Black Holes and Accretion Disks in Active Galactic Nuclei: Microlensing, Caustics, and Collisional Stellar Dynamics

Thesis by

Kevin Patrick Rauch

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1995

(Submitted August 31, 1994)

"The Road goes ever on and on
Down from the door where it began.
Now far ahead the Road has gone,
And I must follow, if I can,
Pursuing it with eager feet,
Until it joins some larger way
Where many paths and errands meet.
And whither then? I cannot say..."

—B. Baggins

Acknowledgements

It's funny how quickly a year can go by—or four. I guess time is relative after all. As I write this it's difficult for me to think back to when I first arrived here; the memory seems quite distant and highly demagnified. But I believe I've learned more here than I realize—though not all that Caltech has to teach. Oh well, Roger can't afford to keep me here *that* long...

I first must thank my mom and dad for letting me pursue my own goals even when they don't really know precisely what it is I do. That's OK, some days *I* can't even figure out what I'm doing... ☺ To Wal, thanks for the Southern Comfort—better aged than never! AJ, it was nice having you in the area; I'll miss the biannual camping trips. But why did it always rain when you drove?? I'd also like to thank the process of plate tectonics for creating mountains—I can't imagine existence without them any more (though I may be reminded awfully soon). Thanks also to Melvyn for computing the SPH models for me—but not for consuming all available disk space in the process!

I am, needless to say, most indebted to my advisor, Roger Blandford, who I believe has forgotten more good ideas than I may ever think of.

To Bettina, for Vol. I.

Abstract

Interactions between and the structure of black holes, accretion disks, and dense star clusters are investigated. Observed rapid gravitational microlensing variability in the quasar Q2237+0305 is used in conjunction with numerical simulations of microlensed quasar accretion disks to determine whether the observations constrain theoretical accretion disk models. It is found that blackbody disks are at least three times too large to account for the observed variability, and on that basis it is argued that the optical emission is either nonthermal or optically thin.

Accurate, efficient, and general-purpose routines to compute geodesic trajectories in the Kerr spacetime describing rotating black holes are implemented and applied to several problems. The optical caustic structure of the Kerr metric describing rotating black holes is determined and its possible relevance to rapid X-ray variability in active galactic nuclei is discussed. It is found that the (primary) caustic is a small tube with an astroid cross section which extends behind the black hole asymptotically parallel to the optic axis but displaced from it by an amount proportional to the spin of the hole, and that the angular magnification is unexpectedly high everywhere inside the caustic. Sample point source light curves and the appearance of thick accretion disks around Kerr black holes are calculated and the influence of caustics on them is assessed.

The dynamical evolution of the core of a dense star cluster around a Kerr black hole and under the influence of star-disk interactions is examined. It is shown that there are astrophysically plausible regimes in which star-disk interactions can dominate all other dynamical processes. The effects of star-disk interactions on single orbits are illustrated. It is found that star-disk interactions steepen the initial density profile towards an equilibrium r^{-3} profile and simultaneously increase the central density by up to two orders of magnitude. It is argued that this process could self-limit when densities climb to such a level that collisions between stars become important.

Simulations of the dynamical evolution of the density cusp of a star cluster around a massive black hole in a regime where stellar collisions dominate other dynamical processes are performed. The calculations are done using a discrete cluster of stars and a fully relativistic formalism. Versatile numerical methods are developed and applied to this problem. A modified form of Kepler's Equation asymptotically valid in the Kerr geometry is derived. It is found that collisions produce a constant density core which is mainly populated by stars on highly radial orbits, in contrast to previous Fokker-Planck analyses in which an $r^{-1/2}$ profile has been found. Collisional refilling of the loss cone is seen. Additional applications of the numerical algorithms are suggested.

Table of Contents

1

Introduction

One of the charms of astrophysics, I find, is in its ability to intertwine an enormous range of physical scales into the solution of a single problem, such as using the physics of nuclear reactions to explain the heavy element abundance of the interstellar medium. One of the investigations presented in this thesis (Chapter 2) contains a similar spirit; the predominant motif, however, is one of interaction—specifically, the interaction of three probable components of active galactic nuclei: massive black holes, accretion disks, and dense nuclear star clusters. The following chapter uses observations of gravitational microlensing occurring on cosmological scales together with the results of simulations to determine the extent to which accretion disk models of active galactic nuclei (AGNs) can be constrained by such observations. Attention then shifts to the massive black holes (MBHs) presumed to be the central engines of AGNs and to investigations of the optical structure of the spacetime around a rotating (Kerr) black hole and how it could influence observed AGN emission, followed by studies of the dynamical evolution of dense star clusters around MBHs, in which evolution is dominated either by interactions with an accretion disk or by physical collisions between the stars themselves. Particular attention is paid to the influence and importance of uniquely relativistic effects from Kerr black holes, and considerable (as well as versatile) numerical machinery is developed to allow accurate calculation of these effects. The present chapter reviews the basic issues important to the analysis.

1.1. MICROLENSING: INTRINSIC AND EXTRINSIC

The idea that gravity should bend light rays just as it does the trajectories matter can be traced back nearly three centuries to Newton (1704) himself. The

issue was not considered seriously, however, until after the publication of General Relativity by Einstein (1915). Einstein's prediction that light rays grazing the solar limb should be deflected by $\approx 1.75''$ was at least crudely confirmed by Eddington's (1919) observations of stellar positions near the sun during the total solar eclipse of 1919. Ensuing years saw elaboration of the basic theory, including the realization that gravitational lensing could produce multiple images of a single object (Eddington 1920; Einstein 1936). Zwicky (1937) argued that lensing of distant galaxies should be detectable. It was not until 1979, however, that theory and observation finally met, when the chance discovery of the doubly-imaged quasar 0957+561 was announced (Walsh, Carswell, & Weyman 1979) and in which the two images, separated by $\sim 6''$, are the result of lensing by an observable foreground galaxy. The field has expanded rapidly since then, both theoretically and observationally, and today there are a few dozen secure or proposed lenses, including, in addition to multiply-imaged quasars, the newer classes of radio rings (Hewitt et al. 1988; Langston et al. 1989) and luminous arcs (Soucail et al. 1987; Lynds & Petrosian 1989), both of which result when an extended background source is nearly collinear with the observer and lens. Recent reviews of this subject can be found in Blandford and Narayan (1992) and Refsdal and Surdej (1994); a detailed treatment of gravitational lens theory is given in the monograph by Schneider, Ehlers, and Falco (1992).

The qualitative image characteristics of most observed lenses can be understood in terms of a very simple lens model in which the deflector is approximated by a point mass. In this case the lens equation determining the apparent image positions can be written as

$$\theta^2 - \phi\theta - \theta_{\rm E}^2 = 0, \tag{1.1}$$

where θ is the apparent angular separation of an image from the optic axis (the line connecting the observer and deflector), ϕ is the true separation of the source from the axis, and $\theta_{\rm E}$ is the so-called Einstein ring radius given by

$$\theta_{\rm E} = \left(\frac{4GM}{c^2 D}\right)^{1/2} \approx 3 \left(\frac{M}{10^{12} M_\odot}\right)^{1/2} \left(\frac{D}{1\ {\rm Gpc}}\right)^{-1/2} {\rm arcsec}, \tag{1.2}$$

where $D = D(z_{\rm L}, z_{\rm S})$ is an effective lens distance depending on the lens and source redshifts and on the assumed cosmological model. Thus for off-axis sources, $\phi \neq 0$, two images are produced by the lens, one on each side, while for on-axis sources there is axisymmetry about the optic axis and a ring image (of radius the Einstein ring radius) is produced. If an extended source lies only slightly off-axis, two arcs will be produced. Viewing the lens equation as a (multi-valued) mapping $\phi \rightarrow \theta$, it is easy to see that the magnification of an image is given by the Jacobian of the transformation (in *solid angle*),

$$\mathcal{M} = \left| \frac{\partial \vec{\Omega}_{\rm src}}{\partial \vec{\Omega}_{\rm obs}} \right|^{-1} = \left| \frac{\phi \, d\phi}{\theta \, d\theta} \right|^{-1} = \left| 1 - \left(\frac{\theta_{\rm E}}{\theta} \right)^4 \right|^{-1} . \tag{1.3}$$

Thus a ring or arc image (which has $\theta \approx \theta_{\rm E}$) suffers high magnification. Note that $\theta_{\rm E}$ acts as an effective 'size' of the lens: for $\phi \lesssim \theta_{\rm E}$ the two images have similar magnifications and separations from the optic axis while for $\phi \gg \theta_{\rm E}$ there is one nearly unaltered image at $\theta \approx \phi$ for which $\mathcal{M} \approx 1$ with the other image at $\theta \ll \theta_{\rm E}$ having $\mathcal{M} \ll 1$.

Gravitational microlensing (Chang & Refsdal 1979; Young 1981; Paczyński 1986) is a variation of the basic lensing phenomenon which can occur when the distribution of mass within the deflecting object (normally a galaxy or cluster of galaxies for cosmological lenses) is inhomogeneous on small scales, as in a galaxy composed of individual stars, and when the lensed object is sufficiently small (roughly speaking, when its angular size is less than the ring radius of the individual stars). In this situation the lensed image seen telescopically (the 'macroimage') actually consists of very many unresolved microimages, which are the multiple images produced by each individual mass concentration in the deflector; the characteristic separation of these microimages is the ring radius, which for cosmologically distant sources and stellar mass deflectors is $\theta_{\rm E} \sim 10^{-6}$ arcsec. It may seem at first that the existence of microlensing is of only theoretical interest, since individual microimages have no chance of being resolved optically in the foreseeable future, but their existence can be indirectly inferred (and directly utilized) through the

quite observable *variations* in intensity that microlensing can produce. Suppose for the moment that the deflector is a galaxy whose mass is mainly in the form of stars. Microlensing-induced changes in macroimage intensity occur when there is transverse motion between the background source and the lensing galaxy, which over time alters the particular configuration of stars producing the microimages. As the star field slowly changes, some microimages will disappear, others will be created, and all will suffer some change in magnification. Thus the aggregate macroimage will appear to vary even if the source itself does not, just as atmospheric inhomogeneities cause stars to twinkle. Note that this is more than a passing analogy—the gravitational lensing equation (in the weak field limit) can be written in a form identical to that appropriate for physical optics by defining an effective index of refraction $n = 1 - 2\Phi/c^2$, where Φ is the (Newtonian) gravitational potential of the lens. This type of microlensing can be termed 'extrinsic' because the lensing objects lie outside the source and are unrelated to it; the motivation for making this distinction will become clear below.

An important concept for the theoretical understanding of microlensing (as well as macrolensing) is that of *caustics*. In precise analogy to physical optics, the caustics of any gravitational lens are the continuity of source positions for which the observer sees the (point) source infinitely magnified, in the geometric optics limit. Isolated caustics can be classified using catastrophe theory into types such as folds, cusps, and swallowtails according to the severity of the catastrophe (e.g., Berry & Upstill 1980). One fundamental fact of interest is that the number of microimages changes by two whenever a source crosses a caustic surface, with a pair of images being either created or destroyed at the interface. In real situations, of course, finite source sizes and photon wavelengths allow only finite maximum magnifications, although those magnifications can still be quite large. The utility of caustics in the present case is the ease with which microlensing variations can be understood using this framework. Imagining that the source motion lies in a plane, it is clear that for each position in this source plane the observed macroimage will suffer some net magnification as seen by an observer, which will change as the

source position changes. Abstractly one can imagine magnification contours being superimposed on the source plane, and in this way the apparent magnification of any extended source lying anywhere in the source plane is easily computed by convolving the source profile with the magnification map. The caustics of the lensing distribution (actually their intersection with the source plane) appear as high-magnification ridges in such a magnification map; an example is shown in Fig. 1.1.

The prospect of observing microlensing variations in astrophysical lenses is exciting because microlensing probes scales $\sim \theta_E$, which as noted above is $\sim 10^{-6}$ arcsec for stellar deflectors. For a distant quasar this works out to a linear size of $\sim 10^{16}$ cm ~ 0.01 pc, about the same size as that inferred for the continuum emission region in these objects (§1.2). There is thus the possibility of gleaning information on conditions in cosmologically distant objects from regions which cannot be resolved telescopically in even the nearest AGNs! Just such microlensing variations were seen in the quadruple lens system Q2237+0305 in 1988 (Irwin et al. 1989); Chapter 2 examines whether the observations can provide useful constraints on AGN accretion disk models. A more general review of additional microlensing effects (not mentioned here) and their usefulness in determining source structure is given in Rauch (1993).

