 2004. *Cassini*-VIMS observations of Saturn's rings at SOI. AGU Fall Meeting Abstracts 4.

Nitter, T., Havnes, O., Melandsø, F. 1998. Levitation and dynamics of charged dust in the photoelectron sheath above surfaces in space. Journal of Geophysical Research 103, 6605–6620.

Onose, M. 1996. Master's Thesis, Hokkaido Univ.

Porco, C. C., and 34 colleagues 2005. *Cassini* Imaging Science: Initial Results on Saturn's Atmosphere. Science 307, 1243–1247.

Porco, C. C., and 34 colleagues 2005. *Cassini* Imaging science: initial results on Saturn's rings and small satellites. Science 307, 1226–1236.

Porco, C. A., Danielson, G. E. 1982. The periodic variation of spokes in Saturn's rings. Astronomical Journal 87, 826–833.

Shukla, P. K., Mamun, A. A. 2002. Introduction to dusty plasma physics. Introduction to dusty plasma physics P. K. Shukla, A. A. Mamun. Bristol; Philadelphia: Institute of Physics Publishing, c2002. (Series in plasma physics) .

Smith, B. A., and 26 colleagues 1981. Encounter with Saturn — *Voyager 1* imaging science results. Science 212, 163–191.

Smith, B. A., and 28 colleagues 1982. A new look at the Saturn system: The *Voyager 2* images. Science 215, 505–537.

Spencer, J. R., Calvin, W. M., Person, M. J. 1995. CCD spectra of the Galilean satellites: molecular oxygen on Ganymede. Journal of Geophysical Research 100, 19049–19056.

Spilker, L. J., and 11 colleagues 2004. *Cassini* CIRS: preliminary results on Saturn's rings. AGU Fall Meeting Abstracts 5.

Tagger, M., Henriksen, R. N., Pellat, R. 1991. On the nature of the spokes in Saturn's rings. Icarus 91, 297–314.

Waite, J. H., and 10 colleagues 2005. Oxygen ions observed near Saturn's A ring. Science 307, 1260–1262.

Yamamoto, S., Mukai, T. 1998. Dust production by impacts of interstellar dust on Edgeworth-Kuiper Belt objects. Astronomy and Astrophysics 329, 785–791.

Young, D. T., and 42 colleagues 2005. Composition and dynamics of plasma in Saturn's magnetosphere. Science 307, 1262–1266.

Yung, Y. L., McElroy, M. B. 1977. Stability of an oxygen atmosphere on Ganymede. Icarus 30, 97–103.

5.14 Appendix: Velocity drifts in electric fields

5.14.1 With neutral collisions

We derive the expressions in eqs. 5.58 and 5.59 for the mean drift velocities of charged particles when an electric field E_x is applied perpendicular to the magnetic field.

Starting from rest after a collision, the equation of motion for a charged particle reads

$$m\dot{v}_x = q(E + v_y B/c), \;\; m\dot{v}_y = q(-v_x B/c) \tag{5.94}$$

Combining the two differential equations and solving for v_y, with $v_y(0) = 0, \dot{v}_y(0) = 0$, we obtain

$$v_y = \frac{e}{m}\frac{E}{\Omega_g}(\cos\omega t - 1), \tag{5.95}$$

where Ω_g is the gyrofrequency of the particle around B.

The particles suffer collisions on average every $t_c = 1/\nu_c$. To obtain the mean velocity between collisions, we integrate over the collision probability distribution:

$$\bar{v}_y = \int_0^\infty e^{-\nu_c t} v_y(t)/t_c, \tag{5.96}$$

giving

$$\bar{v}_y = \frac{eE}{m\Omega_g} \frac{-\Omega_g^2}{\Omega_g^2 + \nu_c^2}. \tag{5.97}$$

The velocity in the direction of the electric field is

$$v_x = \frac{cE}{B} \sin \Omega_g t. \tag{5.98}$$

Integrating this over the collision distribution, the mean velocity is

$$\bar{v}_x = \frac{cE}{B} \frac{\nu_c \Omega_g}{\Omega_g^2 + \nu_c^2}. \tag{5.99}$$

5.14.2 Including Coulomb collisions

In the previous treatment, we calculated the drift velocity of particles based on their collisions with a species that was not drifting in the frame of reference, such as the neutral atmosphere of the rings. Coulomb collisions between charged particles may also be important. The effect of these collisions depends on the relative drift between different charged species; if there is no relative drift then the collisions have no effect on average. This requires that a current already be flowing for Coulomb collisions to be effective.

Illustrating a different technique for deriving conductivities than the above, we solve the ion and electron equations of motion in steady state, i.e., we look for drift velocities u:

$$\begin{aligned} 0 = \dot{u}_{ex} &= -\frac{eE}{m_e} - \omega_e u_{ey} - \nu_{en} u_{ex} + \nu_{ec}(u_{ix} - u_{ex}), && (5.100) \\ 0 = \dot{u}_{ey} &= \omega_e u_{ex} - u_{ey}\nu_{en} + \nu_{ec}(u_{iy} - u_{ey}), && (5.101) \\ 0 = \dot{u}_{ix} &= \frac{eE}{m_i} + \omega_i u_{iy} - \nu_{in} u_{ix} + \frac{m_e}{m_i}\nu_{ec}(u_{ex} - u_{ix}), && (5.102) \\ 0 = \dot{u}_{iy} &= -\omega_i u_{ix} - \nu_{in} u_{iy} + \frac{m_e}{m_i}\nu_{ec}(u_{ey} - u_{iy}), && (5.103) \end{aligned}$$

where $\nu_{en} = v_e/\lambda_{in}$ and $n_{in} = v_i/\lambda_{in}$ are the rates of electron-neutral and ion-neutral collisions respectively, where v_e and v_i are the thermal, not drift, ion and electron velocities. The rate of Coulomb collisions is $\nu_{ec} = v_e/\lambda_c$, determined by the thermal velocity of the electrons. Coulomb collisions remove only a fraction m_e/m_i of the ion momentum relative to the electron drift frame, whereas in one collision the electron's momentum is isotropized with respect to the ion drift frame. Equations 5.100 to 5.103 must be solved simultaneously to find the drift velocity for each species.

For the case in which $\nu_{en} \ll \Omega_{g,e}$, Coulomb collisions are more important than neutral collisions in determining u_{ex} if

$$\frac{\nu_{ec}}{\nu_{en}} > \frac{u_{ey}}{u_{ey} - u_{iy}}. \qquad (5.104)$$

Coulomb collisions are less important for determining u_{iy}, because m_i/m_e of them are required to isotropize the ion momentum, while ion-neutral collisions are only a factor $\sqrt{m_i/m_e}$ slower than electron-neutral ones.

Using the values relevant for our liberal atmosphere over the A ring from §5.8.2.2, $\nu_{en} \sim 0.04$ s^{-1}, $\nu_{ec} \sim 0.02$ s^{-1},

$$\frac{\nu_{ec}}{\nu_{en}} \sim 0.5, \quad \frac{u_{ey}}{u_{ey} - u_{iy}} \sim 5 \times 10^4. \qquad (5.105)$$

Coulomb collisions are thus unimportant in determining the conductivity.

5.15 Appendix: the 2-stream instability

In this Appendix we derive conditions for the operation of the two-stream instability as applied to the dusty plasmas over the rings of Saturn.

5.15.1 The fluid equations

We begin by solving the equations of motion for any number of fluids of charged particles moving through each other. We assume that a neutralizing background removes any net charge density in the equilibrium state. The electric field in the system depends on the sum of charge distributions from all fluids. We adopt streams of cold fluids with unperturbed velocities v_i, giving

$$\text{Poisson: } \nabla^2\phi = -4\pi \sum_i n_i' q_i \tag{5.106}$$

$$\text{Equation of Motion: } \frac{\partial u_i}{\partial t} + v_i \cdot \nabla u_i = -\frac{q_i}{m_i}\nabla\phi \tag{5.107}$$

$$\text{Continuity equation: } \frac{\partial n_i'}{\partial t} + v_i \cdot \nabla n_i' + n_i \nabla \cdot u_i = 0 \tag{5.108}$$

We now take $v = v_y$ only, and Fourier transform the above equations with $k = k_y$ only, i.e., $e^{i(k_y y - \omega t)}$. We simplify to two fluids, and make the free choice $v_1 = v$, $v_2 = -v$, resulting in the dispersion relation

$$1 = \frac{\omega_{p1}^2}{(\omega - kv)^2} + \frac{\omega_{p2}^2}{(\omega + kv)^2}, \tag{5.109}$$

where $\omega_{p,i}^2 = 4\pi n_i q_i^2/m_i$. To add further fluids, one simply adds further terms to this relation.

For convenience, we now define

$$\omega_p^2 = \omega_{p1}^2 + \omega_{p2}^2,\ \omega_p^2\Delta^2 = \omega_{p1}^2 - \omega_{p2}^2, \tag{5.110}$$

and

$$\nu = \frac{\omega}{\omega_p},\ x = \frac{kv}{\omega_p}. \tag{5.111}$$

5.15.2 Identical fluids

When the two fluids have the same plasma frequency, the dispersion relation becomes

$$\nu^2 = \frac{1}{2} + x^2 \pm \frac{1}{2}\sqrt{1+8x^2}. \tag{5.112}$$

This gives $\nu^2 < 0$, and hence pure growth, for $|x| < 1$, i.e., $|kv| < \omega_p$. The phase velocty is zero, so the growing peak of the density perturbation stays stationary in the lab frame. The maximum growth rate is of order ω_p, when $k \sim v/\omega_p$.

5.15.3 Extreme plasma frequency ratio

When the fluids have very different plasma frequencies, we can make an analytic approximation to the instability criterion by perturbing about the solution for $\Delta = 1$ in which one of the plasma frequencies is zero. The unperturbed roots are

$$\nu = -1 + x, -x, -x, 1 + x. \tag{5.113}$$

It is the $\nu = -x$ solution that will become unstable, and so we perturb around this one. The dispersion relation becomes

$$-(1+2x)(\nu')^2(1-2x) \simeq 2(1-\Delta^2)x^2. \tag{5.114}$$

Solving for ν', we obtain

$$\nu \simeq -x + \nu' \simeq -x + ix\frac{\sqrt{1-\Delta^2}\sqrt{2}}{1-4x^2} \simeq -x + ix\sqrt{2}\left(\frac{\omega_{p2}}{\omega_{p1}}\right)\frac{1}{1-4x^2}. \tag{5.115}$$

In this case, the instability will move with a phase velocity matching the velocity of the heavier component. For $x \sim 1$, the growth rate will be characteristic of the heavier component's plasma frequency. The range of unstable modes, however, is set by the plasma frequency of the lighter component.

5.15.4 Including light fluids

Thus far we have been implicitly assuming that only electric fields are relevant on the timescale of the instability, even though it is the magnetic field that has brought some of the particles into corotation. This is valid so long as $\Omega_g \ll \omega$. This is not the case for lighter particles such as electrons and ions; these are tied to vertical field lines in the sense that their guiding centers are prevented from accelerating in the direction of a horizontal electic field. One might then expect their participation in the 2-stream instability to be negligible. Indeed, when we include the gyrational motion of particles in the equations of motion, the term for each fluid in the dispersion relation Eq. 5.109 becomes

$$\frac{\omega_{p1}^2}{(\omega - kv)^2 - \Omega_g^2}. \tag{5.116}$$

For $\Omega_g \gg kv$ this term is negative and roughly constant, adding a constant to the left hand side of the dispersion relation. If such a light fluid is present along with a pair of fluids that would otherwise be unstable to the 2-stream instability, then the light fluid's presence would have little effect on the situation.

However, we have not yet considered the possibility of motion *along* the field lines, which is relevant if $k_z \neq 0$. We expect finite k_z because the dust layer is of finite height. A light fluid trapped on field lines can still move in response to vertical electric fields. Taking $u_y = 0$ (i.e., assuming $\Omega \gg \omega$) for the light fluid, and calling this fluid 1, the resulting dispersion relation is

$$1 = \frac{k_z^2}{k_{tot}^2} \frac{\omega_{p1}^2}{(\omega - kv)^2} + \frac{\omega_{p2}^2}{(\omega + kv)^2}. \tag{5.117}$$

For $k_y \sim k_z$, the light fluid's behavior is analogous to that which it would have were no magnetic field present. Because ω_{p1} is so large, it dominates the dispersion relation, and (provided there are other fluids present) the maximum unstable v_p will be determined by the plasma frequency of the light fluid. If this fluid conists of electrons with $n_e \sim 1$ cm^{-3}, then its plasma frequency $\omega_p \sim 6 \times 10^4$

s^{-1}, and the critical velocity difference is $\sim 60,000$ km s^{-1}. In this case the instability would not be confined to the region of corotation.

5.15.5 Saturation of the instability

An even more serious effect for spoke formation caused by the presence of light fluids is the reduction of the maximum electric field that can be produced by the instability. Given an unstable situation, the instability can operate until it goes non-linear, i.e., $\Delta n \sim n$ for any fluid component (giving a maximum electric field $E \sim nZe\lambda$), or when the electric field turns around a fluid particle over a wavelength, $EZe\lambda \sim mv^2$. Non-linearity begins at the smaller of these two electric fields.

The dust particles just light enough to corotate, given a wavelength of 1 km as assumed in the simplest example, would be turned around (for $v_p \sim 1$ km s^{-1}) by an electric field $\sim 6 \times 10^{-3}$ statvolt cm^{-1}. Their charge density is able to produce a field $\sim 2 \times 10^{-4}$ statvolt cm^{-1}. Therefore the saturation field is $\sim 10^2$ times the Debye sheath field, as quoted in §5.11.

However, if electrons are present, then the electrons go non-linear when the field is only $E \sim 2 \times 10^{-13}$ statvolt cm^{-1}, much smaller than even the electric field in the Debye sheath during the day. The instability may then occur, but its effect on the rings would be negligible, and we would not expect a spoke to be produced.

It is possible that, if the charge density on dust grains exceeds that on the plasma, then the instability could continue to progress once the plasma has gone non-linear. This will require further investigation.

Chapter 6

Spoke Formation under Moving Plasma Clouds

Submitted to Icarus 28 March 2005 as

Alison J. Farmer & Peter Goldreich,

"Spoke formation under moving plasma clouds"

Abstract

Goertz and Morfill (1983) propose that spokes on Saturn's rings form under radially moving plasma clouds produced by meteoroid impacts. We demonstrate that the speed at which a plasma cloud can move relative to the ring material is bounded from above by the difference between the Keplerian and corotation velocities. The radial orientation of new spokes requires radial speeds that are at least an order of magnitude faster. The model advanced by Goertz and Morfill fails this test.

6.1 Introduction

The nature of the "spokes" in Saturn's rings remains a matter of speculation 25 years after their discovery by the *Voyager* spacecraft. A brief summary of their properties is given here; for more details see Mendis et al. (1984) and references therein.

1. Spokes are transient radial albedo features superposed on Saturn's rings.

2. Spokes are composed of dust with a narrow size distribution centered at 0.5 micron.

3. Spokes have optical depths of about 0.01.

4. Spokes are only seen near corotation, which is where the Keplerian angular velocity of the ring particles matches the planet's rotational angular velocity. Corotation occurs in the outer B ring.

5. Individual spokes measure about 10,000 km in radial length and 2,000 km in azimuthal width; they extend over about 10% of the ring radius.

6. Spokes are seen preferentially on the morning ansa of Saturn's rings, and are most closely radial there.

7. Spokes fade and are distorted by differential rotation as they move from morning toward evening ansa.

8. A few observations have been interpreted as showing the birth of individual spokes within 5 minutes along their entire lengths. This timescale implies a propagation velocity of at least 20 km s^{-1}.

9. Spokes are only observed at small ring opening angles to the Sun (McGhee et al. 2005).

The above observations suggest that spoke formation involves the sudden lifting of a radial lane of dust grains from the surface of the rings. Their subsequent fading and distortion is compatible with the elevated dust grains moving on Keplerian orbits that intersect the ring plane half an orbital period (i.e., about 5 hours) later.

Currently the most popular model for spoke formation is that of Goertz and Morfill (1983, GM). GM propose that the formation of a spoke is initiated when a meteoroid impacts the ring and creates a dense plasma cloud. Electrons from the cloud are absorbed by the ring producing a large electric field which levitates negatively charged dust grains. The grains enter the cloud where they absorb additional electrons. Overall charge neutrality is maintained by the net positive charge of the plasma.

Because the dust grains are massive, they move on Keplerian orbits. The plasma in which they are immersed is, however, tied to the magnetic field lines that pass through the ionosphere of Saturn. The motion of the negatively charged dust relative to the positively charged plasma produces an azimuthal electric field that causes the plasma cloud to drift radially. GM argue that the plasma cloud will continue to levitate dust grains as it moves. According to their calculations, the plasma cloud drifts in the radial direction, away from corotation, at 20–70 km/s. This velocity is sufficient to account for the formation of a radial spoke of length 10,000 km within 5 minutes.

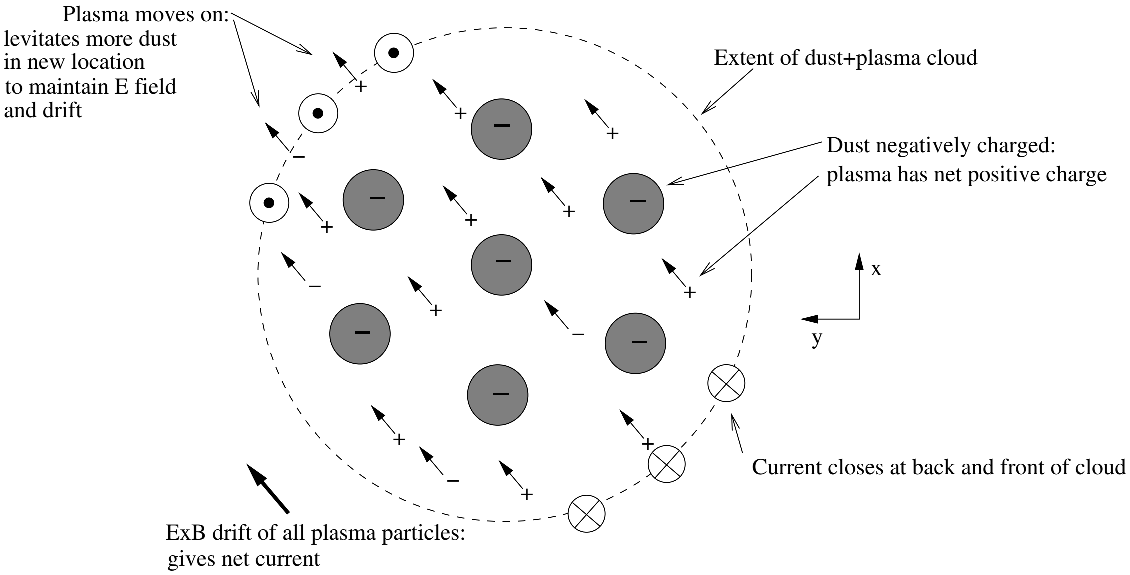
Plasma moves on:
levitates more dust
in new location
to maintain E field
and drift
Extent of dust+plasma cloud
Dust negatively charged:
plasma has net positive charge
x
y
Current closes at back and front of cloud
ExB drift of all plasma particles:
gives net current
+
−

Figure 6.1: The dusty plasma cloud, viewed from above in the local rest frame of the ring particles.

We perform a self-consistent calculation of the plasma cloud drift velocity in §6.2. It establishes that the drift velocity cannot exceed the difference between the local Keplerian and corotation velocities. This upper limit is of order 1 km s^{-1} in the region where spokes are observed. We reveal the source of GM's error in §6.3 and estimate the correct drift velocity in §6.4. A short summary in §6.5 concludes our paper.

6.2 Drift of a plasma cloud

The essence of GM's model is illustrated in Fig. 6.1. The part of the ring plane at the base of the plasma cloud has a finite Hall conductivity: in the presence of an electric field in the local rest frame of the ring particles, the net positively charged plasma will drift relative to the negatively charged dust grains which are embedded within it. Away from the plasma cloud the ring plane has negligible electrical conductivity.

Overall charge neutrality of the dusty plasma implies that currents must close. Although the magnetospheric plasma maintains as equipotentials the magnetic field lines linking the ring plane to the ionosphere, it cannot carry currents across the field lines. Thus currents that flow through the base of the plasma cloud must close in Saturn's ionosphere. In the wake of the cloud, the levitated dust rapidly combines with the positive ions that balance its charge, leaving the spoke trail behind the cloud non-conducting.

6.2.1 The model

We analyze a simple model that captures the relevant features of the system of a dusty plasma cloud, ionosphere, and magnetosphere. A 2D strip of material in the xy-plane at $z = 0$ represents the dust grains (plus neutralizing ions) at the base of the plasma cloud at a given time.[1] Infinite sheets of a different material at $z = \pm L$ take the place of the ionosphere. The magnetosphere consists of a vertical magnetic field of strength B embedded in a massless plasma that maintains the field lines as equipotentials. All calculations are done in the rest frame of the strip relative to which the ionosphere moves with velocity $\mathbf{v} = (0, v, 0)$ (see Fig. 6.2). Thus x represents the radial direction in the rings, and y is azimuthal.

We wish to determine the drift velocity of the plasma in and above the strip. This is equivalent to finding the horizontal electric field in the system since the drift velocity $\mathbf{v}_p$ measured in the same frame as $\mathbf{E}$ satisfies

$$\mathbf{v}_p = c\frac{\mathbf{E} \times \mathbf{B}}{B^2}\,. \tag{6.1}$$

6.2.2 Finding the drift velocity

In the rest frame of a conducting sheet, the height-integrated current density $\mathbf{J}$ perpendicular to the magnetic field is given by

$$\mathbf{J} = \sigma\mathbf{E} + \Sigma\frac{\mathbf{B} \times \mathbf{E}}{|B|}, \tag{6.2}$$

where σ and Σ are, respectively, the height-integrated direct (Pedersen) and Hall conductivities. Because the ionosphere consists of two sheets in parallel, its effective conductivity is twice that of a single sheet.

Subscripts s and i are used to denote properties of the strip and ionosphere. The electric field

[1] Defined in this way, the strip resembles a strip of metal, in which the dust grains are analogous to the ion lattice and the neutralizing ions are like the electrons that balance the charge on the lattice.

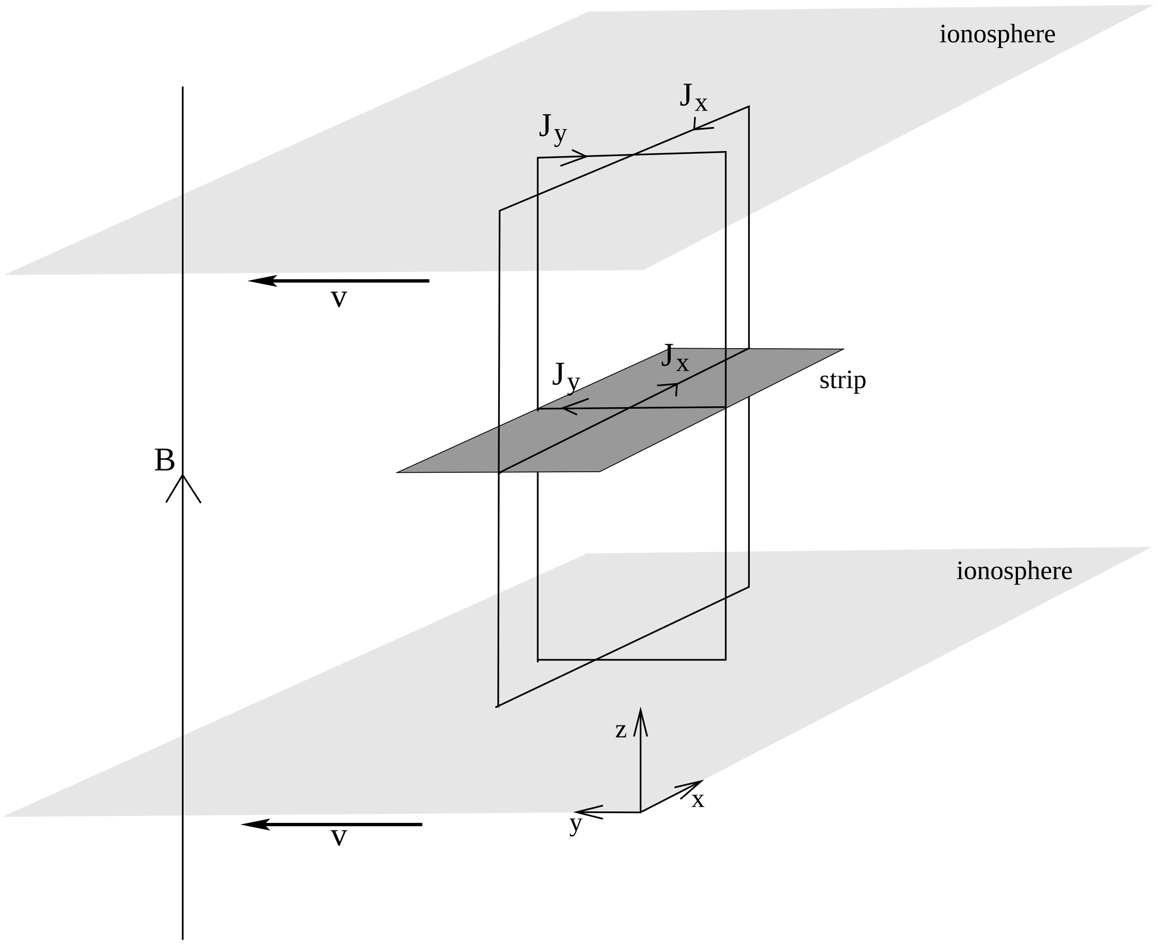

Figure 6.2: The ionosphere-strip configuration described in §6.2.1

in the rest frame of the ionosphere is related to that in the rest frame of the strip by