In astronomical lenses deflection angles are always small ($\sim 1''$) and gravitational fields are always weak. In AGNs, however, the presence of MBHs, if in fact true, will lead to lensing in their vicinity in which deflection angles are not small and gravity is strong. Depending on the detailed optical structure of the spacetime around an MBH, such lensing could have noticeable effects on the emission produced in its vicinity. This kind of lensing can be called 'intrinsic' because the lens in this case is an integral part of the object itself, as opposed to cosmological lenses in which the deflector is unrelated to the source. For Schwarzschild (uncharged, non-rotating) black holes, there is axisymmetry about the line of sight and the optical behavior is similar to that of the (weak field) point mass described above. The major difference in this case is that there can be more than two images of the

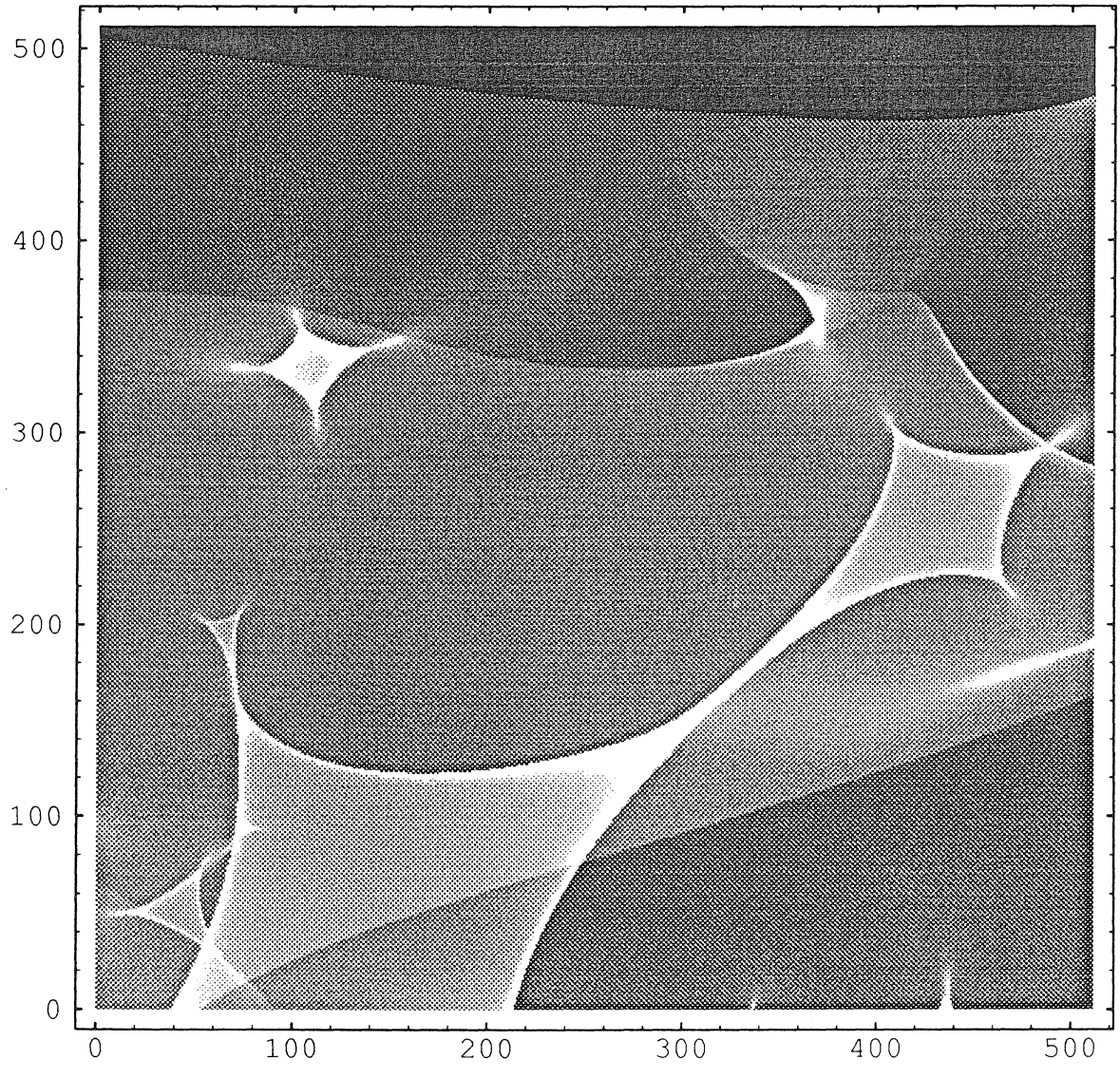

Fig. 1.1: A portion of a microlensing caustic network (lensing magnification pattern projected onto the source plane) produced by an intervening screen of stars (point masses), showing both cusps (tips of diamonds) and interconnecting folds. Axes scales are arbitrary.

source; in fact, there are always an *infinite* number of images, although almost all will be so highly demagnified that they are of no practical importance. To understand how this comes about, we need to step back for a moment and consider more formally the propagation of light in general relativity.

In general relativity (and under conditions analogous to the geometric optics limit of ordinary mechanics), the trajectories of light rays and all other massless particles are described by null geodesics, which satisfy the equations of motion which can be derived for any particular spacetime from the corresponding metric components. The equations of geodesic deviation determine how the cross-section of a small bundle of rays changes due to shearing and focusing effects as the bundle propagates along its central null geodesic; caustics can be located by finding the conjugate points of the ray bundle, which are the points at which the cross-sectional area of the bundle vanishes.

In the spherically symmetric Schwarzschild spacetime, all geodesic motion is confined by symmetry to a plane. The solution for null geodesics (which reach infinity) in this case can be written

$$\left(\frac{du}{d\theta}\right)^2 = 2u^3 - u^2 + b^{-2}, \tag{1.4}$$

where $r = 1/u$ is the coordinate distance from the hole, θ is the azimuthal angle in the plane of the trajectory, b is the impact parameter of the ray, and r and b are assumed to be in units of GM/c^2. Examination of (1.4) reveals that for $b^2 < 27$ the cubic, whose roots determine the radial turning point of the trajectory, has no positive roots and hence the photon is captured by the black hole, for $b^2 = 27$ there is a double root at $u = 1/3$ and the photon asymptotes to a circular orbit at $r = 3$, and for $b^2 > 27$ the photon escapes to infinity. For $|b^2 - 27| \ll 1$, the photon circles the hole many times (near $r = 3$) before being captured or escaping—infinitely many times in the limit $b^2 \to 27$. This means that for any given source position, there are an unlimited number of rays which will circle the hole just the right number of times to eventually intersect the source (almost all having $b^2 \approx 27$), or in other words, the black hole creates an unlimited number of images of any

source. As the number of revolutions made by the photon around the black hole is extremely sensitive to b when that number is large, however, these images will be quite strongly demagnified, since tiny changes in b (apparent image position) lead to large changes in position in the source plane.

For a Kerr black hole, the symmetry about the optic axis is broken, photon trajectories no longer lie in a plane, and the optical structure is more complex. Examination of the optical properties in this case requires that the full machinery of general relativity be used, with photon trajectories being accurately followed using the equations of motion for null geodesics in the Kerr metric and with the equations of geodesic deviation needing to be properly integrated to determine caustic locations and local magnifications. The determination of the optical (caustic) structure of Kerr black holes and its possible consequences for AGNs is the subject of Chapter 3.

1.2. UNCOVERING THE INNER STRUCTURE OF AGNS

Active galactic nuclei have been observed in one or another of their various incarnations for over half a century (e.g., Seyfert 1943, who referred to his objects as "peculiar extragalactic nebulae"). Objects falling under the AGN umbrella include Seyfert galaxies (Types 1, 2, 1.5, 1.8, ... ; e.g., Osterbrock 1993), BL Lac objects and the related blazars and optically violent variables (OVVs), quasars or quasi-stellar objects (QSOs; now essentially synonymous, originally the terms were applied to radio loud and radio quiet objects, respectively), radio galaxies (extended Fanaroff-Riley (FR) type I and II or compact core sources), and low ionization emission line galaxies (LINERs). Starburst galaxies and ultraluminous IR galaxies can also be included. A 'typical' AGN is characterized by strong continuum emission spanning several decades in frequency, from far IR to X-rays, with most of the power residing in the infrared and in the so-called Big Blue Bump centered at UV energies (where the absolute emission is at a maximum); some AGN are also strong radio sources, and a number have also been detected

in gamma rays (Hunter et al. 1993; Bertsch et al. 1993). Very roughly, the continuum flux as a function of frequency can be described by a power law of index $\alpha \approx 0.5 - 1$ ($F_\nu \propto \nu^{-\alpha}$) from IR to X-ray energies, although this is definitely an oversimplification (e.g., Wills et al. 1985). In addition, all classes except BL Lacs show some combination of broad (widths $\sim 3,000 - 30,000$ km s^{-1}) and/or narrow (widths $\sim 200 - 1000$ km s^{-1}) emission lines, which are believed to be produced by the central (unresolved) continuum source energizing surrounding gas clouds; the physical regions within the AGN in which the lines are produced are termed the broad line region (BLR) and narrow line region (NLR). Observationally, the main characteristics delineating the different classes are the strength or absence of both their radio emission (and its morphology) and their broad or narrow emission lines, and their variability and polarization properties (both of which can be rather extreme in some objects). Although initially studied as separate phenomena on the basis of their varied traits, more recently the trend has been to combine this rather daunting list of objects into a unified framework in which differences in viewing angle, opacity, etc., of a prototype object lead to bifurcations in observed properties (e.g., Lawrence 1987). At the present time substantial unification seems viable. A recent, detailed review of this subject can be found in Antonucci (1993).

While progress has been made over the years in elucidating the physical conditions present in AGNs, many of the constraints remain rather generic. This state of knowledge is probably to be anticipated at this stage, for two basic reasons. From the theorist's point of view, real AGNs are expected to be very 'messy' systems, with MBHs, accretion disks, star clusters, magnetic fields, relativistic jets, emission line clouds, star formation regions, and perhaps yet unknown components all being potentially important, making the development of realistic models a very complicated and uncertain proposition. Observationally, the issue is challenging because in almost all cases the active nucleus, and in particular the region within which its power output is being generated, is spatially unresolved, so that observations tend to produce only circumstantial evidence to which indirect (and usually model-dependent) arguments must be applied before physical conclusions

can be made. Thus while a rough estimate of the mass contained in the inner core of an AGN (or normal galactic nucleus) can be convincingly made using simple dynamical reasoning, the widely debated question (Tonry 1987; Dressler & Richstone 1988; Kormendy 1994) of whether most of the mass is in the form of an MBH has yet to be resolved with confidence, and new lines of investigation have an opportunity to produce useful insights. On the theoretical side, this motivates explorations to determine what effects are produced by specific AGN model components, which of them are likely to be of importance in real systems, and which are not; the present work constitutes one such exploration of a small region of 'paradigm space.' Discussion of some currently known constraints, and how they have been arrived at, will help to put things into context.

The discovery of quasar redshifts (Schmidt 1963) was a major breakthrough in constraining the nature of these most energetic members of the AGN family. Spirited debate in the years after this discovery was given to how the large observed redshifts were being produced, the main contenders being Doppler effects (of either the source as a whole or of material within it) and gravitational redshift due to emission inside a deep gravitational potential. The cosmological interpretation of AGN redshifts, in which the redshifts are Doppler in nature and due to the large recession velocity of the source as a whole, is now both well accepted and well supported, the COBE results on the nature of cosmic background radiation (Smoot et al. 1992) being one recent example in favor of the standard cosmological interpretation of redshifts. This interpretation is important because it allows distances to these objects to be estimated, which in turn determines the scale of one of the most important physical parameters of an AGN: its power output. The luminosities so obtained, assuming isotropic emission in the AGN rest frame, range from $L_{\rm bol} \sim 10^{41}$ erg s^{-1} for a modest Seyfert galaxy to $L_{\rm bol} \sim 10^{47}$ erg s^{-1} for a high power quasar.

A significant constraint on what could be generating such prodigious amounts of power comes from the variability of the observed continuum emission, which at X-ray frequencies can vary by factors of order unity on timescales of minutes

to hours in some objects (Kunieda et al. 1990; Edelson 1994). By making the simple (and perhaps naive) argument that significant variations cannot occur more quickly than the light travel time across the source, emission regions $\lesssim 0.01$ pc in size—and in some cases of order the size of the solar system—are obtained. If the equally simple and naive argument is made that the emission, to be quasi-steady, should not be greater than the Eddington limit, then a lower limit on the mass that must be contained within this volume can be found, specifically, $M \gtrsim 10^8 L_{46} M_\odot$, where $L_{46} = L_{\rm bol}/(10^{46}$ erg s$^{-1})$—a substantial amount of matter to pack into a region so small! Squeezing such large amounts of power out of such small volumes on a sustained basis requires a process with high efficiency in converting rest mass to radiant energy, apparently higher than nuclear or atomic processes can reasonably provide, leading to the hypothesis that the ultimate energy source is gravitational in nature (e.g., Robinson et al. 1964). Nearly all AGN models still in serious contention work under this assumption.

Estimates of physical conditions in the BLR and NLR can be gotten in several ways. Observations of ionization states of different lines and of line strength ratios, for instance, can be used together with photoionization models to compute the temperature and density of the emitting gas (Ferland & Netzer 1983; Aldrovandi & Contini 1985), which for the NLR leads to temperatures of $\sim 10^4$ K and densities (of the electrons) of $\sim 10^3 - 10^4$ cm^{-3}, and for the BLR $\sim 10^4$ K and $\sim 10^9 - 10^{10}$ cm^{-3}, respectively. The sizes of these regions—which in reality are not really distinct and separate but represent average conditions at two characteristic distances from the central ionizing source—can also be estimated from photoionization models, giving sizes of $\sim 100 - 1000$ pc for the NLR and $\sim 0.1 - 1$ pc for the BLR. For a few nearby Seyfert galaxies, the NLR can be resolved observationally (Osterbrock 1991), and their sizes so obtained are also ~ 100 pc. Although the BLR is too small to be resolved telescopically, reverberation mapping, in which the observed lag time between variations in the continuum source and the changes in line intensities in response to those variations is used with BLR models to determine the separation between the two, has been used to

derive BLR sizes, generally giving results of $\sim 0.01 - 0.1$ pc, rather smaller than the above order of magnitude estimate (Peterson 1993 and references therein). From the sizes and the absolute line luminosities, filling factors for the emission line clouds of ~ 0.01 are obtained.

In the 'standard' model for AGNs, which can satisfy all of the preceding conditions, the central engine consists of a massive black hole in which fueling occurs via an accretion disk, which is also responsible for generating the ionizing continuum source upon which the BLR and NLR depend. This is the basic framework that will be assumed in subsequent chapters. Theoretical arguments in favor of accretion disks include their ability to solve the angular momentum problem—that is, how to remove enough angular momentum from ambient material on large scales to allow it to be accreted by the black hole—and perhaps also their creation of an axisymmetric geometry, which might aid the formation of bipolar jets or other outflows (Begelman, Blandford, & Rees 1984). Observationally, disks can be argued for as the source of the Big Blue Bump, and they also appear able to explain the properties of the low ionization emission lines (Ulrich 1989). The black hole model is attractive because of its ability to generate high conversion efficiency of rest mass to energy (up to $\sim$ 40% for a maximally rotating black hole) and because a black hole is a stable state gravitationally. Other models of AGNs involving supermassive stars (Hoyle & Fowler 1963), star clusters (Spitzer & Saslaw 1966; Colgate 1967), or dense starbursts (e.g., Terlevich 1992) do not share the latter property, and are almost certain to collapse and form a black hole during subsequent evolution; as indicated schematically in an oft-reproduced diagram (Fig. 1 of Rees 1984), all roads lead to massive black holes.

Additional support for the black hole model exists in observations of normal galactic nuclei. If the black hole model is correct, then many (and perhaps almost all) normal galactic nuclei will possess MBHs in their centers, remnants of once active nuclei that have exhausted their gas supply. The presence of these black holes can be inferred by observing the light and velocity dispersion profiles of the nuclear stars and the rotation curves of nuclear gas clouds; if an MBH resides in

the center, cusps will develop in these curves as the center is approached (see §1.3 for additional information). Observations of several nearby galaxies, such as M31, M32, and our own galaxy (Lugger et al. 1992; Genzel, Hollenbach, & Townes 1994) are consistent with their centers harboring black holes with masses ranging from $\sim 10^6 M_\odot$ for the Milky Way to $\sim 3 \times 10^7 M_\odot$ for Andromeda. A recent HST observation (Harms et al. 1994) of M87 gives strong evidence for a central mass of $\sim 3 \times 10^9 M_\odot$; interestingly, this is essentially the same value proposed by Sargent et al. (1978) in one of the earliest observational searches of this kind. Overall, while it has not been definitively proven that MBHs inhabit either normal or active galactic nuclei, it is fair to say that the case for them is already strong on both observational and theoretical grounds, and appears to be getting continually stronger.

Of the alternative models for AGNs mentioned above, the starburst model is perhaps the most plausible. In this model, commencement of AGN activity is envisioned as a step in the final formation of the young galaxy. The photoionizing continuum is then supposed to be due to early-type main sequence stars, with "warmers" (rapidly evolving, very massive ($\sim 100 M_\odot$) stars), supernovae, and supernova remnants leading to the production of the narrow emission lines, non-thermal emission and the BLR, and additional broad permitted lines, respectively (Terlevich & Melnick 1985; Terlevich et al. 1987; Filippenko 1989). Low luminosity Seyferts might contain only one active supernova remnant, allowing for rapid variability, while in a luminous quasar the number could be of order 100 (Terlevich 1989). Although not a model of AGN activity itself, it has also been suggested that the ultraluminous IRAS galaxies may be an early stage in the development of quasars (Sanders et al. 1988).

The starburst and black hole models for AGNs lead to very different interpretations of how such objects evolve. In the black hole model the evolution is determined mainly by the continuity of the supply of accreted gas, where dormant phases during times of depleted gas supply are punctuated by episodes of renewed activity when the hole is refueled, as by close interactions with a neighboring galaxy

(Byrd et al. 1987). The black hole continues to grow during these outbursts, which implies that its mass increases exponentially with an e-folding time of $\sim 10^8$ yr, if the growth is Eddington limited; note that this happens to be similar to galactic dynamical timescales, which set the rate at which large scale accretion flows can be funneled into the nucleus to support such growth (e.g., Hernquist 1989). In the starburst model evolution is dictated by stellar evolution and the interactions of supernovae and their remnants, and activity declines soon after star formation rates have dropped to where the central cluster can no longer produce these objects at the required rate. As in reality stars and black holes are likely to coexist in AGNs, these models should perhaps be considered more complementary than antagonistic, with starburst nuclei such as M82 being predominantly stellar-driven and high power quasars corresponding to accretion-dominated objects.

Even within the standard black hole model, however, there are numerous uncertainties. Is the accretion disk geometrically thick or thin? Is the local emission blackbody? How is angular momentum transported within the disk? Is there a Comptonizing corona? What is the nature of the intercloud region in the BLR? Are magnetic fields important? As a general relativistic object, what special influences does the black hole impart on its surroundings? The list continues. To gain insight into questions such as these requires careful, involved analysis of the issue under investigation. This thesis explores a few specific ones.

1.3. DYNAMICAL PROCESSES IN GALACTIC NUCLEI

The final two chapters of this work are concerned with the dynamical evolution of dense, relativistic star clusters around an MBH residing in a (possibly active or inactive) galactic nucleus. This section introduces the physical processes which can affect such evolution. The history of this subject dates back nearly 30 years to the work of Spitzer and Saslaw (1966) and Spitzer and Stone (1967), who simulated the evolution of a compact cluster of stars undergoing collapse due to gravitational relaxation and subsequent mass liberation from high velocity (disruptive) stellar

collisions. These early models did not include black holes; the motivation at this time was the possibility that energy release from the collisions themselves could power QSOs (Woltjer 1964). Soon after (Colgate 1967) it was realized that the cluster would undergo a merger period in which collision velocities were still small enough that the stars would not disrupt, but coalesce instead. In this scenario, the growth of massive stars would produce, on stellar evolutionary timescales, high supernova rates that were hoped could account for observed quasar luminosities. The work of Colgate (1967) suggested they could; that of Sanders (1970) argued against it in all but low luminosity objects. The discrepancy appears to arise from the somewhat different assumptions made regarding collision outcomes, which both authors estimated using simple analytical rules. Soon after this time the black hole hypothesis described previously began its rise to prominence. Hills (1975) suggested that tidal disruption of stars in the vicinity of an MBH and subsequent accretion of gas might be the energy production mechanism in AGNs. A flurry of activity by several groups (Bahcall & Wolf 1976; Frank & Rees 1976; Lightman & Shapiro 1977; Young et al. 1977; Bahcall & Wolf 1977; Frank 1978; Young 1980) began to examine in greater detail the effects of an MBH on the distribution function of a surrounding star cluster and the importance of tidal mass loss induced by relaxation, which occurs when orbits diffuse into ones passing so close to the MBH that the stars are tidally disrupted near periapse. The general conclusion of these studies was that tidal mass loss alone was insufficient to fuel higher power AGNs unless stellar densities were so high that other processes, in particular stellar collisions, became even more important. Observational searches for MBHs in galactic nuclei were also begun around this time (e.g., Young 1979). Begelman and Rees (1978) reconsidered the work of Spitzer and Saslaw (1966) and related studies and concluded that the evolution of such systems would result in the formation of an MBH. Begelman, Blandford, and Rees (1980) explored AGN models involving binary MBHs. The formation of initial 'seed' black holes by collapse of a relativistic cluster of (collisionless) compact objects such as neutron stars has also been investigated (Shapiro & Teukolsky 1986; Quinlan & Shapiro 1989;

Lee 1993). More recent calculations of the dynamical evolution of galactic nuclei (McMillan, Lightman, & Cohn 1981; Duncan & Shapiro 1983; David, Durisen, & Cohn 1987a,b; David & Durisen 1989; Murphy, Cohn, & Durisen 1991; Quinlan & Shapiro 1990) have tended to focus on performing Fokker-Planck simulations (in which the star cluster is described by a phase-space distribution function whose time evolution is governed by Boltzman's Equation with various source and sink terms) incorporating increasing numbers of active processes, including the effects of growing black holes (modeled as Newtonian point potentials), stellar collisions, tidal mass loss, stellar evolution, mergers, binary formation and heating, star formation, and two-body relaxation. In the simulations performed as part of this thesis, the dominant dynamical processes are stellar collisions and interactions with an accretion disk (not mentioned above) and the resulting evolution of the cluster distribution function (e.g., changes in the density profile), with relaxation and tidal effects also making appearances; the remainder of this section discusses these processes in greater detail.

Perhaps the most basic dynamical process present in a star cluster (with or without a central black hole) is that of two-body gravitational relaxation, in which gravitational encounters between stars lead to slow modifications in individual orbital parameters. Relaxation leads to the development of a core-halo structure in which energy is continually transferred from core stars with high binding energy to more loosely bound halo stars, causing the core to evolve towards even higher densities and binding energies. The timescale for relaxation to occur in a cluster of equal mass stars is given by (Spitzer & Hart 1971)

$$\begin{aligned} t_{\rm r} &= \frac{v_{\rm rms}^3}{4\pi(3/2)^{1/2}G^2 m_\star \rho(r)\ln(0.4N_\star)} \\ &\sim 2\times 10^{11}\left(\frac{v_{\rm rms}}{10^3\ {\rm km\ s^{-1}}}\right)^3\left(\frac{m_\star}{M_\odot}\right)^{-1}\left(\frac{\rho(r)}{10^6\ M_\odot\ {\rm pc^{-3}}}\right)^{-1}\ {\rm yr} \end{aligned} \tag{1.5}$$

for $\ln(0.4N_\star) \approx 15$, where $N_\star$ is the number of stars in the cluster and $v_{\rm rms}$ is the velocity dispersion; the logarithm term is essentially a Coulomb-type logarithm expressing the fact that the many distant encounters contribute as much to the

relaxation as the occasional close encounters (at least when $N_\star \gg 1$). In real galactic nuclei $t_{\rm r}$ can be either shorter or longer than a Hubble time.

The essential effect of an MBH on a cluster of stars in dynamical equilibrium around it is to create a power law cusp in the density profile and velocity dispersion in the vicinity of the hole (i.e., where the gravitational potential of the MBH dominates). The precise value of the equilibrium power law depends somewhat on the specific assumptions made, such as how the black hole came into being. An early scaling argument due to Peebles (1972), in which dynamical equilibrium was taken to imply a radius-independent flow rate of stars towards smaller radii (which occurs by gravitational relaxation), leads to a cusp of the form $\rho(r) \propto r^{-9/4}$. As shown by Bahcall and Wolf (1976), the Peebles solution does not satisfy the proper boundary conditions at large and small radii and actually implies diffusion of stars *away* from the MBH; the self-consistent solution they proceed to derive has the form $\rho(r) \propto r^{-7/4}$, which they show holds for both isotropic and anisotropic (in velocity space) distribution functions, as long as the form of the anisotropy is independent of radius. In these cases equilibrium is reached on a relaxation timescale. If one assumes that an initially small black hole grows adiabatically in the core of a larger cluster (due to gradual consumption of cluster stars), meaning that the mass doubling timescale is much longer than the dynamical timescale but much shorter than the relaxation timescale for the cluster, then for isotropic velocities one finds $\rho(r) \propto r^{-3/2}$ (Peebles 1972; Young 1980) and for purely circular initial orbits $\rho(r) \propto r^{-9/4}$ (Young 1980); velocity dispersions in the isotropic case rise as $r^{-1/2}$. In this case the cusp develops on the mass doubling timescale.

A modification of the $r^{-7/4}$ law results when tidal effects are taken into account (Frank & Rees 1976; Lightman & Shapiro 1977). Recall first that a star will be disrupted by tidal forces when it passes within a distance $r_{\rm t} \approx 2R_\star(M/m_\star)^{1/3}$ of the black hole (of mass M), at which point the star begins to overflow its Roche lobe. As noted in Lightman & Shapiro (1977), the relaxation time can be thought of as the timescale for the rms angular momentum of an orbit of binding energy $\mathcal{E}$ to change by $L_{\rm max}(\mathcal{E})$, the angular momentum of a circular orbit of energy $\mathcal{E}$.

Thus within a relaxation time orbits will eventually diffuse into the 'loss cone,' which consists of those low angular momentum orbits, $L < L_{\rm min}$, which pass inside the star's tidal radius; a star residing in the loss cone will be disrupted in a dynamical time, unless the rms angular momentum transfer per orbital time ΔL exceeds $L_{\rm min}$, in which case only a fraction $L_{\rm min}^2/L^2$ remain in the loss cone long enough to suffer disruption. As shown in the referenced papers, inclusion of loss cone effects leads to a flattening of the $r^{-7/4}$ cusp inside a critical radius (the radius where $\Delta L \sim L_{\rm min}$), although the change in slope remains small until one approaches relatively close to the tidal radius.

Physical collisions between the stars in a cluster occur on a timescale $t_{\rm coll} \sim (n_\star \sigma_{\rm coll}(v_{\rm rel}) v_{\rm rel})^{-1}$, where $n_\star$ is the stellar number density, $v_{\rm rel}$ is the typical relative velocity, and $\sigma_{\rm coll}(v_{\rm rel}) \approx 4\pi R_\star^2 (1 + v_{\rm esc}^2/v_{\rm rel}^2)$ is the collision cross section between identical stars of radius $R_\star$ and surface escape velocity $v_{\rm esc}$ (the first term represents the geometric cross section and the latter is the enhancement due to gravitational focusing). The amount of mass released in a collision depends, naturally, on both the impact parameter and on $v_{\rm rel}$, as well as on the density profiles of the stars themselves. Qualitative differences in the outcomes of collisions occur when $v_{\rm esc} < v_{\rm rel}$ and $v_{\rm esc} > v_{\rm rel}$, as in the former situation collisions are highly disruptive (unless the collision is at grazing incidence) while in the latter stars tend to coalesce with relatively little mass loss. Early simulations such as Spitzer and Saslaw (1966) computed the mass loss from collisions using simple analytic prescriptions. In recent years the development of numerical hydrodynamics codes, and in particular application of the method of smooth particle hydrodynamics (Benz & Hills 1987; Davies, Benz, & Hills 1991, 1992; Lai, Rasio, & Shapiro 1993) has allowed much more precise estimates of mass loss from stellar collisions to be made, which the simulations described in Chapter 5 will make use of.