$$E_{xi} = E_{xs} + \frac{vB}{c}\,, \quad E_{yi} = E_{ys}. \tag{6.3}$$

We assume, as GM implicitly did, that the currents are small enough so as not to significantly perturb the externally imposed magnetic field B, i.e., $4\pi J/c \ll B$. We also take $v \ll c$.

Then, using Eq. 6.2 for the currents in the rest frame of each conductor, and expressing current conservation by

$$\mathbf{J}_s + \mathbf{J}_i = 0\,, \tag{6.4}$$

we solve the resulting simultaneous equations to give

$$E_{xs} = -\frac{vB}{c}\left[\frac{\Sigma_i\Sigma_t + \sigma_i\sigma_t}{\Sigma_t^2 + \sigma_t^2}\right], \tag{6.5}$$

$$E_{ys} = \frac{v|B|}{c}\left[\frac{\Sigma_s\sigma_i - \Sigma_i\sigma_s}{\Sigma_t^2 + \sigma_t^2}\right], \tag{6.6}$$

where $\Sigma_t = \Sigma_s + \Sigma_i$ and $\sigma_t = \sigma_s + \sigma_i$.

The velocity $\mathbf{v}_p$ at which the plasma cloud moves then follows from Eq. 6.1:

$$v_{px} = v\left[\frac{\Sigma_s\sigma_i - \Sigma_i\sigma_s}{\Sigma_t^2 + \sigma_t^2}\right]b, \tag{6.7}$$

$$v_{py} = v\left[\frac{\Sigma_i\Sigma_t + \sigma_i\sigma_t}{\Sigma_t^2 + \sigma_t^2}\right], \tag{6.8}$$

where we have introduced $b = B/|B| = \text{sign}(B)$.

The main message from eqs. 6.7 and 6.8 is that $|v_{px}|, |v_{py}| < v$.[2] This finding contradicts the result obtained by GM (i.e., $|v_{px}| = 20 - 70$ km s^{-1}), because in the region in which spokes appear, $v \lesssim 1$ km s^{-1}. The reasons for this discrepancy are detailed in §6.3.

Another interesting limit is one in which *either or both* of the direct and Hall ionospheric con-

[2] Because Hall conductivities can be negative, we can envisage the contrived case $\Sigma_s \approx -\Sigma_i$, in which we can in principle have $|v_{px}| > v$. However, this requires the unlikely cancellation of the two unrelated Hall conductivities to a high degree of accuracy. Moreover, $|v_{px}| > v$ requires $\Sigma_i > \sigma_i$, which is not the case for Saturn's ionosphere (see section 6.4).

ductivities are much larger than both components of the strip's conductivity. In this case, $|v_{px}| \ll v$ and $v_{py} \approx v$; i.e., the plasma cloud basically corotates with the ionosphere.

6.3 Where GM erred

GM neglected several terms when solving equations analogous to our Eqs. 6.1–6.4. The components of the height integrated currents in both the strip and ionosphere as a function of the electric fields in their respective rest frames satisfy:

$$\begin{aligned} J_{xs} &= [\sigma_s E_{xs}] - [b\Sigma_s E_{ys}]\,, && (6.9) \\ J_{ys} &= [\sigma_s E_{ys}] + b\Sigma_s E_{xs}\,, && (6.10) \\ J_{xi} &= \sigma_i E_{xi} - [b\Sigma_i E_{yi}]\,, && (6.11) \\ J_{yi} &= \sigma_i E_{yi} + [b\Sigma_i E_{xi}]\,. && (6.12) \end{aligned}$$

Terms left out in GM's analysis are enclosed in square brackets. Since the strip has negligible direct conductivity, neglecting terms proportional to σ_s does no harm. The Hall current in the ionosphere should be included since the Hall conductivity is comparable to the direct conductivity in the ionosphere. However, this neglect is not a large source of error. The serious omission is that of the strip's Hall current from J_{xs}, as indicated in Eq. 6.9.

The above equations, with the bracketed terms omitted, yield the incorrect result

$$v_{px} = v\frac{\Sigma_s}{\sigma_i}b\,, \tag{6.13}$$

from which $|v_{px}| > v$ follows for $|\Sigma_s| > \sigma_i$.

Physically, GM's error is described by their incorrect statement that "the motion of the plasma cloud does not constitute a perpendicular (to $\mathbf{B}$) current as the plasma cloud is charge neutral." Because the dust grains are negatively charged the plasma must have a net positive charge. The azimuthal electric field set up by the differential motion of the plasma and the dust grains does cause the plasma to drift radially, but because the plasma is net positive, this constitutes a radial Hall current. This radial current closes in the ionosphere, and thus modifies the radial electric field in both the ionosphere and the strip. But the radial electric field is responsible for driving the azimuthal Hall current in the strip, and this modification is not taken into account in GM. They therefore incorrectly fix $v_{py} = v$, leading to a severe overestimate of v_{px}. All these factors are accounted for in the self-consistent calculation outlined in §6.2.

6.4 Recalculation of drift velocity

We recalculate the drift velocity of a plasma cloud from eqs. 6.7 and 6.8.

The direct conductivity in the strip is very small because the dust grains have low mobility and the plasma is tied to the magnetic field lines. The Hall conductivity, which results from the $\mathbf{E} \times \mathbf{B}$ drift of the positively charged plasma, is correspondingly high, of order

$$\Sigma_s \sim -\frac{N_p e c}{|B|}\,, \tag{6.14}$$

where $N_p e$ is the height integrated charge density of the plasma.[3] Using the approach of GM, we estimate the height integrated charge density to be of order 10^1 esu cm^{-2}, which gives $\Sigma_s \sim -10^{14}$ cm s^{-1}.[4]

Saturn's ionospheric conductivities vary with latitude and with time, with typical dayside values of the height-integrated direct conductivity being $\sigma_i \simeq 10^{12}-10^{13}$ cm s^{-1}, and nightside values about 100 times smaller (Cheng and Waite 1988). The Hall conductivity is about an order of magnitude smaller, e.g., for the auroral region we have $\sigma_i = 5 \times 10^{13}$ cm s^{-1} and $\Sigma_i = 8 \times 10^{12}$ cm s^{-1} (Atreya et al., 1983). The direct current is carried predominantly by protons, and the Hall current by electrons, which suffer fewer collisons per gyroperiod than the protons, so $\mathbf{E} \times \mathbf{B}$ drift more freely. Collision frequencies of electrons and protons in Saturn's ionosphere are smaller than the respective

[3] The Hall conductivity is negative because it is defined as positive for the commonly encountered case in which electrons are the dominant current carriers.

[4] This number may be spuriously high, because the approach of GM gives more charge on the dust grains than was originally present in the plasma.

gyrofrequencies, so the Hall conductivity is smaller than the direct conductivity.

We substitute into eqs. 6.7 and 6.8 the typical dayside values $\sigma_i/2 = 3 \times 10^{12}$ cm s^{-1} and $\Sigma_i/2 = 3 \times 10^{11}$ cm s^{-1} (the factors of two account for both hemispheres of the ionosphere, although we note that we do not expect north-south symmetry of conductivities). We then obtain

$$v_{px} = 6 \times 10^{-2} v,\; v_{py} = -2 \times 10^{-3} v, \tag{6.15}$$

where we have used $b = -1$ as is appropriate for Saturn, and where $v \lesssim 1$ km/s. With these velocities, spokes will not form quickly or radially. The strip Hall conductivity is high enough to drag the plasma column in an almost Keplerian orbit.

6.5 Conclusions

We have studied the physical situation in which a strip of conducting material moves between two parallel sheets of a different material, to which it is joined by perpendicular magnetic field lines (Fig. 6.2). The plasma on these field lines will drift in the plane of the strip, both parallel and perpendicular to the relative velocity vector. We have shown that the magnitude of this drift velocity cannot exceed that of the relative strip-sheet velocity.

Application of this limit to the most popular model for spoke formation demonstrates that the model rests upon a gross overestimate of the velocity at which a plasma cloud can drift.

Acknowledgements

This research was supported by NSF grant AST 00-98301. We thank J. H. Waite for information on Saturn's ionospheric conductivity.

Bibliography

Atreya, S. K., Donahue, T. M., Nagy, A. F., Waite, J. H., McConnell, J. C. 1984. Theory, measurements, and models of the upper atmosphere and ionosphere of Saturn. Saturn 239–277.

Cheng, A. F., Waite, J. H. 1988. Corotation lag of Saturn's magnetosphere - Global ionospheric conductivities revisited. Journal of Geophysical Research 93, 4107–4109.

Goertz, C. K., Morfill, G. 1983. A model for the formation of spokes in Saturn's rings. Icarus 53, 219–229.

Mendis, D. A., Hill, J. R., Ip, W.-H., Goertz, C. K., Gruen, E. 1984. Electrodynamic processes in the ring system of Saturn. Saturn 546–589.

McGhee, C. A., French, R. G., Dones, L., Cuzzi, J. N., Salo, H. J., Danos, R. 2005. HST observations of spokes in Saturn's B ring. Icarus 173, 508–521.

Chapter 7

Understanding the Behavior of Prometheus and Pandora

Abstract

Goldreich & Rappaport (2003) discovered that the wayward motions of Prometheus and Pandora are attributable to the chaotic dynamics of their 121:118 mean motion resonance. While the observed system behavior has been reproduced in numerical integrations by these and other authors, a physical understanding of the origin of its features has been lacking. In this work we fill that gap. Specifically, we use the analogy with a parametric pendulum to demonstrate that the "kinks" observed in the satellite longitudes at apse antialignment are intervals of libration between stretches of circulation. In addition, we show that the system may display drastically different behavior in as little as 20 years. Changes in mean motions will not always occur at times of apse antialignment.

7.1 Introduction

Discovered by the *Voyager* spacecraft in 1980 and 1981, Prometheus and Pandora are two small moons of Saturn. Starting in 1995, a series of observations showed them to be in very different positions in their orbits than predicted using the *Voyager* data. Further, an abrupt change in their angular velocities was observed in 2000. The mystery of these discrepancies was solved in 2003 by Goldreich & Rappaport, who showed that the motions of Prometheus and Pandora could be understood in terms of mutual interactions via their 121:118 mean motion resonance. They found that this interaction results in chaos, and attributed the jump in mean motions to the stronger interactions that occurred when the orbits' apses were antialigned.

Prometheus and Pandora are back in the limelight because of the prospect of precision measurements by the *Cassini* spacecraft, which arrived at Saturn in July 2004 and will return data for at least 4 years. Measurements made with *Cassini* have the potential to enable further study of the system dynamics. There has been considerable recent work in predicting the orbital positions of Prometheus and Pandora, so that *Cassini* may be pointed in the correct direction to observe them. These studies have confirmed that the 121:118 resonance dominates the dynamics.