Finally, there is the effect of star-disk interactions (Syer, Clarke, & Rees 1991). If an accretion disk is present around the MBH in an AGN that also contains a dense star cluster, its dynamical effects on the cluster can be substantial. It is obvious that each star will hit the disk twice per orbit. Assuming that the disk is

relatively thin and its orbits nearly circular, the drag force experienced by the stars during each plane crossing will preferentially remove radial and perpendicular (to the disk plane) kinetic energy from the stars' orbits, causing them to be slowly ground down into increasingly circular orbits with smaller and smaller inclinations to the disk plane. This process occurs on a 'grinding' timescale given by $t_{\rm grind} \sim t_{\rm dyn}\Sigma_\star/(2\Sigma_{\rm d})$, where $\Sigma_{\rm d}$ is the surface density of the disk and $\Sigma_\star = m_\star/(\pi R_\star^2)$ is the effective surface density of the star. The grinding timescale is essentially the time it takes a star to cumulatively interact with an amount of disk material equal to its own mass. There are astrophysically plausible conditions under which either $t_{\rm grind} < t_{\rm coll}$ or $t_{\rm grind} > t_{\rm coll}$ could prevail, making this effect of potential importance in the dynamical evolution of an AGN; the case $t_{\rm grind} \ll t_{\rm coll}$, taking account of relativistic effects, is explored in Chapter 4.

REFERENCES

Aldrovandi, S. M. V., & Contini, M. 1985, A&A, 149, 109

Antonucci, R. 1993, ARA&A, 31, 473

Bahcall, J. N., & Wolf, R. A. 1976, ApJ, 209, 214

———. 1977, ApJ, 216, 883

Begelman, M. C., Blandford, R. D., & Rees, M. J. 1980, Nature, 287, 307

———. 1984, Rev Mod Phys, 56, 255

Begelman, M. C., & Rees, M. J. 1978, MNRAS, 185, 847

Benz, W., & Hills, J. G. 1987, ApJ, 323, 614

Berry, M. V., & Upstill, C. 1980, Prog. Optic, 18, 258

Bertsch, D. L. et al. 1993, ApJ, 405, L21

Blandford, R. D., & Narayan, R. 1992, ARA&A, 30, 311

Byrd, G. G., Sundelius, B., & Valtonen, M. 1987, A&A, 171, 16

Chang, K., & Refsdal, S. 1979, Nature, 282, 561

Colgate, S. A. 1967, ApJ, 150, 163

David, L. P., & Durisen, R. H. 1989, ApJ, 346, 618

David, L. P., Durisen, R. H., & Cohn, H. N. 1987a, ApJ, 313, 556

———. 1987b, ApJ, 316, 505

Davies, M. B., Benz, W., & Hills, J. G. 1991, ApJ, 381, 449

———. 1992, ApJ, 401, 246

Dressler, A., & Richstone, D. O. 1988, ApJ, 324, 701

Duncan, M. J., & Shapiro, S. L 1983, ApJ, 268, 565

Eddington, A. S. 1919, Observatory, 42, 119

Eddington, A. S. 1920, Space, Time, and Gravitation (Cambridge: Cambridge University Press)

Edelson, R. 1994, in Courvoisier, T. J.-L., Blecha, A., eds, Multi-wavelength Continuum Emission of AGN (Dordrecht: Kluwer), p. 113

Einstein, A. 1915, Sitzunber. Preuß Akad. Wissensch. ersten Halbband, p. 831

Einstein, A. 1936, Science, 84, 506

Ferland, G. J., & Netzer, H. 1983, ApJ, 264, 105

Filippenko, A. V. 1989, AJ, 97, 726

Frank, J. 1978, MNRAS, 184, 87

Frank, J., & Rees, M. J. 1976, MNRAS, 176, 633

Genzel, R., Hollenbach, D., & Townes, C. H. 1994, Rep. Prog. Phys, 57, 417

Harms, R. J. 1994, ApJ, in press

Hernquist, L. 1989, Nature, 340, 687

Hewitt, J. N., Turner, E. L., Schneider, D. P., Burke, B. P., Langston, G. I., & Lawrence, C. R. 1988, Nature, 333, 537

Hills, J. G. 1975, Nature, 254, 295

Hoyle, F., & Fowler, W. A. 1963, MNRAS, 165, 129

Hunter, S. D. et al. 1993, ApJ, 409, 134

Irwin, M. J., Webster, R. L., Hewett, P. C., Corrigan, R. T., & Jedrzejewski, R. I. 1989, AJ, 98, 1989

Kormendy, J. 1994, in Harris A., Genzel, R., eds, The Nuclei of Normal Galaxies: Lessons from the Galactic Center

Kunieda, H., Turner, T. J., Awaki, H., Koyama, K., & Mushotzky, R. 1990, Nature, 345, 786

Lai, D., Rasio, F. A., & Shaprio, S. L. 1993, ApJ, 412, 593

Langston, G. I., Schneider, D. P., Conner, S., Carilli, C. L., Lehár, J., Burke, B. P., Turner, E. L., Gunn, J. E., Hewitt, J. N., & Schmidt, M. 1989, AJ, 97, 1283

Lawrence, A. 1987, PASP, 99, 309

Lee, M. H. 1993, ApJ, 418, 147

Lightman, A. P., & Shapiro, S. L. 1977, ApJ, 211, 244

Lugger, P. M., Cohn, H. N., Cederblom, S. E., & Lauer, T. R. 1992, AJ, 104, 83

Lynds, R., & Petrosian, V. 1989, ApJ, 336, 1

McMillan, S. L. W., Lightman, A. P., & Cohn, H. 1981, ApJ, 251, 436

Murphy, B. W., Cohn, H. N., & Durisen, R.H. 1991, 370, 60

Newton, I. 1704, Optics, 2nd edition, query 1

Osterbrock, D. E. 1991, Rep Prog Phys, 54, 579

———. 1993, ApJ, 404, 551

Paczyński, B. 1986, ApJ, 301, 503

Peebles, P. J. E. 1972, Gen. Rel. and Grav., 3, 63

Peterson, B. M. 1993, PASP, 105, 247

Quinlan, G. D., & Shapiro, S. L. 1989, ApJ, 343, 725

———. 1990, ApJ, 356, 483

Rauch, K. P. 1993, in Surdej, J., Fraipont-Caro, D., Gosset, E., Refsdal, S., Remy, M., eds, Gravitational Lenses in the Universe, proc. of the 31st Liège Int. Astroph. Coll., p. 385

Rees, M. J. 1984, ARA&A, 22, 471

Refsdal, S., & Surdej, J. 1994, Rep Prog Phys, 56, 117

Robinson, I., Schild, A., & Schucking, E. 1964, eds, Quasi-stellar Sources and Gravitational Collapse (Chicago: Chicago Univ. Press)

Sanders, R. H. 1970, ApJ, 162, 791

Sanders, D. B., Soifer, B. T., Elias, J. H., Madore, B. F., Matthews, K., Neugebauer, G., & Scoville, N. C. 1988, ApJ, 325, 74

Sargent, W. L. W., Young, P. J., Boksenberg, A., Shortridge, K., Lynds, C. R., & Hartwick, F. D. A., ApJ, 221, 721

Schmidt, M. 1963, Nature, 197, 1040

Schneider, P., Ehlers, J., & Falco, E. E. 1992, Gravitational Lensing (Berlin: Springer)

Shapiro, S. L., & Teukolsky, S. A. 1986, ApJ, 307, 575

Smoot, G. et al. 1992, ApJ, 396, L1

Soucail, G., Fort, B., Mellier, Y., & Picat, J. P. 1987, A&A, 172, L14

Spitzer, L., & Hart, M. H. 1971, ApJ, 164, 399

Spitzer, L., & Saslaw, W. C. 1966, ApJ, 143, 400

Spitzer, L., & Stone, M. E. 1967, ApJ, 147, 519

Syer, D., Clarke, C. J., & Rees, M. J. 1991, MNRAS, 250, 505

Terlevich, R., & Melnick, J. 1985, MNRAS, 213, 841

Terlevich, R., Melnick, J., & Moles, M. 1987, in Khachikian, E. Y., Fricke, K. J., & Melnick, J., eds, Observational Evidence for Activity in Galaxies (Dordrecht: Reidel), p. 499

Terlevich, R. 1989, in Beckman, J. E., Pagel, B. E., eds, Evolutionary Phenomena in Galaxies (Cambridge: Cambridge Univ. Press), p. 343

Terlevich, R. 1992, in Filippenko, A., ed, Relationships Between Active Galactic Nuclei and Starburst Galaxies (San Francisco: Astronomical Society of the Pacific), p. 133

Tonry, J. L. 1987, ApJ, 322, 632

Ulrich, M.-H. 1989, in Meyer, F. et al., eds, Theory of Accretion Disks (Dordrecht: Kluwer), p. 3

Walsh, D., Carswell, R. F., & Weymann, R. J. 1979, Nature, 279, 381

Wills, B. J., Netzer, H., & Wills, D. 1985, ApJ, 288, 94

Woltjer, L. 1964, Nature, 201, 803

Young, P. J. 1979, Ph.D. thesis, Cal. Inst. Tech., Pasadena

———. 1980, ApJ, 242, 1232

———. 1981, ApJ, 244, 756

Young, P. J., Shields, J. A., & Wheeler, J. C. 1977, ApJ, 212, 367

Zwicky, F. 1937, Phys Rev Lett, 51, 679

2

Microlensing and the Structure of Active Galactic Nucleus Accretion Disks†

Kevin P. Rauch and Roger D. Blandford

California Institute of Technology

ABSTRACT

Rapid variability has been reported in two of the four, gravitationally-lensed images of Q2237+0305 and this is attributed to microlensing caused by the intervening stars. The associated constraints on the source size and properties are studied and compared with a variety of stationary accretion disk models. It is found that graybody disks have a half-power radius $r \sim 2\times10^{16}(\nu/\nu_L)^{-1.7}\mathcal{L}_{L,46}^{0.5}(\epsilon_{-1})^{-0.5}$ cm where ν is the frequency, $\mathcal{L}_{L,46}$ is the luminosity per log frequency in units of 10^{46} erg s^{-1}, at the Lyman limit ν_L, and $\epsilon = 0.1\epsilon_{-1}$ is the emissivity relative to a blackbody. The reported microlensing variation in Q2237+0305 requires the disk size to be over three times smaller than a blackbody disk of similar luminosity implying that the optical emission is either non-thermal or optically thin. An exploration of non-stationary disk models including orbiting, transient, hot spots leads to a similar conclusion. Implications for models of AGN optical continua are briefly discussed.

Subject headings: gravitational lenses - quasars

† First published in ApJ, 381, L39 (1991).

2.1. INTRODUCTION

The lens system Q2237+0305 is at present the only one in which microlensing effects have been convincingly detected; in 1988 the A image was observed to brighten by $\approx 0.2^m$ within 26d (Irwin et al. 1989). More recent monitoring has produced evidence of a 0.3^m decline in image A over 3 months, a 0.2^m decline in image B over 3 years, and a 0.4^m decline in image C over a similar timescale (Corrigan et al. 1991). Several subsequent studies (e.g., Wambsganss, Paczyński, & Schneider 1990; Witt, Kayser, & Refsdal 1993; Wambsganss & Paczyński 1991) have shown that the frequency and amplitude of the variations are fairly insensitive to the masses of the stars producing the microlensing, due to the rather large optical depths involved. This work examines the constraints one can hope to place on the source itself. The next section discusses simple AGN accretion disk models. Section 2.3 discusses the Q2237+0305 microlensing simulations. Conclusions are presented in Section 2.4.

2.2. ACCRETION DISK MODEL

It is commonly supposed that the continuum emission from an AGN originates from the surface of an accretion disk (e.g., Shields 1978, Czerny & Elvis 1987, Laor & Netzer 1989). However, there is still no satisfactory model that can account for the observed spectrum, polarization, and variability in detail. This has led some researchers to doubt that a disk is even a good first approximation to the flow (e.g., Rees 1984 and references therein). In the simplest type of accretion disk model, the binding energy liberated by the infalling gas is emitted locally as black body radiation. If this happened over several octaves of radius, then a flux $F_\nu \propto \nu^{1/3}$ rising with frequency would be emitted, contrary to the steeper, falling spectra generally observed.

In some disk models this contradiction has been resolved by postulating a non-thermal power-law continuum extending from the far infrared to X-ray energies. The observed optical spectrum would be the sum of the disk and the non-thermal

radiation, (augmented with hydrogen continuum radiation and blended iron line radiation). However, there is now good evidence that the infrared emission longward of 1μ in normal quasars and Seyfert galaxies is re-radiation by hot dust (Sanders et al. 1989).

Despite this, the observed continuum may still be quasi-thermal emission from an accretion disk if the binding energy is radiated non-locally. There are several mechanisms for transporting energy from small disk radius, where most of the binding energy is liberated, to large radius. These include viscous transport (e.g., Shakura & Sunyaev 1973), direct irradiation of a thick (e.g., Pringle 1981) or warped (e.g., Bardeen & Petterson 1975) disk, and backscattering from high altitude (e.g., Shields 1989). The net effect is to increase (steepen) the spectral index. A second generalization of the original thermal disk models acknowledges that a real disk is likely to possess a scattering atmosphere so that it radiates less efficiently than a black body, increasing its local color temperature.