All of the abovementioned studies use numerical integrations to reproduce the observed features of the orbits, or to predict their behavior over the next 10 years or so. The mutual interactions of the satellites are now included in predictions for *Cassini*, but physical explanations for the mean motion jumps have not advanced beyond those proposed by Goldreich & Rappaport. According to these, the mean longitudes would continue to display a drift-jump behavior.

In this work we provide a physical understanding of the observed system behavior. We start by briefly reviewing both resonant orbital dynamics and chaos, in §7.2 and §7.3. The chaotic dynamics of a parametric pendulum are examined in §7.4. In §7.5 we describe the Prometheus-Pandora system in detail and review previous work in the area. Then we show that the system is analogous to a parametric pendulum. Using this analogy, we explain such features as the abrupt changes in mean motion. An understanding of the chaotic dynamics leads naturally to the prediction that the current drift-jump pattern of mean longitudes will not continue indefinitely, and may change in as little as 20 years. In §7.6 we conclude.

7.2 Resonant interactions

Here we provide a brief overview of resonant interactions in orbital dynamics. For more detail and derivations, see Murray & Dermott (1999).

Two orbiting bodies are in a mean motion resonance if the ratio n'/n of their mean motions is a rational number.[1]

Alternatively we can write[2] $\dot{\psi} = (p+q)n - pn' = 0$, or $\psi = (p+q)\lambda - p\lambda' = \text{constant}$, where p and q are integers, and where λ is the mean longitude, measured relative to an inertial line. Resonance implies that conjunctions[3] occur at a discrete number (q) of longitudes relative to periapse in the

[1] The mean motion $n = (GM/a^3)^{1/2}$.

[2] Neglecting in this paragraph the precession of the orbits.

[3] At intervals of $2\pi/|n - n'|$, the bodies reach conjuction, i.e., they are at the same longitude.

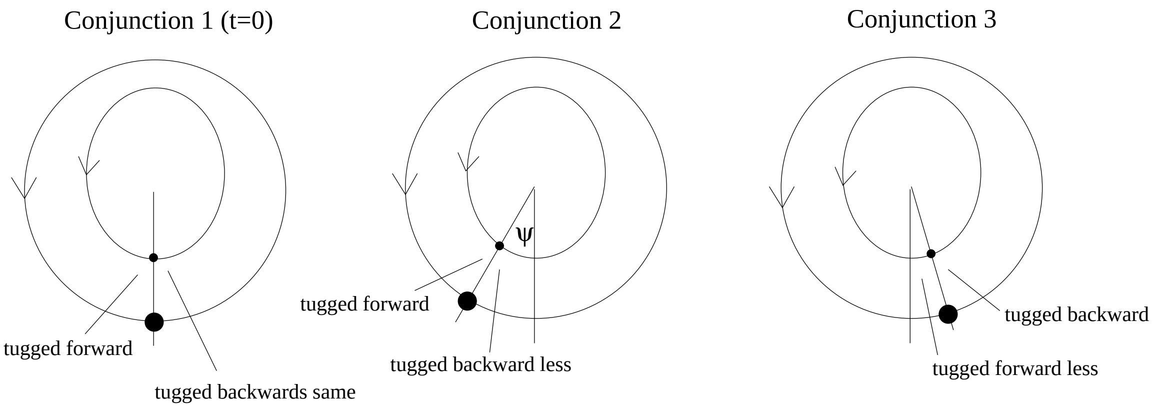

Figure 7.1: Near-resonant interactions as a pendulum. Consider the case in which a massive body orbits a central object on a circular orbit, and a test particle has an eccentric interior orbit. Let the bodies be near a 2:1 mean motion resonance, with the inner body going a little too fast, so that $\dot{\psi} = 2n - n' < 0$. If we start the bodies at conjunction at $\psi = 0$, then the next conjunction will occur earlier ($\psi < 0$) than it would for the perfect resonance. As the test particle sweeps by the massive body (being the particle with higher angular velocity), it first experiences a pull forwards, when the massive body is ahead of it in its orbit, and once it has passed, it experiences a backwards pull. Because the particles are further apart after the conjunction, the pull forwards is larger, and the net effect is to increase the angular momentum, and hence reduce the angular velocity, of the test particle, so that $\dot{\psi}$ is closer to zero than before. If the test particle is slowed sufficiently, then conjunction 3 (or later) will occur on the other side of periapse. In this case the net kick experienced by the test particle will speed it up, so that the longitude of conjunction is always being driven towards periapse. This libration of ψ around periapse is exactly analogous to the libration of a pendulum bob about its potential minimum. Analogous to a pendulum bob with high initial velocity, ψ can circulate around the whole orbit if the initial $\dot{\psi}$ is too large for the kicks to reverse its sign.

orbits of the bodies. Near resonances, bodies can exchange orbital energy and angular momentum in the same sense over many orbits, provided at least one of the orbits is eccentric or inclined. The direction of this exchange depends on the relation between the positions of conjunctions and the apsidal lines of the orbits.

The dynamics of ψ resemble those of a pendulum. If we are close enough to exact resonance ($\dot{\psi}$ small) then interactions can cause ψ to librate about the exact resonance, so the location of subsequent conjunctions oscillates about a fixed point (see Figure 7.1). Further from exact resonance, the angle ψ circulates, and the location of conjunction circulates around the orbit at a rate which is modified by the interactions.

Now consider the case of orbits that precess, at a rate $\dot{\varpi}$. We now have a choice of relations between $n, n', \dot{\varpi}$, and $\dot{\varpi}'$, which will give conjunctions in only a few discrete locations relative to periapse in the orbits. Consider the 2:1 resonance discussed in Figure 7.1, for the case in which both orbits are eccentric and precessing. We can have a resonance for which $\psi_1 = 2\lambda - \lambda' - \varpi =$ constant, corresponding to conjunctions occurring at the same place in one of the orbits. Alternatively, we could have $\psi_2 = 2\lambda - \lambda' - \varpi' =$ constant, when the conjunctions occur at the same place in the *other* body's orbit. We can derive these relations by simply moving into the precession frame of each orbit in turn, and consdering resonances in that frame. In general, if both orbits are eccentric, precession splits a $p + q : p$ resonance into $q + 1$ components.

Interactions at different mean motion resonances are described by Fourier coefficients C_i of the gravitational perturbation to the potential of the two satellites. The equations of motion for the system contain terms proportional to $C_i \sin \psi_i$. Far from resonance, $\dot{\psi}_i$ is large and so $\sin \psi_i$ oscillates rapidly, and the term with phase ψ_i can safely be neglected. Precession rates are typically much smaller than mean motions, i.e. $\dot{\varpi}, \dot{\varpi}' \ll n, n'$, so the splitting of each $p{+}q : p$ mean motion resonance into its individual components is small enough that all $q + 1$ components may require inclusion in the equations of motion. Splitting of resonances is the cause of chaos in the Prometheus-Pandora system.

7.3 Chaos—a simple introduction

The study of chaotic dynamics is extensive and well-documented in such texts as Tabor (1989) and Lichtenberg & Lieberman (1983). Chaotic dynamics in the solar system are reviewed by Wisdom (1987). Here we give only a brief overview of the topic, as is relevant to its application in this study.

Chaos in a system is manifested by extreme sensitivity to initial conditions: slightly different starting conditions lead, after some time, to vastly different phase space trajectories. Often trajectories exhibit what appear to be random changes. The divergence in phase space between two neighboring trajectories is exponential, with time constant $1/\ell$, where ℓ is the Lyapunov exponent.

An otherwise periodic system that experiences impulsive "kicks" may become chaotic. Suppose the system is kicked at phase ϕ_1. If the kick is small, then the motion is not much affected and the next kick will occur at a phase ϕ_2 that is essentially determined by the unperturbed motion. The phases at which kicks occur will be smoothly distributed throughout the period of the motion, and their effects will cancel out. The motion will be regular, not chaotic.

However, if the kick at ϕ_1 is such that it causes a large change in ϕ_2 (that is, if ϕ_2 is more than a radian from where it would have been had the first kick been small), then chaos will ensue. In a chaotic system there is extreme sensitivity to the place at which each kick occurs, and small deviations in the phase of the initial kick lead to the rapid separation of neighboring trajectories. This sensitivity appears as a randomness in the kicks: we cannot predict them exactly and so they appear random, although the system is completely deterministic. If we start two systems out with slightly different initial conditions, their trajectories will separate exponentially in phase space as they experience progressively more different kicks. Between kicks, their paths separate linearly, but at the kicks, there is an impulsive increase in their separation and rate of separation in phase space.

Chaotic behavior is observed in many different systems. Here we focus on one of the simplest and most common systems encountered in physics: the parametric pendulum. The “kicks” in this system will be phase delays near the separatrix.[4] Because the period is infinite on the separatrix, trajectories initially separated very slightly in phase space will typically become much more separated when they cross the separatrix. A pendulum which repeatedly passes through the separatrix will be chaotic. Separatrix crossings are induced by perturbations to the potential in which the pendulum moves, hence the parametric pendulum can be chaotic.

7.4 Pendulum equations

In the next section we show that the dynamics of the Prometheus-Pandora system is essentially that of a parametric pendulum.[5] The dynamical features are somewhat obscured in the Prometheus-Pandora case because the system is not very adiabatic, i.e., there is not a large separation between the period of small amplitude oscillations and the timescale of variation of the potential. It is

[4] The separatrix is the trajectory in phase space that separates libration from circulation. The oscillation period is infinite on this trajectory.

[5] A parametric pendulum is one that moves in a time-variable potential.

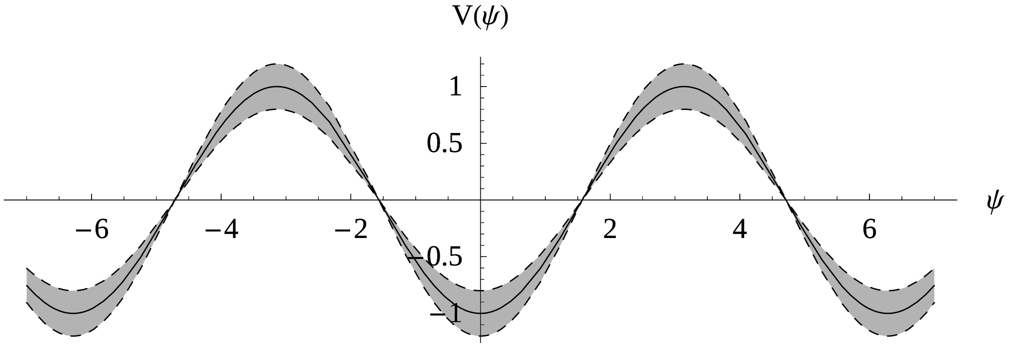

Figure 7.2: The potential $V(\psi)$ in which the particle moves, for $\alpha = 0.2$. Shaded area shows range over which potential varies.

instructive to first consider a parametric pendulum whose equation satisfies

$$\frac{d^2\psi}{dt^2} = -A(t)\omega_0^2 \sin\psi \tag{7.1}$$

where $A(t) = 1 + \alpha\cos(\kappa t)$. We choose $\alpha < 1$ and $\kappa \ll \omega_0$, so that for small oscillations the system is adiabatic, with slowness parameter $\epsilon \sim \alpha\kappa/\omega_0 \ll 1$. Without loss of generality, we now take $\omega_0 = 1$. The potential in which the system moves, given by

$$V(\psi, t) = -A(t)\cos\psi, \tag{7.2}$$

is illustrated in Figure 7.2.

7.4.0.1 Integration

The equation of motion (Eq. 7.1) is integrated in *Mathematica* using a Runge-Kutta formalism with fixed step size. The numerical accuracy of the routine is sufficient so that the divergences of trajectories is dominated by imposed differences in initial conditions and not by rounding errors.

7.4.1 System behavior and the separatrix

A pendulum can exhibit motion of two kinds, illustrated in Figure 7.3: libration when the energy $E < |V|$, and circulation when $E > |V|$ in which the system passes over the “hills” of the potential.

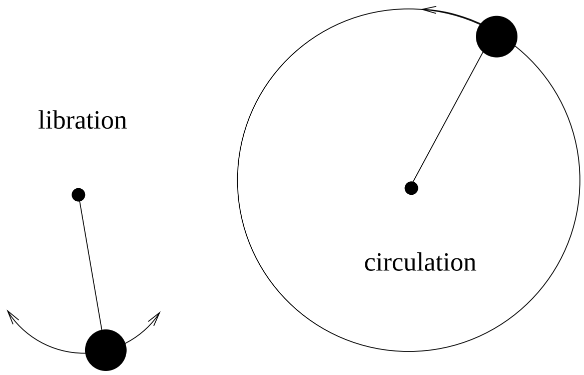

Figure 7.3: Libration and circulation of a pendulum

When $A(t)$ varies, the motion can switch between the two regimes. For example, starting with libration of large amplitude, if the potential becomes shallower then the system may be able to make it over the crest of the new lower potential hill and into circulation. If it does this, is has crossed the separatrix. A particle on the separatrix has just barely enough energy to make it to the crest of the hill, so it takes a very long time to do so. Motion on the separatrix formally has infinite period. It is this property that leads to separatrix crossings giving large enough phase delays to cause chaos.

7.4.2 Action and locating separatrix crossings

A system is on the separatrix when the energy

$$E(t) = \dot{\psi}^2/2 - A(t)\cos\psi \tag{7.3}$$

is exactly that required to make it to the top of the potential hills, i.e.,

$$E_{\text{sep}}(t) = |A(t)|. \tag{7.4}$$

In this we implicitly assume that the potential varies little over the period of the motion (i.e., $\epsilon \ll 1$), so that we treat the potential as fixed when considering whether the system is in libration or circulation.

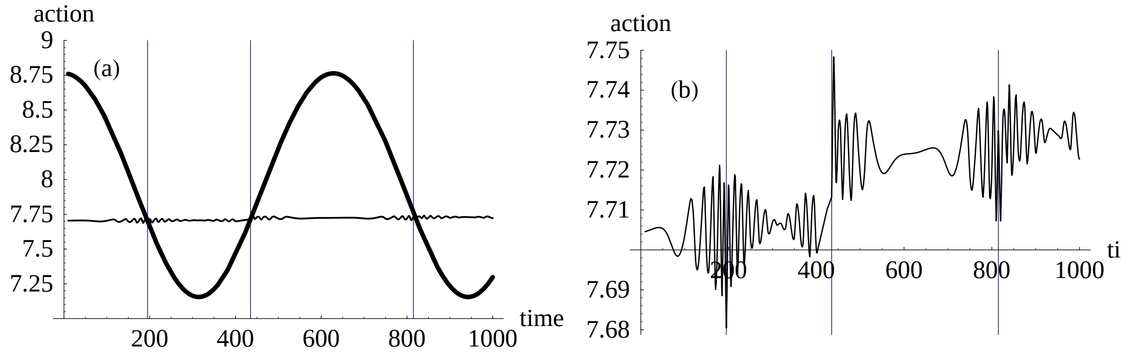

Figure 7.4: Action as a function of time for the system with $\alpha = 0.2$, $\kappa = 0.01$. (a) Thick line shows separatrix action, and thin line shows system action. (b) Zoom-in on system action, showing jumps in action at separatrix crossings.

It is more elegant to consider action rather than energy, because for adiabatic systems the action is constant as $A(t)$ varies. We calculate the action J by freezing the potential and integrating over a period.[6]

$$J(t) = \oint p\, dq = \oint \sqrt{2[E(t) - V(\psi')]}\, d\psi'. \tag{7.5}$$

We obtain the period in a similar way, from $P(t) = \oint dq/\dot{q}$.

To find the separatrix crossings, we compare the system action with the separatrix action;

$$J_{\text{sep}}(t) = 8A(t)^{1/2}. \tag{7.6}$$

This is shown in Figure 7.4 for $\kappa = 0.01$, $\alpha = 0.2$, i.e., slowness parameter $\epsilon = 0.002$. We see that J is constant over the timescales on which the potential varies, except at the separatrix crossings. This is to be expected: near the separatrix the system is not adiabatic, because the period is very long. We therefore do not expect the action to be conserved.

In Figure 7.5 we plot the period and the evolution of ψ as a function of time. Transitions between libration ($J < J_{\text{sep}}$) and circulation ($J > J_{\text{sep}}$) are evident. We also see that when $J = J_{\text{sep}}$, the period tends to infinity.

[6] When the system is in libration, we integrate over only half a period of the motion, because otherwise the action changes by a factor of 2 across the separatrix at fixed energy.

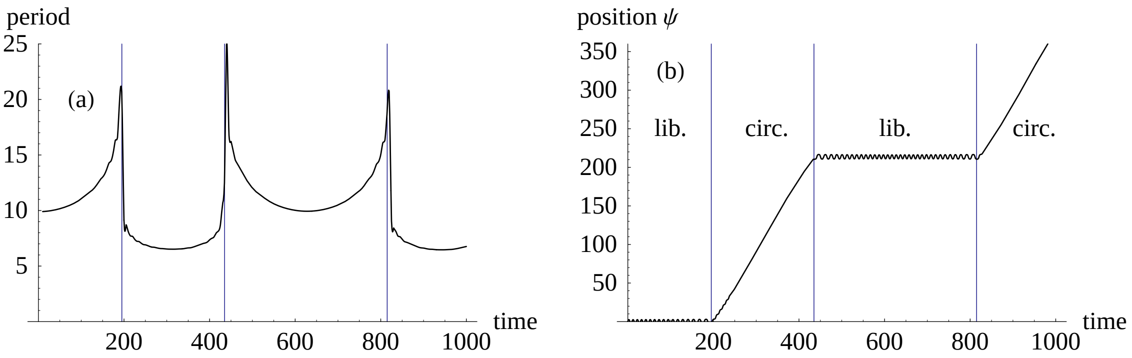

Figure 7.5: For the system with $\alpha = 0.2$, $\kappa = 0.01$: (a) Period as a function of time, clearly showing location of separatrixes. (b) Evolution of $\psi(t)$, showing transitions from libration to circulation and vice versa at separatrix crossings.

7.4.3 Separation in phase space and Lyapunov exponent

The separation of trajectories in phase space is of practical interest in chaotic systems because it limits our ability to make predictions about the future. Small errors in initial conditions rapidly amplify to the point at which predictions are useless.

We can track the separation of two neighboring trajectories by integrating them separately and calculating the distance between them. However, since their separation increases exponentially, at some point they can no longer be considered neighboring. To avoid this it is necessary to repeatedly decrease the phase space separation, because finite numerical accuracy prevents us from starting with orbits close enough that they will remain neighboring throughout the integration.

A better method is to Taylor expand the equation of motion to first order in the separation $\Delta\psi$:

$$\frac{d^2\Delta\psi}{dt^2} = -A(t)\omega_0^2\Delta\psi\cos\psi. \tag{7.7}$$

Because this is a linear equation in $\Delta\psi$, it describes the separation as though it is always infinitesimally small. The separation in phase space as a function of time

$$S(t) = \left[\Delta\psi^2 + \left(\frac{d\Delta\psi}{dt}\right)^2\right]^{1/2}. \tag{7.8}$$

is calculated by integrating Eq. 7.7 simultaneously with the equation of motion.

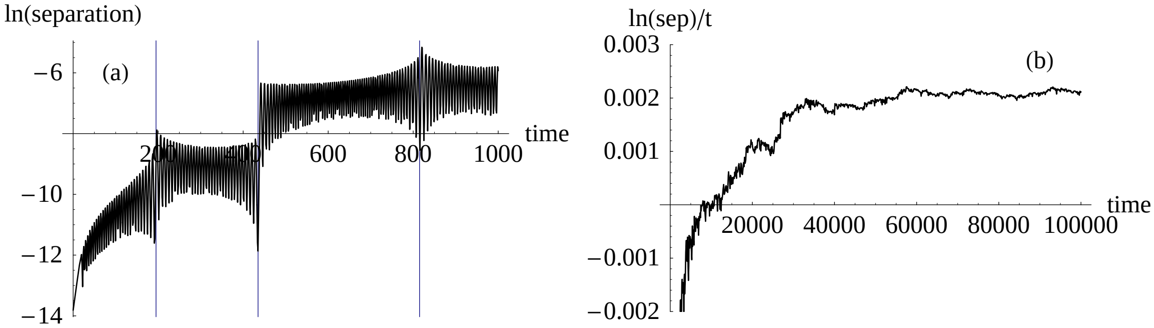

Figure 7.6: Separation of trajectories in phase space (a) The jumps upon separatrix crossing of phase space separation and rate of separation between two neighboring trajectories. (b) A longer integration gives the Lyapunov exponent, which asymptotes to about 0.0022.

The Lyapunov exponent is calculated by taking the limit at late times of $(\ln S)/t$, as shown in Figure 7.6. We find $\ell \simeq 0.0022$, which is roughly the rate of separatrix crossing. In other words, the separation between trajectories with different initial conditions roughly doubles every time a separatrix is crossed.

The Lyapunov exponent represents the mean divergence of trajectories over time, but we may also ask how this divergence takes place on shorter timescales. For this we refer to Figure 7.6a, showing the separation as a function of time through the first few separatrix crossings. At a separatrix crossing, the separation in phase space undergoes a sudden jump, corresponding to the phase delay introduced at the separatrix. After the crossing, the rate of linear separation of the trajectories is different from before. On average random jumps result in an increase of separation. The separation upon entering the next separatrix crossing is on average larger than it was the previous time, and so the size of the jump is larger (the fractional size of the jump is the same). When viewed on long timescales, the phase space separation undergoes an exponential increase. This increase is mediated by sudden jumps in S and $\dot{S}$ at separatrix crossings. The increase in S between crossings is linear with time (when averaged over the oscillations of the periodic motion, which introduce modulations because the periods are slightly different). This is important to remember for the Prometheus-Pandora case, which we consider shortly.

7.4.4 Excursions in the chaotic zone

As mentioned in §7.3, only trajectories that come close to the separatrix are chaotic. The range in initial action over which there is chaotic behavior is

$$\Delta J_{\mathrm{cz}} \simeq 8(A_{\mathrm{max}}^{1/2} - A_{\mathrm{min}}^{1/2}) \simeq 8\alpha. \tag{7.9}$$

If a system starts with action outside this zone, it never crosses the separatrix and its motion will be regular. A system that starts within the chaotic zone eventually explores the entire range of chaotic actions. We see in Figure 7.4 that at a separatrix there is a jump in the otherwise conserved action. These jumps are highly phase sensitive, and appear random. In time the system explores the whole chaotic zone (but does not leave it). Characteristic of all chaotic motion is uniform coverage of the phase plane in the chaotic region. This is known as ergodicity.

Jumps in action on separatrix crossings are of order the slowness parameter ϵ (Cary, Escande & Tennyson, 1986),

$$\frac{\Delta J_{\mathrm{jump}}}{J} \sim \epsilon \sim \alpha\frac{\kappa}{\omega}. \tag{7.10}$$

Separatrix crossings are spaced in time by $\Delta t \sim \pi/\kappa$ (about twice every potential variation cycle), so the time taken to cross the zone by a random walk is

$$t_c \sim \frac{\pi}{\kappa}\left(\frac{\Delta J_{cz}}{\Delta J_{\mathrm{jump}}}\right)^2 \sim \frac{\pi\omega^2}{\kappa^3}, \tag{7.11}$$

which is independent of α. We illustrate this in Figure 7.7, for which we have set $\kappa = 0.1$ and $\alpha = 0.2$, so that $\epsilon \sim 0.02$. We have decreased the adiabaticity of the example system in order to obtain a large change in action during a reasonable length of integration. The zone crossing time seen in Figure 7.7 is in good agreement with that predicted by Equation 7.11, $t_c \sim 3000$.