We first make a simple, empirical model of continuum emission from an accretion disk by supposing that the power radiated per unit radius by the disk is arbitrary, but that there is an effective radius-to-frequency mapping $\nu(r) = 10^{15}\nu_{15}\mathrm{Hz}$. If the local disk flux is $f_\nu = f(r)\delta(\nu - \nu(r))$, then the observed angle-averaged flux F_{ν_0} is given by

$$4\pi d_L^2 \nu_0 F_{\nu_0} = \nu L_\nu = 4\pi\nu r(\nu) f(r(\nu)) \left|\frac{dr}{d\nu}\right|, \tag{2.1}$$

where $\nu_0 = (1+z)^{-1}\nu$ is the observed frequency, L_ν is the spectral luminosity, and d_L is the luminosity distance to the AGN. More generally, we can use the observed spectrum in an individual AGN to infer both f(r) and $r(\nu)$, given a suitable assumption about the emissivity. We also suppose that the emission is gray so that the effective frequency varies according to $\nu(r) \simeq 3k[f(r)/\epsilon\sigma]^{1/4}/h$, where σ is the Stefan-Boltzmann constant, and $\epsilon < 1$ (taken for simplicity to be constant over radius and frequency) is the emissivity relative to a black body. If we model the spectrum shortward of 1μ as a power-law, $F_{\nu_0} \propto \nu_0^{-\alpha}$, and normalize the spectral luminosity to its value at the Lyman continuum, $\mathcal{L} = 10^{46}\mathcal{L}_{46}$ erg

$s^{-1} = \nu_L L_{\nu_L}$, where $\nu_L = 3.3 \times 10^{15}$ Hz is the Lyman continuum frequency, then we can relate the radius of the source at a given frequency to $\mathcal{L}$ using

$$r(\nu) \propto \epsilon^{-1/2} \mathcal{L}^{1/2} \nu^{-(3+\alpha)/2}. \tag{2.2}$$

The associated disk flux variation is

$$f(r) \propto \epsilon^{-\left(\frac{1-\alpha}{3+\alpha}\right)} \mathcal{L}^{\frac{4}{3+\alpha}} r^{\frac{-8}{3+\alpha}}, \tag{2.3}$$

ignoring edge effects. This spectrum will only extend up to the frequencies that are radiated close to the inner edge of the accretion disk from an area proportional to the square of the hole mass. This depends upon the Eddington ratio $\ell = L/L_{Edd}$. The frequency, $\nu_{\max}$, where $\nu L_\nu$is maximized satisfies the scaling law

$$\nu_{\max} \propto \ell^{\frac{2}{5-\alpha}} \mathcal{L}^{\frac{-1}{5-\alpha}} \epsilon^{\frac{-1}{5-\alpha}}. \tag{2.4}$$

For a thin disk, $\ell \lesssim 0.3$.

In order to estimate constants of proportionality in these scaling relations and to test their robustness to detailed assumptions about the disk emissivity, we have computed more sophisticated disk models which incorporate relativistic corrections (Doppler and gravitational redshifts and light deflection near the black hole) in a Schwarzschild metric, with a minimum disk radius equal to that of the least stable circular orbit, and an emissivity $f_\nu(\nu(r)) = \pi \epsilon B_\nu(T(r))$. For a nominal spectral index $\alpha \sim 0.5$, $\ell \sim 0.3$, and $\epsilon = 0.1\epsilon_{-1}$, we find that half the luminosity is emitted within a radius

$$\begin{aligned} r(\nu) &\sim 2 \times 10^{16} (\epsilon_{-1})^{-0.5} \mathcal{L}_{46}^{0.5} \nu_L^{-1.7} \text{ cm}, \\ \nu_{15} &< \nu_{\max,15} = 7 \left(\frac{\ell}{0.3}\right)^{0.4} \mathcal{L}_{46}^{-0.2} (\epsilon_{-1})^{-0.2}. \end{aligned} \tag{2.5}$$

We find that 90 percent of the luminosity is radiated within $\sim 3r(\nu)$. (See Fig. 2.1.)

Several comments should be made about equations (2.2)-(2.5). Firstly, in order to reproduce a spectral index $\alpha = 0.5$, the power emitted per octave of disk radius $\propto r^2 f$ must decay slowly with radius, typically $\propto r^{-0.3}$. Secondly,

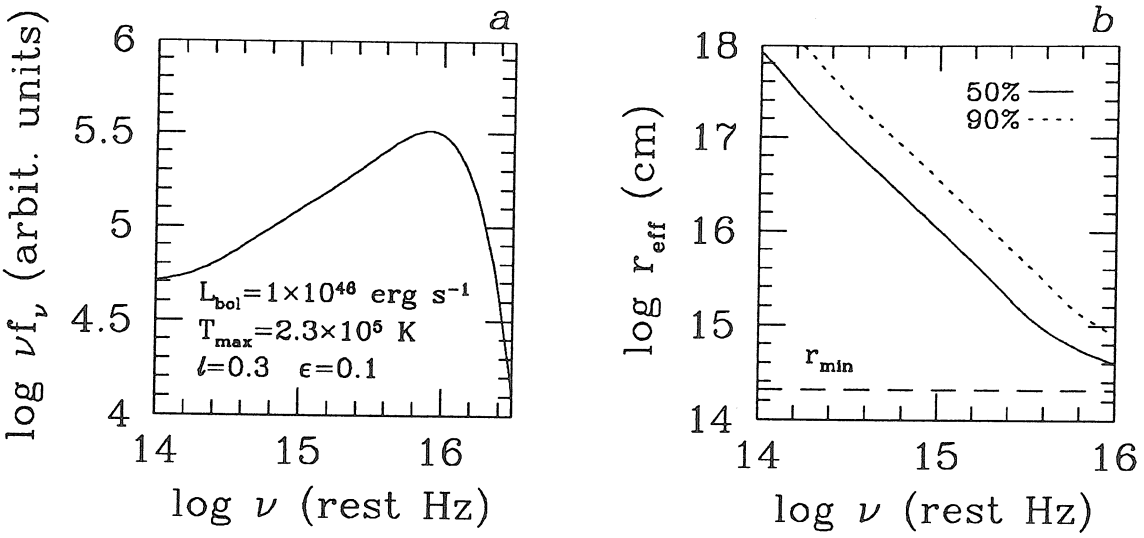

Fig. 2.1: (*a*) Integrated spectrum νF_ν for an empirical thin accretion disk model constructed to produce a spectral index $\alpha = 0.5$ in the optical-near UV with luminosity appropriate to a moderately powerful quasar. The spectrum at frequencies below $\sim 3 \times 10^{14}$ Hz is presumed to be re-emission by thermal dust; that above $\nu_{\rm max} \sim 10^{16}$ Hz must be produced non-thermally. (*b*) Associated half-power and 90 percent power radii for the same disk model. The size at high frequency is sensitive to the assumptions made about the inner edge of the disk at $r_{\rm min}$. However, the optical spectrum is quite insensitive to these details.

the disk effective radius at a given frequency increases with decreasing emissivity. Black body emission produces the most compact sources. Thirdly, although the spectrum near $\nu_{\rm max}$ is not observed directly, there are indications that the continuum extends, with approximately constant slope beyond ν_L (Steidel & Sargent 1987). In the highest redshift and luminosity quasars (e.g., HS 1700+6416), the spectrum may extend to too high frequency to be compatible with a simple disk spectrum (Reimers et al. 1989). The frequency $\nu_{\rm max}$ is not sensitive to the Eddington ratio and the emissivity as the scaling laws show. However, it is sensitive to the spin of the hole and the inclination. Nevertheless, these sensitivities are not apparent at lower frequencies where existing and anticipated microlensing ob-

servations have been made. Fourthly, this simple gray atmosphere model cannot describe the complexity of real disks and certainly fails to encompass the diversity of the models. The inner regions of putative quasar accretion disks are radiation pressure-dominated. The opacity is determined by the derived photospheric density, which depends upon the assumed viscosity prescription and (at temperatures of interest) can be dominated by electron scattering (e.g., Czerny & Elvis 1987) or bound-free transitions of hydrogen and helium (e.g., Ferland & Rees 1988, Laor & Netzer 1989).

We have adopted the assumptions used in a variety of published generic disk models to derive half-power radii. The emergent spectra are sensitive to the vertical structure and radiative transfer which, in turn, depend upon the prescription for angular momentum transport and energy release. Models in which the ratio of the shear stress to the total pressure $\alpha_V \gtrsim 0.1$ are usually too tenuous to cool via free-free and bound-free emission and a hot, Comptonizing corona may be produced. These disks will radiate predominantly in the X-rays and, if viewed directly, will not account for observed quasar spectra. Isothermal, homogeneous disks with $\alpha_V \lesssim 0.1$ and electron scattering behave roughly like gray bodies with $\epsilon \gtrsim 0.3$ and have similar sizes. Disks in which the shear stress is magnetic and limited in magnitude to the gas pressure (e.g., Sakimoto & Coroniti 1981, cf. Balbus & Hawley 1991) are effectively black (Laor & Netzer 1989). Provided that $r(\nu) \gtrsim 30GM/c^2$, it is fairly insensitive to the details of the energy release near the inner edge of the disk. Models in which an underlying power-law continuum is re-processed by a blanket of cool, dense gas clouds (Ferland & Rees 1988) also radiate nearly thermally and so have an effective emissivity at radius r approximately equal to their covering factor.

We also considered a scenario in which lens motion is small compared to the motion of a hot spot on an accretion disk. We must compare the variation of a microlensed disk with that produced by an unlensed control disk. A hot spot on an accretion disk will orbit the black hole with a speed $v \sim$

$35,000\, M_8^{1/2}\, r_{15}^{-1/2}\,\mathrm{km\,s^{-1}} = 120,000(r/r_{\rm min})^{-1/2}$ km s^{-1}, faster than any conceivable transverse velocity between the source, lens and observer, where M_8 is the hole mass in units of $10^8 \mathrm{M}_\odot$. Since accretion disks rotate differentially, hot spots will be transient, with brightness evolving on a disk-shear timescale, $\tau = 1.5(r^3/(GM))^{1/2} \sim 0.13 M_8 (r/r_{\rm min})^{3/2}$ d. We modeled this variation by supposing that several hot spots are formed on the disk surface. At regular intervals (of order 2τ), a bar was introduced at a random azimuthal angle, of width ~ 0.5 rad and between radii $[r/2, 2r]$. This perturbation was taken to increase the local effective temperature by a factor of up to five at peak, and from zero perturbation to peak and back to zero on a time scale of $\sim 4\tau \sim \tau_{orb}$. During this time, differential rotation sheared the bar into a spiral pattern. In this model the disk was most variable at frequencies $\sim \nu(r)$; at these frequencies the intrinsic disk variability was ≈ 0.05 mag. The precise form of the perturbations did not significantly influence the results.

2.3. APPLICATION TO Q2237+0305

The foreground galaxy was modeled following Wambsganss et al. (1990), i.e., as a planar distribution of point masses. A Salpeter mass function was assumed for the stars, $f(m)\,dm \propto m^{-2.35}\,dm$, with an upper mass cut-off at $10\,\mathrm{M}_\odot$ and a lower cut-off of 0.1 $\mathrm{M}_\odot$. The stars were placed uniformly and independently within a circular aperture whose size was such that the total flux of the source in microimages produced outside this area (or more importantly, its *variation*) was negligible. The general lensing equation for this situation is

$$\vec{\theta}_{src}(\vec{\theta}) = \begin{bmatrix} 1+\gamma & 0 \\ 0 & 1-\gamma \end{bmatrix} \vec{\theta} - \frac{4G}{c^2}\left(\frac{D_{LS}}{D_{OS}D_{OL}}\right) \sum_{stars} m_i \frac{(\vec{\theta}-\vec{\theta}_i)}{\left|\vec{\theta}-\vec{\theta}_i\right|^2}\,. \qquad (2.6)$$

Here γ is the normalized shear and m_i is the mass of the star at position $\vec{\theta}_i$; the average surface mass density (assumed to be all in stars) will be denoted σ. The values of σ, γ and angular distances D_{ij} ($\Omega = 1$) were chosen as those for the A image of Q2237+0305, as estimated by Schneider et al. (1988).

The spectrum observed from Q2237+0305 has an index $\alpha \sim 0.7$ (corrected for reddening) over the range 5×10^{14} Hz $\lesssim \nu \lesssim 2 \times 10^{15}$ Hz in the rest frame (Nadeau et al. 1991). Yee (1988) gives the dereddened i-band magnitudes for images A, B, C, and D of 16.89, 16.95, 16.94, and 17.44 in that order, for a total of 15.53. Dividing by the total estimated magnification of 15 ± 3 (Schneider et al. 1988 and Kent & Falco 1988) and converting to luminosity units, noting that the luminosity distance to this quasar is $d_L = 3 \times 10^{28}\ h_{75}^{-1}$ cm ($\Omega = 1$), gives $\mathcal{L} = 7 \times 10^{45}\ h_{75}^{-2}$ erg s^{-1}. With the assumed spectrum this implies a bolometric luminosity of $L \approx 3 \times 10^{46}\ h_{75}^{-2}$ erg s^{-1}. We would estimate the error in this value (at fixed h) as about a factor of two based on the uncertainty in the (unmeasured) spectral index in the UV region. With $M_B \approx -25.0 + 5 \log h$, Q2237+0305 is among the brightest 1% of quasars ($M_B < -21.5 + 5 \log h$) in its redshift range (see, e.g., Boyle et al. 1987).

We have adapted our disk models to fit the de-magnified image of Q2237+0305 (Fig. 2.2). The rather high luminosity of this object forced $\nu_{\max} \lesssim 10^{16}$ Hz, if one assumes $\epsilon \gtrsim 0.1$ and $\ell \lesssim 0.3$. Observations at a wavelength $\lambda_{obs} \sim 1\mu$, corresponding to a rest frequency $\nu \sim 10^{15}$ Hz found no color changes in image A when the variations occurred (Nadeau et al. 1991). The half-power radius of a blackbody disk model at this frequency is $\sim 8 \times 10^{15}$ cm. Other disk models are at least as large.

We have repeated the microlensing simulations of Wambsganss et al. (1990) and verify their conclusion that the source radius must be smaller than $\sim 2 \times 10^{15}$ cm in order to (typically) reproduce the independent micro-lensing variations in the four images, assuming a lower cut off in the stellar mass function of ~ 0.1 M$_\odot$. A smaller cut off mass requires an even smaller disk, although the dependence is weak. A larger mass implies an unreasonably large transverse velocity of the lens galaxy. Accretion disks are over three times too large to account for the microlensing variation. In other words the disk brightness temperature must be ~ 10 times the equivalent brightness temperature of a thermal disk, or $\gtrsim 2 \times 10^5$ K in the rest frame for the 1μ observations. This result is quite robust as it will only

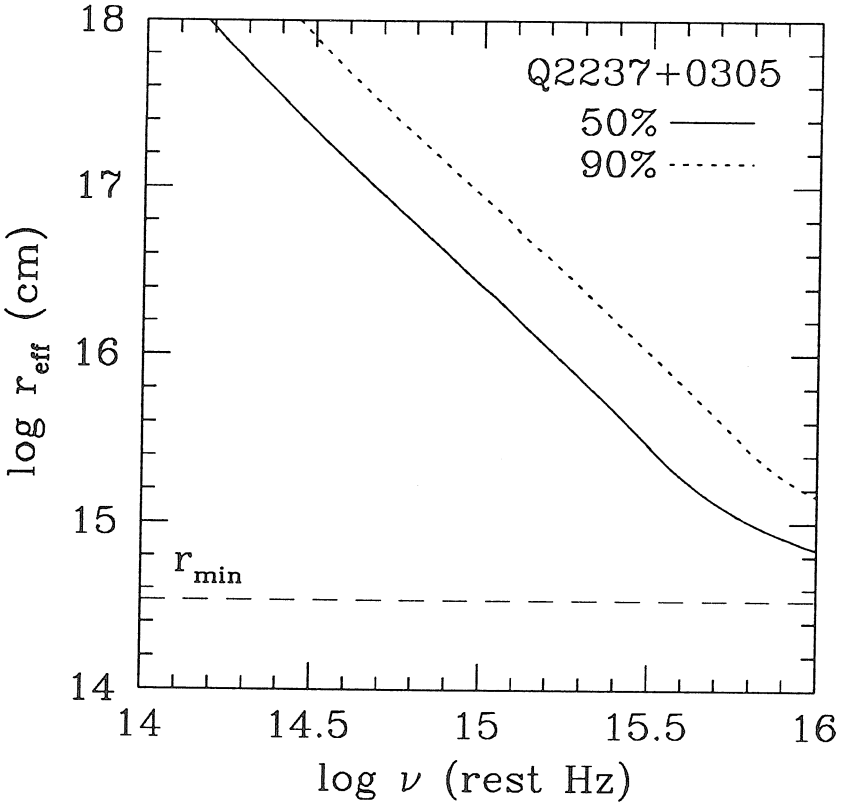

Fig. 2.2: Disk model with gray emissivity $\epsilon \simeq 0.1$, computed to reproduce the observed spectrum of the quadruple-imaged quasar Q2237+0305 (assuming $h_{75} = 1$). The half power disk radius at a rest frequency of $\sim 10^{15}$ Hz associated with the best documented microlensing variation is 2×10^{16} cm., approximately ten times larger than the maximum size necessary to account for these fluctuations. The effective radius of a blackbody disk is $\sim 7 \times 10^{15}$ cm.

be strengthened by a reduction in either the emissivity or the Eddington ratio. An unreasonably large increase in the magnification of all four macroimages to several thousand would be necessary to reduce the disk size to permit microlensing.

We have used a disk containing transient brightness fluctuations to compute its apparent light curve when placed behind the above microlensing screen of stars. To ensure adequate resolution, the ray-shooting density was increased until the results did not change. In addition, the position of the center of the accretion disk

relative to the aperture center was varied over a small grid; this served both to gauge the sensitivity of the variations to the specific area of the caustic pattern in which the disk resided as well as the change in magnitude of the source as a whole.

When the brightness fluctuations were limited to less than five times the (local) mean disk value, the resulting structure-induced light curve variations turned out to be very small, in most cases less than 0.1 mag. The maximum variation in any of the models, produced when two of the disk perturbations happened to cross caustics almost simultaneously, was about 0.2 mag, half the size of the observed variation in image A. It seems unlikely that such hot spots contribute significantly to the observed variations. We therefore increased the amplitude of the temperature fluctuations and varied their number and radial location until variations of $\sim$ 0.5 mag appeared about once or twice a year. This required $\sim$ 5 hot spots located at radius $r \sim 40 r_{\rm min}$ with peak temperature $T \sim 2 \times 10^6$ K, a factor 100 larger than the mean temperature in the smooth disk models at this radius. The spots themselves were $\sim 10^{15}$ cm in size. The high temperatures required again imply that non-thermal emission is required in the source if disk structure produces the observed variations.

We can understand the difficulty in producing large variability from N hot spots of angular size a small compared with the typical inter-caustic spacing b. If the spots move with speed v, then we expect the interval between high magnification events to be $\Delta t \sim b/(vN)$, the peak magnification relative to the mean flux to be $M \sim (b/a)^{1/2} N^{-1}$, and the duration of the event to be $\delta t \sim a/v \sim \Delta t/(NM^2)$. Hence the brightness temperature of each hot spot will be proportional to N^3, while the duty cycle of the high magnification events obeys $\delta t/\Delta t \sim N^{-1} M^{-2}$.

2.4. DISCUSSION

We have shown that the reported variation in the images of Q2237+0305 is not compatible with a thermal accretion disk and requires brightness temperature $\gtrsim 3 \times 10^5$ K at a blue rest wavelength. Although internal source motion may help contribute to the observed variations in image A of Q2237+0305, it is unlikely to be the major cause of them unless the source happens to lie exceptionally near a caustic. A longer set of observations should be able to rule out this possibility. We remark that of the four images, A is the least likely to show variations *a priori* since it has the smallest optical depth. We have examined caustic networks with parameters matching each of the four images, and in none of them is the typical intercaustic spacing smaller than or of order the maximum inferred size of the source. This observation is independent of whether internal structure or bulk source motion is producing the variations in A, since the caustic structure affects both of them equally. It is true that since caustics with shear tend to cluster, one would expect microlensing events in any specific image to cluster as well, but continued monitoring of the images is needed to settle this issue. The characteristic timescale to obtain representative light curves for all images is of order a century (Wambsganss 1991). Because of the small source size relative to the caustic spacing, the magnitude and rapidity of variations are insensitive to both the minimum star mass and the optical depth. What these parameters do determine is the typical frequency of events, by fixing the intercaustic spacing.

In our application to Q2237+0305, we have computed disk models as described in §2 that emit the observed spectral flux over the observed interval of rest wavelength $\sim 1 - 3 \times 10^{15}$ Hz. We find that the required disk emissivity varies $\propto r^{-q}$, where $2.3 \lesssim q \lesssim 2.5$. The Eddington number was limited to $\ell < 0.3$. For the black body model, we find $r(1\mu) = 7 \times 10^{15} h_{75}^{-1}$ cm. Microlensing simulations using a disk of this radius reproduced the reported variation (0.2 mag change in $\sim$ 1 month) typically a few times per century. A gray disk model with $\epsilon \sim 0.1$ had $r(1\mu) \sim 3 \times 10^{16} h_{75}^{-1}$ cm, and an electron scattering disk with $\alpha_V \sim 0.03$ and

bound free opacity as prescribed by, for example, Czerny & Elvis (1987) gave a radius of $\sim 9 \times 10^{15} h_{75}^{-1}$ cm but was unable to reproduce the spectrum.

Our most important deduction is then that the reported individual image variation in Q2237+0305 is unlikely to be produced by microlensing of a thermal (or electron scattering) accretion disk. The source is required to have a supra-thermal brightness temperature and yet not radiate excessively at higher frequency. A significant, non-thermal contribution to the optical emission would satisfy this requirement. Thin disk models are already known to have difficulty in accounting for the observed low orthogonal polarization (Stockman, Angel, & Miley 1979) and the observed absence of intrinsic Lyman continuum edges (Antonucci, Kinney, & Ford 1989). The present result constitutes a third serious objection. In view of the consequences for models of AGN, further monitoring of Q2237+0305 and similar sources is well motivated.

Acknowledgements

We thank Robert Antonucci, Ruth Corrigan, Paul Hewett, Joachim Wambsganss, and Howard Yee for advice and encouragement. Support by the National Science Foundation under grant AST89-17765, NASA under grants NAGW-1301, NAGW-2372 and an NSF graduate fellowship are gratefully acknowledged.

REFERENCES

Antonucci, R. R. J., Kinney, A. L., & Ford, H. C. 1989, ApJ, 342, 64

Balbus, S., & Hawley, J. 1991, ApJ, 376, 214

Bardeen, J. M., & Petterson, J. A. 1975, ApJ, 195, L65

Boyle, B. J., Fong, R., Shanks, T., & Peterson, B. A. 1987, MNRAS, 227, 717

Corrigan, R. T., Irwin, M. J., Arnaud, J., Fahlman, G. G., Fletcher, J. M., Hewett, P. C., Hewitt, J. N., Le Fevre, O., Mc Clure, R., Pritchet, C. J., Schneider, D. P., Turner, E. L., Webster, R. L., & Yee, H. K. C. 1991, AJ, 102, 34

Czerny, B., & Elvis, M. 1987, ApJ, 321, 305

Ferland, G. J., & Rees, M. J. 1988, ApJ, 332, 141

Irwin, M. J., Webster, R. L., Hewett, P. C., Corrigan, R. T., & Jedrzejewski, R. I. 1989, AJ, 98, 1989

Kent, S. M., & Falco, E. E. 1988, AJ, 96, 1570

Laor, A., & Netzer, H. 1989, MNRAS, 238, 897

Nadeau, D., Yee, H. K. C., Forrest, W. J., Garnett, J. D., Ninkov, Z., & Pipher, J. L. 1991, ApJ, 376, 430

Pringle, J. E. 1981, ARA&A, 19, 137

Rees, M. J. 1984, ARA&A, 22, 471

Reimers, D., Clavel, J., Groote, D., Engels, D., Hagen, H., Naylor, T., Wamsteker, W., & Hopp, U. 1989, A&A, 218, 71

Sakimoto, P., & Coroniti, F. V. 1981, ApJ, 247, 19

Sanders, D. B., Phinney, E. S., Neugebauer, G., Soifer, B. T., & Matthews, K. 1989, ApJ, 347, 29

Schneider, D. P., Turner, E. L., Gunn, J. E., Hewitt, J. N., Schmidt, M., & Lawrence, C. R. 1988, AJ, 95, 1619

Shakura, N. I., & Sunyaev, R. A. 1973, A& A, 24, 337

Shields, G. A. 1978, Nature, 272, 706

———. 1989, Ann N.Y. Acad Sci, 571, 110

Steidel, C. C., & Sargent, W. L. W. 1987, ApJ, 313, 171

Stockman, H. S., Angel, J. R. P., & Miley, G. K. 1979, ApJ, 227, L55

Wade, R. A., Hoessel, J. G., Elias, J. H., & Huchra, J. P. 1979, PASP, 91, 35

Wambsganss, J. 1991, private communication

Wambsganss, J., & Paczyński, B. 1991, AJ, 102, 864

Wambsganss, J., Paczyński, B., & Schneider, P. 1990, ApJ, 358, L33

Witt, H., Kayser, R., & Refsdal, S. 1993, A&A, 268, 501

Yee, H. K. C. 1988, AJ, 95, 1331

3

Optical Caustics in a Kerr Spacetime and the Origin of Rapid X-ray Variability in Active Galactic Nuclei†

Kevin P. Rauch and Roger D. Blandford

California Institute of Technology

ABSTRACT

A detailed analysis of the optical caustic structure around rotating black holes is given, including a discussion of their possible relevance to rapid X-ray variability in active galactic nuclei (AGN). It is found that the primary caustic surface takes the form of a small tube with an astroid (four-cusped) cross-section. At large distances behind the hole the tube is oriented parallel to the optic axis, but displaced transversely from it by a distance $a \sin\theta_0$, where a is the spin of the hole in geometrical units and θ_0 is the inclination angle of the distant observer from the pole. The transverse size s and central magnification $\mathcal{M}_0$ of the caustic are found to vary asymptotically as $s \sim 0.34\, a^2 \sin^2\theta_0\, r^{-1}$ and $\mathcal{M}_0 \sim [22M^{1/2}/(a^2 \sin^2\theta_0)]\, r^{3/2}$ (where M is the hole mass), respectively, in agreement with the results of an analytic, asymptotic analysis performed for $\theta_0 = \pi/2$. The magnification is surprisingly high everywhere within the caustic.

A fast, efficient, general-purpose code to trace geodesics in a Kerr metric is described, the calculation of the caustic structure of the Kerr metric representing an initial application. Sample light curves for point sources on geodesic orbits crossing the caustics are computed; in general, the lensing dominates all other features. The influence of caustics in computing the appearance of thick accretion

† First published in ApJ, 421, 46 (1994).

disk models is examined. The importance of caustics in real AGN depends on the geometry of the emitting matter; they could be important for disks and especially for jets, in which a significant fraction of the emitting material can reside near a caustic, for suitably oriented observers.

Subject headings: black hole physics – galaxies: active – accretion disks

3.1. INTRODUCTION

The analysis of null geodesics has proven to be an extremely useful technique for studying and describing black hole spacetimes. These investigations have often taken the form of thought experiments involving point sources of light moving near the event horizon, and have conveyed a good qualitative understanding of the geometry of a spinning black hole (e.g., Misner, Thorne & Wheeler 1973, hereafter MTW). They have done much to persuade physicists that, at least at the classical level, a black hole spacetime is physically realizable and poses no challenge to standard physical notions such as causality.