Jumps in action tend to be smaller towards the edges of chaotic zones, due to "sticking" in the more weakly chaotic boundary layer. Near the edges of the zone, the potential is near an extremum and so is varying more slowly than in the middle of the zone. The system is therefore more adiabatic

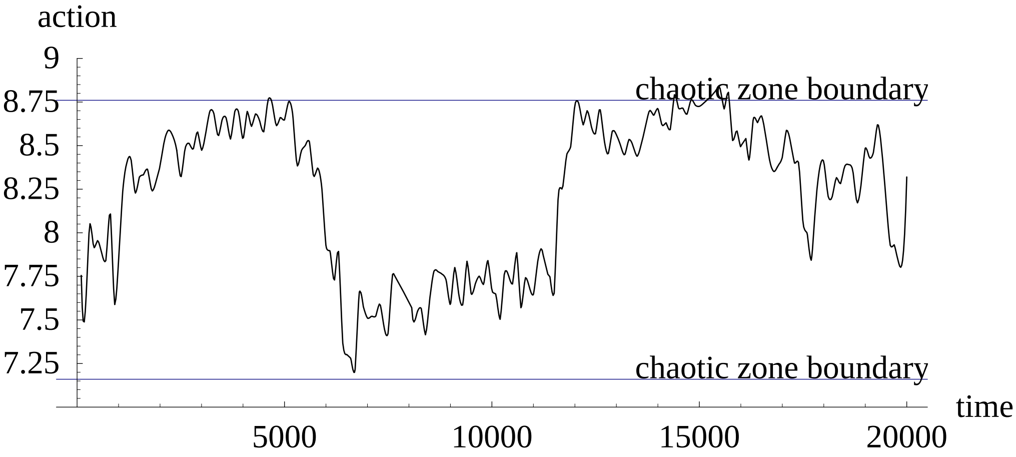

Figure 7.7: Wandering of action as a function of time, over about 600 separatrix crossings, for the case $\kappa = 0.1$, $\alpha = 0.2$.

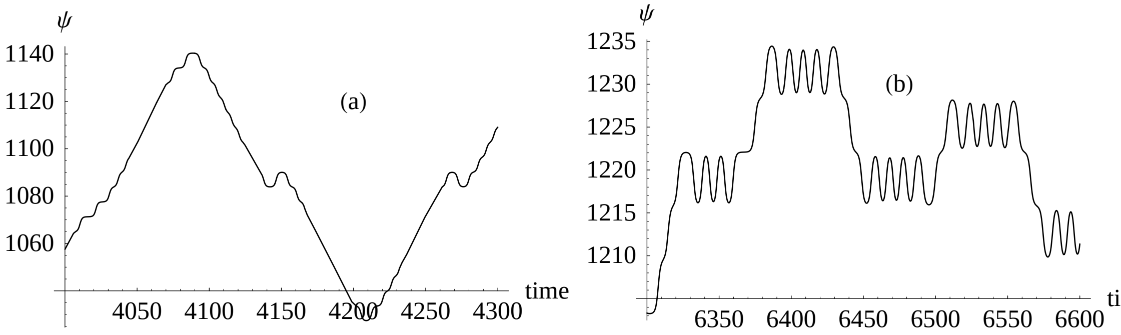

Figure 7.8: Behavior of the pendulum when close to the (a) top and (b) bottom of the chaotic zone, for the same system as in Figure 7.7.

and the size of the jumps in action is smaller (because $\Delta J/J \sim \epsilon$). However, on average the system spends no more time at the boundaries than at any other place in the zone.

Close to the lower boundary of the chaotic zone, the system spends almost all of the time in libration, and is only crossing into circulation when the potential reaches its lowest amplitude. Close to the top of the chaotic zone, the system is mostly in circulation. This is illustrated in Figure 7.8. We expect to find the system sampling each type of behavior.

7.5 Prometheus and Pandora

7.5.1 History

Saturn's narrow F ring (with semimajor axis 2.3 Saturn radii) is flanked by two small moons, Prometheus and Pandora. Discovered in 1980 and 1981 in *Voyager* images, they orbit Saturn in $\sim$ 15 hours. Their low-eccentricity orbits precess in $\sim$ 5 months due to the oblateness of Saturn. Synott et al. (1981, 1983) fitted a freely precessing ellipse to each orbit. Because their semimajor axes are slightly different, the orbits precess relative to each other with a period of $\sim$ 6.2 years. Properties of the satellites and their orbits are summarized in Table 7.1.

During the observational window of the 1995 ring plane crossing, measurements with *HST* showed that Prometheus lagged its expected position based on these fits by about 20 degrees (Bosh & Rivkin, 1996; Nicholson et al., 1996). Five years later, further *HST* observations showed that Pandora was leading its predicted position by a similar amount (McGhee et al., 2000). Furthermore, in 2000, a sudden change was observed in the rate at which the orbits drifted away from the *Voyager* predictions (Figure 7.9). This change occurred when the orbits' apses were antialigned, i.e., the precessional phase at which the bodies make their closest approaches (Figure 7.10).

Various suggestions were made to account for these discrepancies, including perturbations exerted by an undetected coorbital satellite of Prometheus (see French et al., 1998); interactions with clumps in the F ring, or 1 to 5 km objects in the F ring, or the F ring itself (Showalter et al., 1999a,b); long-term resonance dynamics (Dones et al., 1999); and chaos (Dones et al., 2001), before Goldreich & Rappaport (2003a; GR03a), demonstrated via numerical integrations that the deviations were produced by mutual gravitational interactions of the two satellites. Mutual interactions are strongly indicated because of the approximately equal and opposite drifts of the two satellites away from their predicted longitudes. The integrations in GR03a showed gradual drifts away from the *Voyager* orbits, with "kinks" in the drift rates at times of apse antialignment, as seen in Figure 7.9. GR03a further showed that these mutual interactions produce chaos with Lyapunov exponent 0.3 yr^{-1} (see §7.3). In Goldreich & Rappaport (2003b; GR03b), the essential dynamics of the system are captured

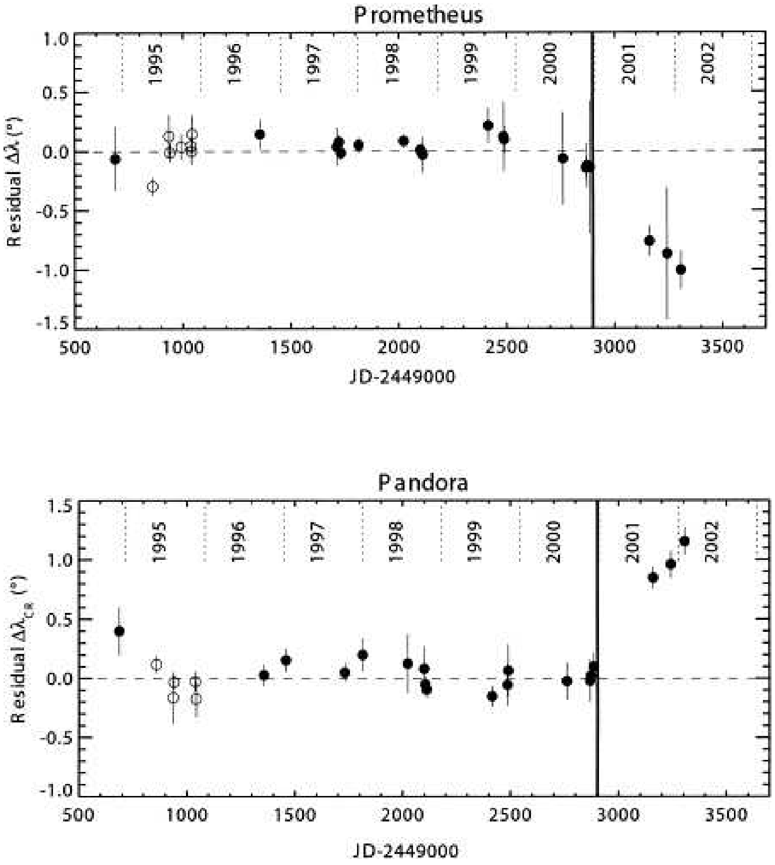

Figure 7.9: Sudden jumps in the positions of Prometheus and Pandora from French et al. (2002).

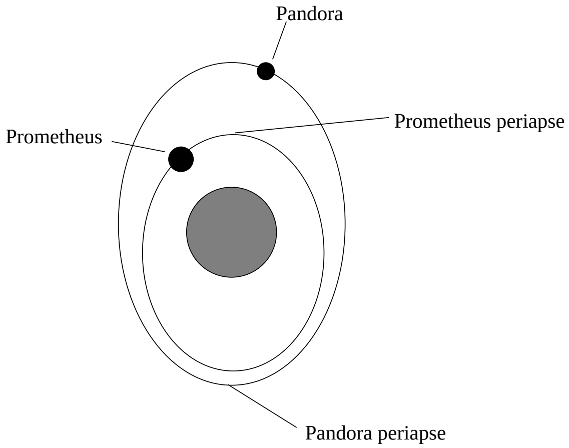

Figure 7.10: Apse antialignment (much exaggerated eccentricities and orbital separations) for Pandora and Prometheus; in reality, the orbits do not cross at any precessional phase.

by integrating a reduced set of equations that includes only interactions due to the 121:118 mean motion resonance. The chaos arises from the overlap of the four closely spaced components of this third order resonance. GR03b state that random mean motion jumps occur at apse antialignments, where the interactions are strongest.

Recently more complete numerical simulations of the system, which include the influences of other Saturnian satellites, have been performed by Cooper & Murray (2004); Jacobson & French (2004); and Renner, Sicardy & French (2005). These show that the simplified dynamics in GR03b is sufficient to describe the chaos in the orbits of Prometheus and Pandora. Our goal is to provide a physical understanding of this fascinating system. Such an understanding is missing from the literature.

All of the above papers cite interactions at apse antialignment as the reason for the kink in the drift rate observed in 2000. Orbital fits avoid the region around apse antialignment due to the stronger interactions that occur there, and it is assumed that between these times the orbits are well fitted by freely precessing ellipses.

In this paper we find that a proper understanding of the mutual interactions between Prometheus and Pandora is more subtle, but the system can be understood by analogy with the behavior of a parametric pendulum. The "kinks" arise during intervals of libration, separated by long stretches of circulation. Both the kinks and the chaos are caused by separatrix crossings.[7] While currently the two separatrix crossings per 6.2 year cycle happen near the position of apse antialignment, we show that this will not always be the case. In as little as 15 years the system may exhibit drastically different behavior.

7.5.2 Similarity of equations

GR03b showed that the mutual interactions of Prometheus and Pandora are well-represented by those of only their 121:118 mean motion resonance. Because of the differential precession of the two eccentric orbits, this resonance splits into four discrete resonances (see Section 7.2). We define the

[7] The frequency of separatrix crossings in relation to the magnitude of the Lyapunov exponent is mentioned in GR03b. Here we focus on the precessional phases at which the crossings occur.

Quantity	Prometheus	Pandora
Symbols	unprimed	primed
m/M	5.80×10^{-10}	3.43×10^{-10}
λ (°)	188.53815	82.14727
n (° s^{-1})	6.797331×10^{-3}	6.629506×10^{-3}
e	2.29×10^{-3}	4.37×10^{-3}
ϖ (°)	212.85385	68.22910
$\dot{\varpi}$ (° s^{-1})	3.1911×10^{-5}	3.0082×10^{-5}

Table 7.1: Properties of Prometheus and Pandora and of their orbits, from GR03b

Resonance q	Period (yr) $= -2\pi/\dot{\psi}_q$	Coefficient C_q
1	1.078	-1.08×10^{-3}
2	1.303	6.26×10^{-3}
3	1.648	-1.21×10^{-2}
4	2.239	7.82×10^{-3}

Table 7.2: Properties of the four components of the 121:118 mean motion resonance, from GR03b

term $\psi = 121\lambda - 118\lambda'$, where λ and λ' are respectively the mean longitudes of Prometheus and Pandora. The four possible resonant arguments are $\psi_q = \psi - \delta_q$, where

$$\delta_1 = 3\varpi,\ \delta_2 = 2\varpi + \varpi',\ \delta_3 = \varpi + 2\varpi',\ \delta_4 = 3\varpi'. \tag{7.12}$$

GR03b note that because of the rapid precession caused by Saturn's oblateness, interactions between the satellites produce negligible effects on their apsidal angles and orbital eccentricities. Therefore it is adequate to only retain changes in mean motions, or equivalently in the angle ψ.