However, despite strong dynamical evidence (e.g., Remillard, McClintock & Bailyn 1992; Kormendy & Richstone 1992), there is still no rigorous proof that black holes actually exist in X-ray binaries or active galactic nuclei (AGN) and have the geometrical properties predicted by general relativity. Such a proof would almost certainly involve direct observation of matter in orbit near the event horizon exhibiting a peculiarly relativistic effect such as Lense-Thirring precession. Now X-ray observations of Galactic black hole candidates and, especially, X-ray loud Seyfert galaxies sometimes show variation on timescales that are not much larger than the time it would take light to cross a black hole of mass large enough to satisfy dynamical or radiative constraints. In other words, we are probably observing some radiation directly from the vicinity of a hole in these objects. This motivates more detailed studies of light propagation in a Kerr metric and the present paper constitutes one such examination.

Light rays are either deflected or captured when they pass near black holes. Black holes can therefore be considered to be gravitational lenses. However, unlike the lenses that are observed acting over cosmological distances (e.g., Blandford & Narayan 1992), black holes in X-ray binaries and AGN can be strong deflectors and the small angle approximation may be invalid. This does not change their qualitative optical properties as gravitational lenses, although it does require that quantitative measures of the lensing action be computed using the full equations of

general relativity. In particular, as we describe below, the caustic surfaces formed behind a Kerr hole have the same four-cusped cross-section as that associated with, say, an elliptical galaxy; however, the point source magnification is surprisingly large within and around this caustic surface.

In §3.2 of this paper, we summarize the connection between the null geodesic equation, the equation of geodesic deviation and caustic surfaces for the particular case of a Kerr metric described using Boyer-Lindquist coordinates. In the following section, we describe a new numerical approach for fast integration of the geodesic equations in this background. This method involves solving some of the differential equations that characterize the rays analytically in terms of elliptic functions and elliptic integrals; the details are given in Appendix A. Although our present concern is mainly with massless particles (i.e., photons), our code is also applicable to timelike geodesics and we make use of this generality in §3.4. We use this approach to compute the shapes of caustics (the loci of point sources whose images an observer sees, in the geometric optics limit, as being infinitely magnified) in a Kerr spacetime. We also compute and display magnification contours for stationary surfaces behind a black hole. In the weak deflection (large impact parameter) limit, these results are in agreement with a perturbative calculation given in Appendix B. In §3.4, we present the first of two illustrative applications, a calculation of the light curves that would be observed from point sources moving on relativistic geodesic orbits passing close to or through a caustic. The second illustration, in §3.5, consists of computing the fluxes that would be observed from thick and slender accretion disks as a function of inclination angle. Although these examples are somewhat artificial, they do give an understanding of the conditions under which emitting elements close to caustics can temporarily outshine relatively unmagnified sources. Finally, in §3.6 we consider possible astrophysical applications of these calculations.

3.2. NULL GEODESICS AND CAUSTIC SURFACES

When a point mass acts as a gravitational lens, light rays are deflected through an angle

$$\alpha = 4GM/bc^2, \tag{3.1}$$

where M is the mass of the lens and b, the asymptotic displacement distance of the ray from the optic axis, is the impact parameter. The optic axis is defined by the straight line extending from the observer through the mass. (Henceforth, we shall use units in which $G = c = M = 1$, so that the deflection angle is now $4/b$.) This expression is only valid in the weak deflection limit when the Robertson expansion of the metric tensor can be used. A source of small extent lying on the optic axis a distance r behind the lens will create an *Einstein ring* image of linear radius $b_E = 2r^{1/2}$ if we assume that the source is much closer to the lens than the observer. If the source is displaced from the optic axis by a small distance ζ, then two diametrically opposite tangential arcs will be created with impact parameters b satisfying $b^2 - \zeta b - b_E^2 = 0$, one just inside the Einstein ring, the other just outside. As surface brightness is conserved (we ignore cosmological corrections for the moment), the flux magnification due to the lens is simply the ratio of the solid angle subtended by the images to that subtended by the source in the absence of the lens at the observer. It is easily calculated to be

$$\mathcal{M} = \left[1 - \left(\frac{b_E}{b}\right)^4\right]^{-1} \tag{3.2}$$

(e.g., Schneider, Ehlers & Falco 1992). The magnification $\mathcal{M} \to \infty$ as $\zeta \to 0$.

If the deflection angle is $O(1)$, then the impact parameter necessarily becomes comparable with the gravitational radius (equal to unity with our units) and the ray must be traced through the general relativistic spacetime. Also, we can no longer treat the mass as a point—it can be no more concentrated than a black hole. The simplest case to consider is the Schwarzschild geometry describing a non-rotating black hole. If we work in Schwarzschild coordinates, then the deflection

angle α can be computed as a function of impact parameter in the integral form

$$\alpha = \int \frac{du}{(b^{-2} - u^2 + 2u^3)^{1/2}} - \pi, \tag{3.3}$$

which can be evaluated analytically in terms of an elliptic integral of the first kind, $\mathrm{K}(\phi|m)$ (Darwin 1959; cf. Fig. 25.7 of MTW).

Non-singular, circularly symmetric, transparent lenses behave slightly differently. Suppose that the lens is axisymmetric on the observer's sky but has a finite core radius with central surface density Σ. Sources more distant than $r_{min} = (4\pi\Sigma)^{-1}$ can be multiple-imaged if they are close enough to the optic axis. On displacement away from the optic axis, the images will be tangentially stretched and magnified, as for a point mass lens. Sources ever more distant from the axis will be radially stretched as they approach the *radial caustic* of the lens and the corresponding bright images converge to the *radial critical curve*, which lies within the Einstein ring (the *tangential critical curve*). The radial caustic is a circle. The tangential caustic remains a degenerate line coincident with the optic axis. Either one or three collinear images will be created by a transparent lens. (It is possible to create more than three images under special circumstances, but this does not happen generically.)

Now, let us break the circular symmetry by giving the gravitational field an elliptical perturbation with ellipticity ϵ. The most immediate consequence is that the images will no longer be collinear. A source located close to the optic axis and sufficiently far from the lens will create four images straddling the tangential critical curve and a fifth one, usually quite faint, close to the optic axis. If the source moves in the vicinity of the optic axis, the four bright images will move around the tangential critical curve, occasionally brightening and coalescing or re-emerging in pairs. The locus of the source positions where these images coalesce is the tangential caustic which has now unfolded from a line to a surface of cross-sectional width $\sim \epsilon b_E \propto \epsilon r^{1/2}$. This surfaces has four cusped *ribs* roughly parallel to the optic axis and is similar to an astroid curve in cross-section. As the source crosses this surface and the two images appear or disappear, the magnification

diverges inversely as the square root of its distance from the caustic. This is known as a *fold catastrophe.* When the source approaches one of the cusped ribs, three images are transformed into one image or *vice versa* and this is known as a *cusp catastrophe* (e.g., Berry & Upstill 1980).

An alternative view of caustic surfaces, which we shall exploit in this paper, is as singularities in bundles or *congruences* of rays. Consider a ray congruence traveling backward in time from the observer. The rays will be sheared (though not converged) by the matter-free spacetime through which they propagate. Eventually the cross section of the congruence will collapse to a line. The place where this happens is said to be *conjugate* to the observer. Rays are tangent to the caustic surface at the conjugate point so the surface can be thought of as the envelope formed by all of the rays leaving the observer.

When the gravitational lens is a Schwarzschild black hole, the core radius disappears, along with the central demagnified ray (which must cross the event horizon), and the radial caustic. Furthermore, as the lens is circularly symmetric, the tangential caustic degenerates to a line. However, there is a new effect (discussed in MTW): the formation of multiple caustics. All rays with impact parameter $b \geq b_1 = 4.4573$ will cross the optic axis behind the hole and form a tangential line caustic outside the horizon. Rays with impact parameter in the range $b_2 = 5.1566 \leq b \leq \sqrt{27} = 5.1962$ will circle the hole and form a second caustic on the optic axis on the observer side of the hole. Rays with impact parameter satisfying $b_3 = 5.1944 \leq b \leq \sqrt{27}$ will form a third caustic behind the hole and so on. The number of caustics formed will tend to infinity as the impact parameter approaches the critical impact parameter for capture, $b_c = \sqrt{27}$, and these rays will circle the hole many times close to the *photon orbit* at $r = 3$. There is a corresponding sequence of unbound rays with $\sqrt{27} \leq b \leq b_2' = 5.3570$, etc., which also form multiple caustics. The existence of multiple conjugate points is illustrated in Fig. 3.1, which plots three rays of equal impact parameter (specifically, $|b| = 5.1980 < b_3' = 5.2027$) but differing impact angles (measured relative to an arbitrary axis in the plane perpendicular to the observer's line of sight); in

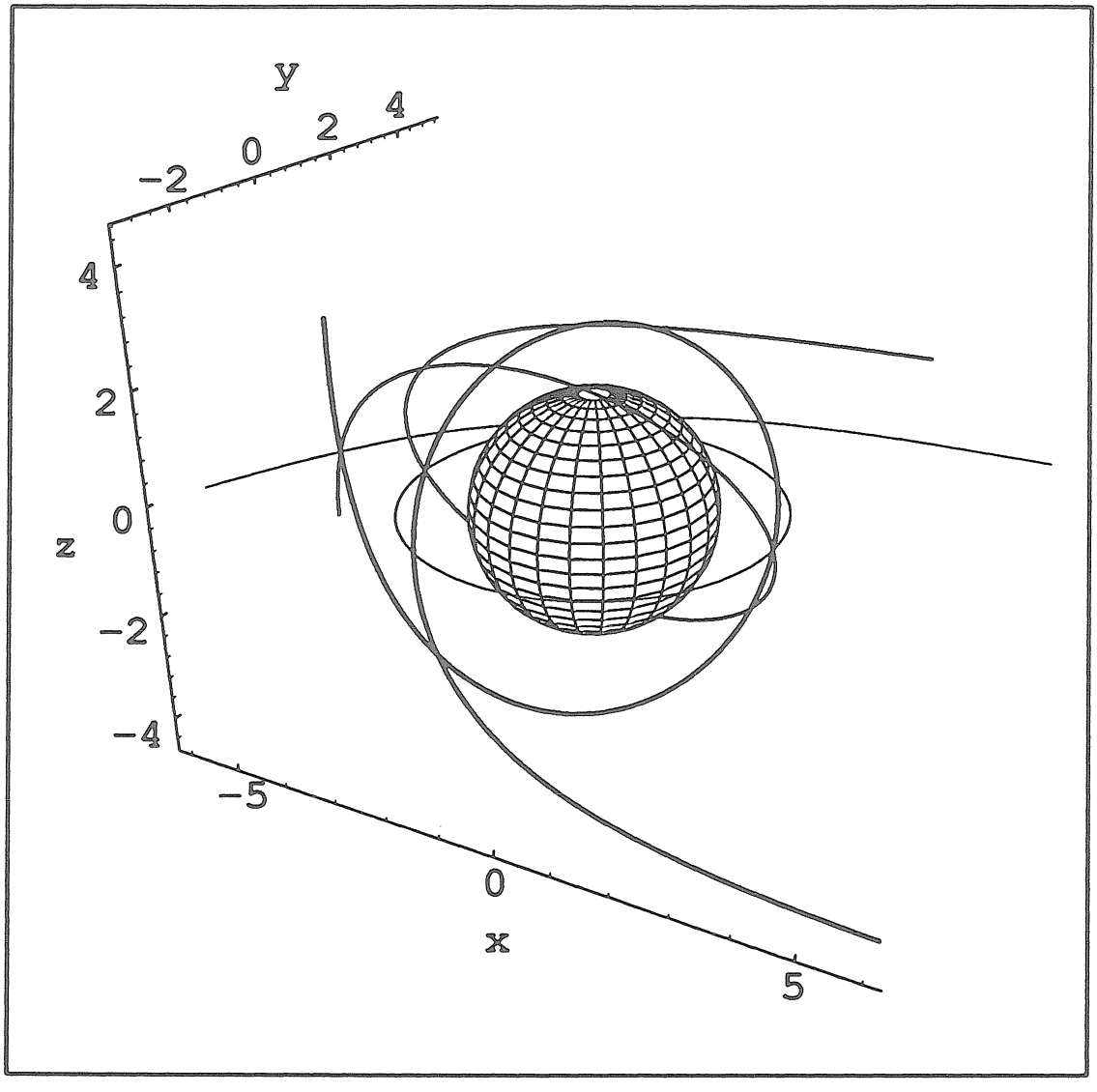

Fig. 3.1: A plot of three rays (distinguished by differing line thicknesses) with impact parameter vectors of equal magnitudes (specifically, $|b| = 5.198$) but differing directions traveling around a Schwarzschild black hole. Points conjugate to a distant observer (residing at $x = \infty, y = 0, z = 0$) are produced where the rays converge; three are produced for this set of rays. See §3.2.

this particular case three conjugate points, where the rays are focused and cross, are produced (the observer lies at $x = \infty$, $y = 0$, $z = 0$, where the three rays also 'converge'). The sphere marks the event horizon of the Schwarzschild hole. There are, therefore, an infinite number of caustic surfaces for any observer. However, although a point source located on any of these caustics will be infinitely magnified (in the geometric-optics limit), the volume of the high magnification region diminishes rapidly as the order of the caustic increases and, in practice, it is only the primary caustic behind the hole that is likely to be associated with large brightness fluctuations.