The equation of motion for ψ reads (GR03b)

$$\frac{d^2\psi}{dt^2} = 3(121n')^2\frac{m}{M}\left[1 + \frac{am'}{a'm}\right]\sum_{q=1}^{4} C_q \sin(\psi - \delta_q), \tag{7.13}$$

where $M = 6 \times 10^{29}$ g is the mass of Saturn, and where the values of terms in Eq. 7.13 are given in Tables 7.1 and 7.2. Because the system is chaotic, it explores all accessible regimes for arbitrary initial data. We find it most convenient to adopt the initial data used by GR03b for easier comparison between their results and ours, although more recent data are of course available.

Eq. 7.13 can be written as

$$\frac{d^2\psi}{dt^2} = -\omega_0^2 A(t)\sin[\psi - \phi(t)], \tag{7.14}$$

at time t, where here $\omega_0 \simeq 3.8$ yr^{-1} and $|A| \leq 1$.[8] We may move to the frame drifting with the potential at $\dot{\phi}$, using $\Psi = \psi - \phi$, to obtain

$$\frac{d^2\Psi}{dt^2} = -\omega_0^2 A(t)\sin\Psi + \ddot{\phi}. \tag{7.15}$$

Provided $\ddot{\phi} \ll \omega_0^2 A$, we can drop the last term, leaving

$$\frac{d^2\Psi}{dt^2} = -\omega_0^2 A(t)\sin\Psi, \tag{7.16}$$

precisely the equation of motion of a parametric pendulum. In this case, the potential is given by

$$V(\Psi, t) = -\omega_0^2 A(t)\cos\Psi. \tag{7.17}$$

Plots of V, ϕ, $\dot{\phi}$ and $\ddot{\phi}$ are displayed in Figure 7.11. The timescale for variation of $|V(t)|$ is 6.2 years $= 2\pi/|\dot{\varpi} - \dot{\varpi}'|$, and the fractional variation in its amplitude is of order unity. The typical oscillation period $\omega \sim 2\pi/\bar{V}^{1/2}$ of the system is on the order of 2 years. The slowness parameter is then $\epsilon \sim 1 \times 2/6.2 \simeq 0.3$. This is a far less adiabatic system than that considered in the previous section, so the "frozen potential" approximation is less valid here. In addition, $\ddot{\phi} \neq 0$, but in our analysis we approximate $\dot{\phi}$ to be constant at -50.7 yr^{-1} when we calculate the action. We pursue the pendulum analogy studied in §7.4 because it can take us a long way toward understanding the behavior of the system. We return to the validity of this treatment in §7.5.4. It is important to note that although much of our analysis rests on the pendulum analogy, we integrate the full equation of motion 7.13 and not the approximate version Eq. 7.16. Thus the behavior $\psi(t)$ we find does not depend on the analogy. Eq. 7.13 is integrated using the same method as described in §7.4.0.1,

[8] Equivalently, the potential for the pendulum in the previous section can be written as a sum of resonant terms: $V(\psi, t) = -A(t)\omega_0^2\cos\psi = -\omega_0^2\cos\psi - (\alpha\omega_0^2/2)\,[\cos(\psi - \kappa t) + \cos(\psi + \kappa t)]$.

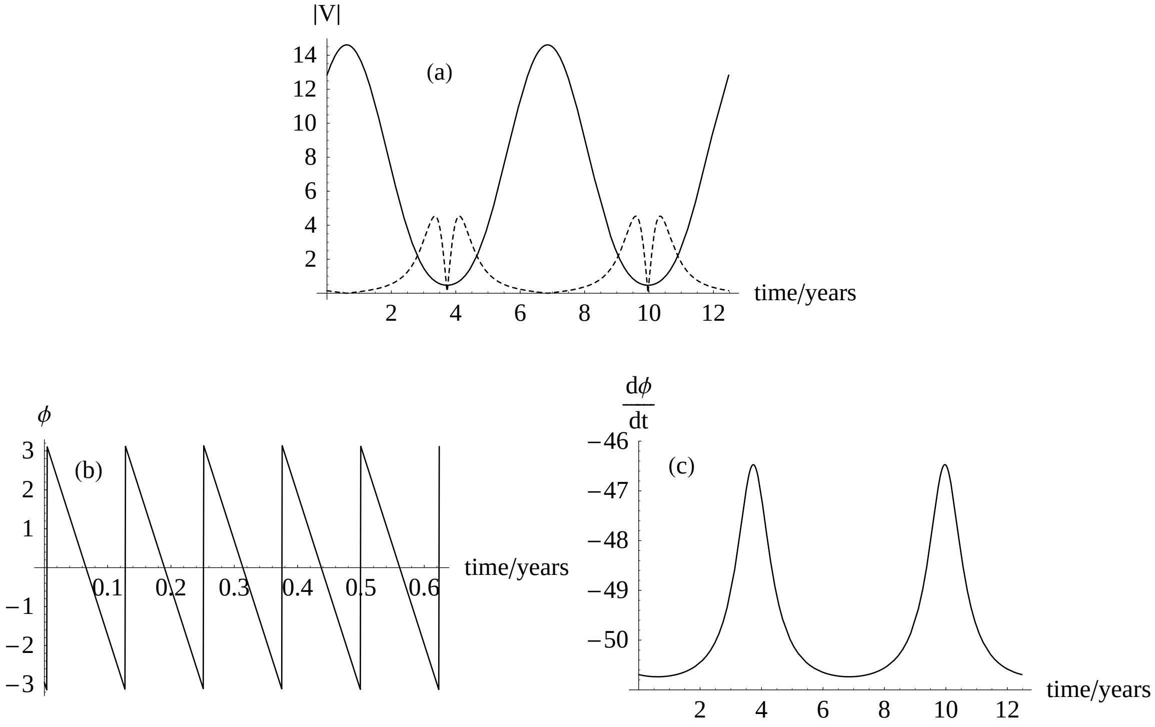

Figure 7.11: Properties of the potential in which the Prometheus-Pandora system moves: (a) solid line: the amplitude of the potential as a function of time; dashed line: $\ddot{\phi}$, (b) Drift phase ϕ as a function of time, (c) Rate of change of the drift phase: this is the drift rate of the hills in the potential. In most of this study we ignore the variation of $\dot{\phi}$, since it is localized to a small fraction of the period.

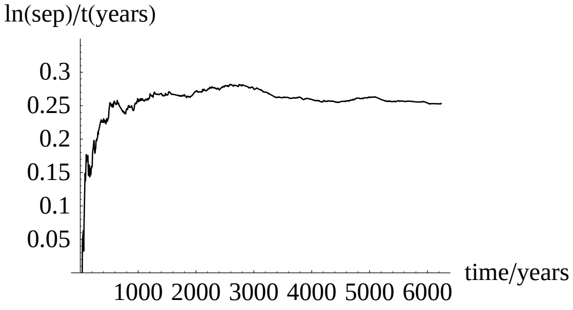

Figure 7.12: Determination of the Lyapunov exponent for the Prometheus-Pandora system.

integrating in parallel the linear term in the phase space separation, analogous to Eq. 7.7. A long integration finds a Lyapunov exponent $\ell \sim 0.3$ yr^{-1}, in agreement with that found by other authors (cf. Figure 7.12).

7.5.3 Similarity of behavior

In Figure 7.13 we plot the behavior of ψ over the first 12 years of integration. This bears a striking resemblance to Figure 7.5, in which the parametric pendulum was switching between libration and circulation, only here the oscillations are less adiabatic. When we compute the action using the same method as in the previous section, and compare it with the separatrix action, we see in Figure 7.13 good agreement between the times of separatrix crossing and the times of switching between the two regimes of motion. As expected, the action is not as constant between separatrix crossings, due to the reduced adiabaticity of the system.

7.5.3.1 Where separatrix crossings occur

GR03ab and subsequent authors state that kinks in the mean motions occur near apse antialignment. Apse antialignment corresponds to the time of maximum amplitude of the potential. We see in Figure 7.13 that separatrix crossings currently occur in that region. A similar behavior is seen in Figure 7.8a, in which the action is close to its maximum value in the chaotic zone. The kinks in mean motion correspond to short intervals of time spent in libration amidst long stretches of circulation

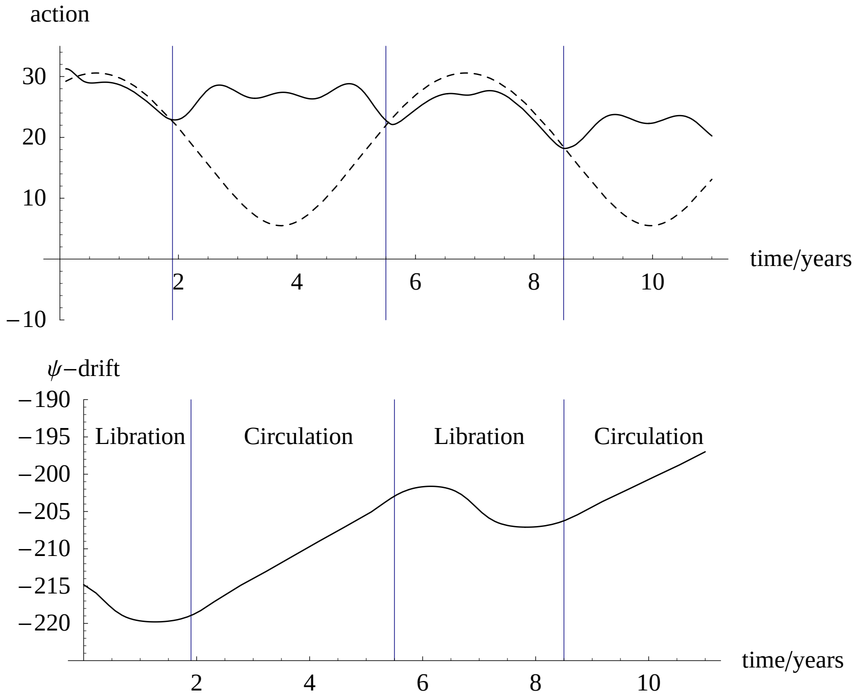

Figure 7.13: Action and ψ for the Prometheus-Pandora system for first 12 years of integration, showing changes in regime of motion at separatrix crossings, indicated by vertical lines.

during which ψ increases monotonically. Subsequent episodes of circulation have different slopes $\dot{\psi}$ because the system re-enters circulation with an altered action, due to the intervening separatrix crossings. The kinks are not *due* to chaos; rather they are due to the separatrix crossings, which *also* give rise to the chaos. We can predict the changes in the circulation rate, but only with an error that increases at every separatrix crossing.

A chaotic system explores the entire chaotic zone in phase space, so we expect to see the system as often in a state like that of Figure 7.8b as like that of 7.8a. Thus, averaged over time, separatrix crossings are not restricted to any particular precessional phase. Our integrations confirm this expectation. As Figure 7.14 shows, the system experiences states in which it is mostly in circulation and mostly in libration. Separatrix crossings, accompanied by sudden changes in the behavior of $\psi(t)$, happen at all precessional cycles.

A long integration confirms that the action wanders throughout the chaotic zone (Figure 7.15, cf. also Figure 7.7).

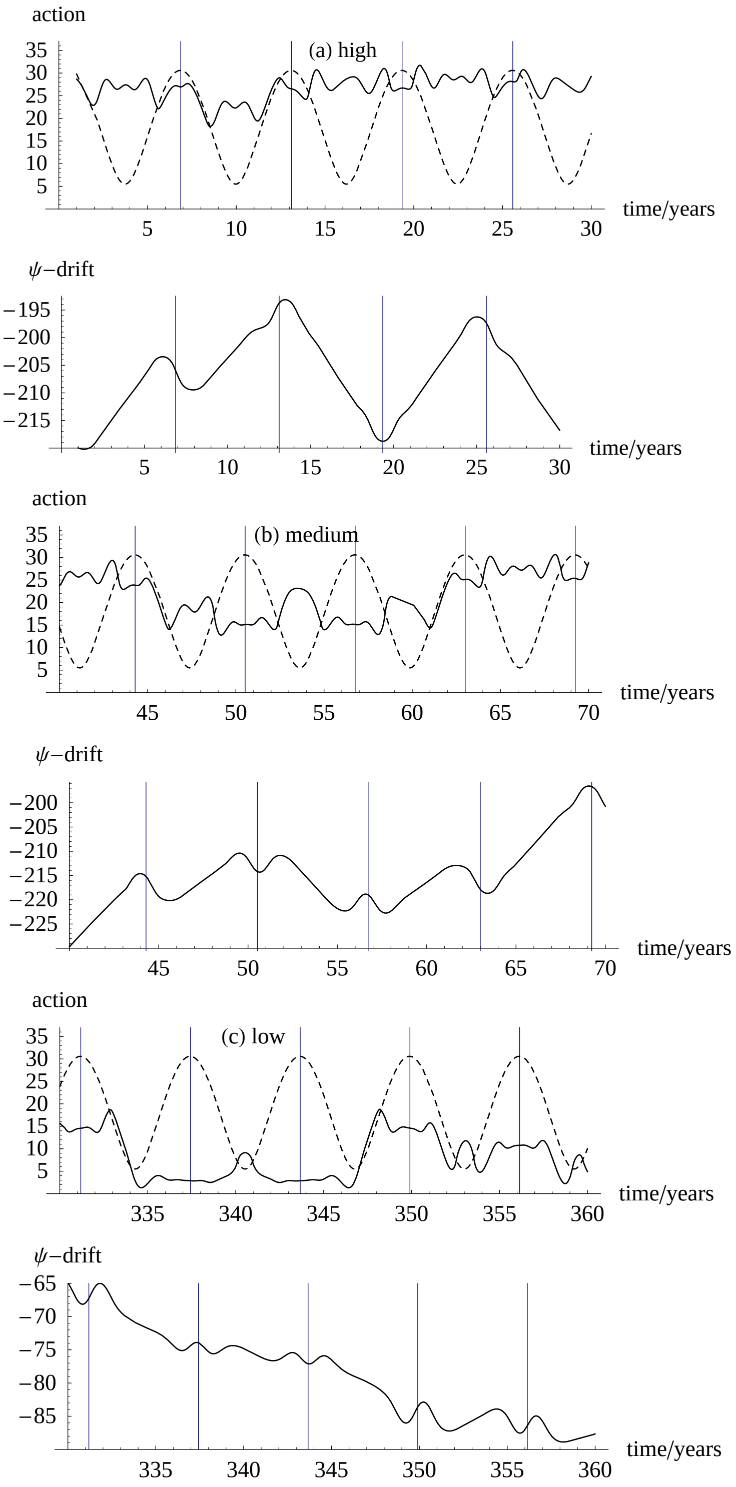

Figure 7.14: Examples of action high, medium, and low, showing separatrix crossings in different places in these three cases. The vertical lines indicate the times of apse antialignments, showing that separatrix crossings do not always occur there.

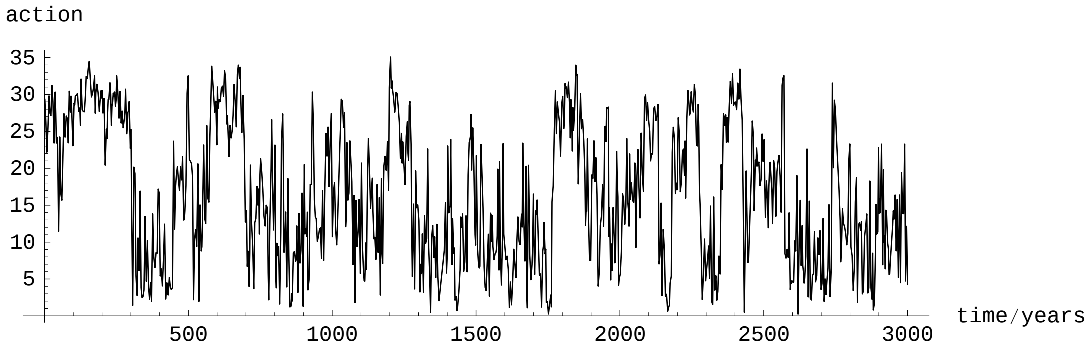

Figure 7.15: Wandering of the action of the Prometheus-Pandora system: this clearly delineates the chaotic zone. Timescale to cross the zone is on the order of 30 years.

Eq. 7.11 predicts that it takes ~ 30 years for the action to diffuse across the chaotic zone, as seen in Figure 7.15. This is an underestimate for our particular choice of initial conditions, because of "sticking" at the top of the chaotic zone (discussed in §7.4.4).

One can fit a variety of orbits to the data, because of measurement errors. Renner et al. (2005) make modern estimates of satellite positions as a function of time, and find that after two separatrix crossings (their 2004 point), there is an uncertainty of 0.2° in the longitude of each satellite.[9]

We start from this level of uncertainty in position, holding all other orbital elements fixed, and we integrate forward in time a set of 3 trajectories spanning the error range. The results are shown in Figure 7.16, in which the action for each trajectory is plotted against the separatrix action. In only 15 years from the start of our integration, one of the trajectories has already transitioned from a state of high to low action. This illustrates that only a small uncertainty in positions can rapidly amplify to give a qualitative change in system behavior.

7.5.4 Limitations of the analogy

The Prometheus-Pandora system is not strongly adiabatic, since only about 3 periods of oscillation take place during the modulation of the potential. Because the amplitude of the potential varies substantially over the period of motion, we expect our frozen potential assumption to lead to uncertainties of order half a period of the motion in the positioning of the separatrix crossings. However, we have seen that the crossing positions are adequately located by this method.

[9] This is about the same as the size of a single *HST* error box at Saturn.

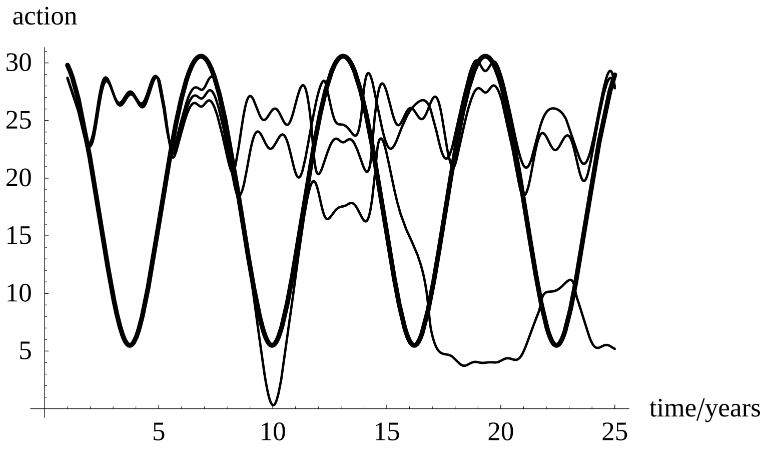

Figure 7.16: Divergence in action of trajectories starting with 0.2° uncertainty in longitudes

A more serious problem is the fact that we do not fulfill the condition $\ddot{\phi} \ll V(t)$ throughout all of the precessional cycle. The dashed line plotted in Figure 7.11 shows the magnitude of the $\ddot{\phi}$ term in comparison with $|V(t)| = |\partial V/\partial\psi|$. For about one third of the time (around apse *alignment*), we do not fulfill this condition, and the pendulum equation does not describe the system well at all. During these times, the system is essentially governed by $\ddot{\Psi} \simeq -\ddot{\phi}$, so Ψ appears to circulate. Perhaps then we cannot define a separatrix crossing near apse alignment. We do find some limited evidence (not described in this chapter) that the jumps in separation of trajectories are larger near apse antialignment than those that occur near apse alignment. Undeniably however, changes in the evolution of $\psi(t)$ occur throughout the precessional cycle (Figure 7.14), and this statement is independent of the pendulum analogy, since it is based on integrations of the full equation of motion, Eq. 7.13.

To progress further, we need to look for the analog of a separatrix for a system that is not oscillating: we need to find the places at which large time lags occur, whether this is by finding the distribution of jumps in separation with time, or by finding places at which the system nearly "stops" at the top of a potential hill. This could form the basis of a future study.

The abovementioned discrepancy in the pendulum analogy is not general to systems of multiple resonances: the term $\ddot{\psi}$ depends on the precession rates of the orbits, while the size of the term

$\omega_0^2 A(t)$ depends on the resonance strengths, though its modulation has a dependence on the relative phases of the resonant terms.

7.6 Summary and conclusions

The equations of motion of the Prometheus-Pandora resonant system can be well represented by those of a parametric pendulum. Despite the low degree of adiabaticity of the Prometheus-Pandora system, the same regimes of behavior exist as found in a more adiabatic model system. Using this analogy, we have explained the nature of the “kinks” in mean motion seen in both observations and simulations as short episodes of libration between long stretches of circulation. We caution that the system will not continue indefinitely to display periods of drift with intervening kinks: in as little as 20 years, the evolution of the mean motions could be drastically different. Sudden changes in behavior (and increased uncertainties in predictions) can happen far from times of apse antialignment.

Bibliography

Bosh, A. S., Rivkin, A. S., 1996. Observations of Saturn's inner satellites during the May 1995 ring-plane crossing. Science, 272, 518–521.

Cary, J. R., Escande, D. F., Tennyson, J. L., 1986. Adiabatic-invariant change due to separatrix crossing. Phys Rev A, 34, 4256–4276.

Cooper, N. J., Murray, C. D., 2004. Dynamical influences on the orbits of Prometheus and Pandora. AJ, 127, 1204–1217.

Dones, L., Showalter, M. R., French, R. G., Lissauer, J. J., 1999. The perils of Pandora. Bull Am Astron Soc, 31, 09.03 (abstract).

Dones, L., Levison, H. F., Lissauer, J. J., French, R. G., McGhee, C. A., 2001. Saturn's coupled companions, Prometheus and Pandora. Bull Am Astron Soc, 33, 29.06 (abstract).

French, R. G., Hall, K. J., McGhee, C. A., Nicholson, P. D., Cuzzi, J., Dones, L., Lissauer, J., 1998. The perigrinations of Prometheus. Bull Am Astron Soc, 30, 02.04 (abstract).

French, R. G., McGhee, C. A., Dones, L., Lissauer, J. J., 2002. Saturn's wayward shepherds: the perigrinations of Prometheus and Pandora. Icarus, 62, 144–171.

Goldreich, P., Rappaport, N., 2003. Origin of chaos in the Prometheus-Pandora system. Icarus, 166, 320–327, GR03b.

Goldreich, P., Rappaport, N., 2003. Chaotic motions of Prometheus and Pandora. Icarus, 162, 391–399, GR03a.

Jacobson, R. A., French, R. G., 2004. Orbits and masses of Saturn's coorbital and F-ring shepherding satellites. Icarus, 172, 382–387.

Lichtenberg, A. J., Lieberman, M. A., 1983. Regular and stochastic motion, Springer-Verlag, New York.

McGhee, C. A., 2000. Comet Shoemaker-Levy's 1994 collision with Jupiter and Saturn's ring plane crossing. Ph.D. dissertation, Cornell University, Ithaca, NY.

Nicholson, P. D., Showalter, M. R., Dones, L., French, R. G., Larson, S. M., Lissauer, J. J., McGhee, C. A., Seitzer, P., Sicardy, B., Danielson, G. E., 1996. Observations of Saturn's ring-plane crossings in August and November 1995. Science, 272, 509–515.

Renner, S., Sicardy, B., French, R. G., 2005. Prometheus and Pandora: masses and orbital positions during the Cassini tour. Icarus, 174, 230–240.

Showalter, M. R., Dones, L., Lissauer, J. J., 1999a. Interactions between Prometheus and the F ring. Bull Am Astron Soc, 31, 09.02 (abstract).

Showalter, M. R., Dones, L., Lissauer, J. J., 1999a. Revenge of the sheep: effects of Saturn's F ring on the orbit of Prometheus. Bull Am Astron Soc, 31, 44.08 (abstract).

Synott, S. P., Peters, C. F., Smith, B. A., Morabito, L. A., 1981. Orbits of the small satellites of Saturn. Science, 212, 191–192.

Synott, S. P., Terrile, R. J., Jacobson, R. A., Smith, B. A., 1983. Orbits of Saturn's F ring and its shepherding satellites. Icarus, 53, 156–158.

Tabor, M., 1989. Chaos and integrability in nonlinear dynamics, Wiley, US.

Wisdom, J., 1987. Urey prize lecture: chaotic dynamics in the solar system. Icarus, 72, 241–275.