Finally, let us turn to the Kerr metric. We again break the circular symmetry and the line caustics of the Schwarzschild hole develop astroid cross sections, just as with a Newtonian elliptic potential. There is, however, one additional effect to be anticipated: the displacement of the caustics (both primary and secondary) from the optic axis by the dragging of inertial frames. These effects will now be demonstrated.

3.3. RAY-TRACING AND CAUSTIC STRUCTURE

3.3.1. *Numerical Method*

Throughout this paper, we assume the geometric-optics limit is applicable, so that all photon trajectories are described by the usual null geodesics. Since integration of the equations of geodesic deviation requires extensive computation of the geodesics being followed, considerable effort was devoted to designing an efficient and accurate code to compute arbitrary (timelike or null) geodesics in a Kerr metric. In particular, the equations of motion were recast as definite integrals in a modified Boyer-Lindquist coordinate system (in which $r \rightarrow u \equiv 1/r$ and $\theta \rightarrow \mu \equiv \cos\theta$), and the associated equations of motion were integrated analytically (in terms of elliptic functions and integrals) wherever practical. Geodesics are parameterized using a pseudo-radial variable which we denote by $\tilde{u}$, monotonic with respect to the affine parameter, which is defined to be simply u unless u is

multiple-valued (as for bound orbits), in which case $\tilde{u}$ is the total (absolute) range of u traversed as the affine parameter is varied monotonically (see Appendix A). In short, this allows $\mu(\tilde{u})$ to be computed analytically. The remaining integrals (for $\phi(\tilde{u})$ and $t(\tilde{u})$) can be written, in principal, in terms of elliptic integrals of the third kind and elementary functions; however, it is faster (and *much* easier) to evaluate them numerically. A Gauss-Kronrod (fixed abscissa Gaussian-style; e.g., Piessens et al. 1983) integration scheme tailored to the general form of the integrals was implemented to obtain both high accuracy and maximal speed. In addition, the program employs several transformations in certain instances (e.g., for geodesics which travel many times around the black hole before escaping to infinity) to maintain accuracy. The resulting routines are very robust and have been extensively tested. The numerical implementation is detailed in Appendix A.

The code was tested in several ways. To test the reduction of the integrals to elliptic functions and integrals, two complete versions of the code, one performing the integrals numerically and one analytically, were written and checked for agreement by tracing sample geodesics (over 10^7 in all, spread more or less uniformly through the parameter space) and comparing the results. Results in the Newtonian and Schwarzschild limits (e.g., orbital precession, orbital times, deflection angles, and time delays) were also checked, as were Kerr-specific results such as Lense-Thirring precession rates for bound orbits and known analytic results (e.g., radii and impact parameters for photon orbits in the equatorial plane). However, the most stringent verification of the code, in our opinion, is the fact that the computed shapes of the caustic surfaces are consistent with the expectations of catastrophe theory and agree with the asymptotic analytic results derived in Appendix B.

3.3.2. *Location of Caustic Surfaces*

To map out the caustic surfaces, the geodesic deviation equations were integrated numerically in a form compatible with the ray-tracing routines. In particular, the metric and connection coefficients (and their derivatives) were computed

in the modified coordinate system and the affine derivatives $d/d\lambda$ were replaced with $d/d\tilde{u}$. Explicitly, the deviation equations become

$$f^2 \frac{d^2 Y^\alpha}{d\tilde{u}^2} + \left[\delta^\alpha_\gamma k^\epsilon f_{,\epsilon} + 2f\Gamma^\alpha_{\beta\gamma} k^\beta\right] \frac{dY^\gamma}{d\tilde{u}} + \Gamma^\alpha_{\beta\delta,\gamma} k^\beta k^\delta Y^\gamma = 0. \qquad (3.4)$$

Here δ^μ_ν is the Kronecker delta, $f(\tilde{u}) \equiv u_{sgn}\rho^{-2}\sqrt{U}$ (i.e., $d/d\lambda = f(\tilde{u})d/d\tilde{u}$; see Appendix A), and Y is the separation 4-vector. These equations can be reduced to a system of first-order ODEs in the usual manner. One disadvantage of using $\tilde{u}$ as independent variable in this situation is that the derivatives $d/d\tilde{u}$ become infinite at a turning point of u; in this situation we rescale $\tilde{u}$ into a new coordinate which removes the square-root divergence in f. We used the Bulirsch-Stoer method (Press et al. 1992) to integrate the ODEs with an accuracy of $\sim 10^{-7}$.

We note that the locations of the caustics can also be computed using the optical focusing equation (Schneider et al. 1992). Although it is simple in abstract form, numerically it is somewhat less convenient to work with than the deviation equations, in particular because the focusing equation requires calculation of the Weyl tensor (the trace-free part of the Riemann curvature tensor) and the Ricci tensor when computing the value of the shear along the ray, whereas the curvature tensor drops out of the deviation equations almost completely, leaving only a derivative of the connection coefficients, as can be seen in equation (3.4).

The caustics are located by propagating two initially orthogonal, spacelike separation vectors Y_1 and Y_2 backward in time along a ray until they become linearly dependent, as measured relative to a pair of orthonormal basis vectors (perpendicular to the ray) which are parallel-transported along the ray. At this *conjugate* point, the magnification, which is inversely proportional to the area spanned by the separation vectors, becomes infinite. There are two ways to set up the two initial separation 4-vectors Y_1^α and Y_2^α. For distant observers ($u_0 \ll 1$), space is essentially Minkowskian and we take the initial $Y_{1,2}$ to be the orthonormal basis vectors themselves, with the initial derivatives $dY/d\tilde{u} = 0$ (conceptually this makes infinity the starting conjugate point, where the ray bundle converges). For relatively nearby observers ($u_0 > 10^{-3}$), we do the opposite, taking $Y_1 = Y_2 = 0$

initially (the usual way to make the observer be one conjugate point on the ray) and their derivatives equal to the initial orthonormal basis vectors. The results presented below are all for distant observers, the relevant case for our applications. We computed a number of caustic slices for observers at a modestly distant $r_0 = 1000$ from the hole for comparison; the results were extremely similar, the caustics for the closer observer being slightly larger and displaced from the caustics for the distant observer.

3.3.3. *Caustics for Equatorial Observers*

For the distant observer we take $u_0 = 0$ and there are then two parameters left to vary in computing the caustic surfaces, the observer inclination θ_0 and the spin of the hole a, which we take to be non-negative without loss of generality. The qualitative results are as follows. As described in §3.2, for $a = 0$ (Schwarzschild geometry) there is symmetry about the line of sight and the caustic surface degenerates into a line (the optic axis) on which point sources are seen infinitely magnified, as an Einstein ring. For $a > 0$, the symmetry is broken and the caustics are no longer degenerate; we find that the line becomes a small tube extending behind the hole, with a simple astroid cross-sectional shape, more or less distorted depending on the distance of the caustic behind the hole. For observers in the equatorial plane of the black hole, $\mu_0 = 0$, the cross-sectional area is maximized, as is the transverse displacement of the caustic tube from the optic axis (which, we find, approaches a far from the hole, in agreement with the asymptotic analysis presented in Appendix B), for any given a; the area also increases monotonically with a. As the observer inclination increases, the size and displacement both decrease, until the tube again degenerates onto the optic axis for observers on the pole, where there is rotational symmetry about the line of sight. Near the horizon, the caustic tube is 'stretched' perpendicular to the equatorial plane and spirals into the horizon like a winding ribbon, making an infinite number of revolutions before crossing the horizon as $a \to 1$ (but a finite number for $a < 1$). The (primary) caustic surface for the maximally rotating case is shown in Fig. 3.2. In this

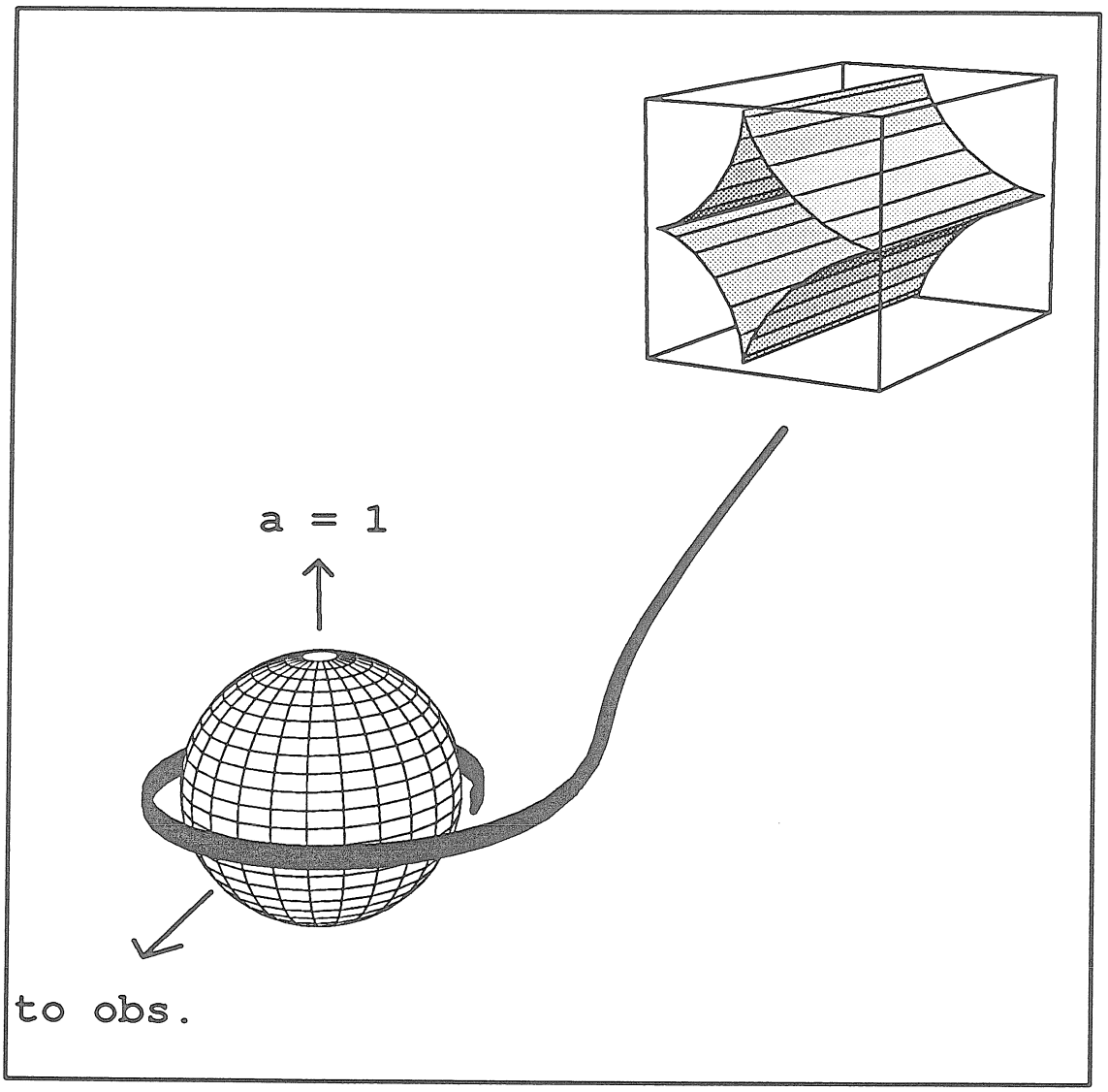

Fig. 3.2: A representation of the primary caustic surface in an extreme Kerr geometry for a distant equatorial observer. The sphere denotes the event horizon. Formally, the caustic winds around the horizon an infinite number of times for this special case ($a = 1$) before crossing it; only one revolution is shown. The inset box shows a magnified view of a small section of the caustic. The surface is a small tube with a 4-cusped cross-section.

figure, the sphere represents the event horizon and the inset box gives a magnified view of a short section of the caustic tube, in which the astroidal cross-section can be easily seen.

3.3.4. *Angular Magnification of a Point Source*

For $\mu_0 \neq 0$, we present the quantitative results as a set of representative caustic slices (tube cross-sections) and magnification contours plus a set of figures showing scaling relations for the displacements from the optic axis and widths of the caustics as a function of a, μ_0, and x'. The slices are plotted using a pseudo-Euclidean coordinate system related to Boyer-Lindquist coordinates by $(x, y, z) \equiv (r\sin\theta\cos\phi, r\sin\theta\sin\phi, r\cos\theta)$, plus a rotation about the y-axis for observers out of the equatorial plane, $(x', y', z') = (x\sin\theta + z\cos\theta, y, z\sin\theta - x\cos\theta)$, so that the observer always lies along the x'-axis (at $x' = \infty$) and sizes, etc., are measured in $y' - z'$ planes perpendicular to the line of sight (planes of constant x'). The line $y' = 0, z' = 0$ defines the optic axis. In Figs. 3.3*a*-*d*, caustic slices for the case of an equatorial observer in an extreme Kerr ($a = 1$; cf. Fig. 3.2) metric are plotted for $x' = -100, -40, -5, 0$; Fig. 3.4 shows a plan view of the intersection of the same caustic surface with the equatorial plane near the horizon, showing the first three revolutions of the ribbon-like caustic around the hole (the inset panel shows the caustic tube at greater distances). Figs. 3.5*a*-*c* show iso-magnification contours corresponding to the slices in Figs. 3.3*a*-*c*, respectively; these are the angular magnifications (of apparent flux) that a point source at that position at rest in the Boyer-Lindquist frame would have, due to the combined effects of multiple-imaging and magnification of each individual image. For moving sources, the angular magnification must be combined with the variable gravitational redshifts and Doppler shifts due to the source motion (obligatory inside the ergosphere). Note that *the magnifications are high everywhere inside the caustic*; this is the result we most wish to emphasize. This point is made clear by Figs. 3.6*a*-*d*, which present the minimum magnification inside the caustic as a function of distance behind the hole, for three values of observer inclination

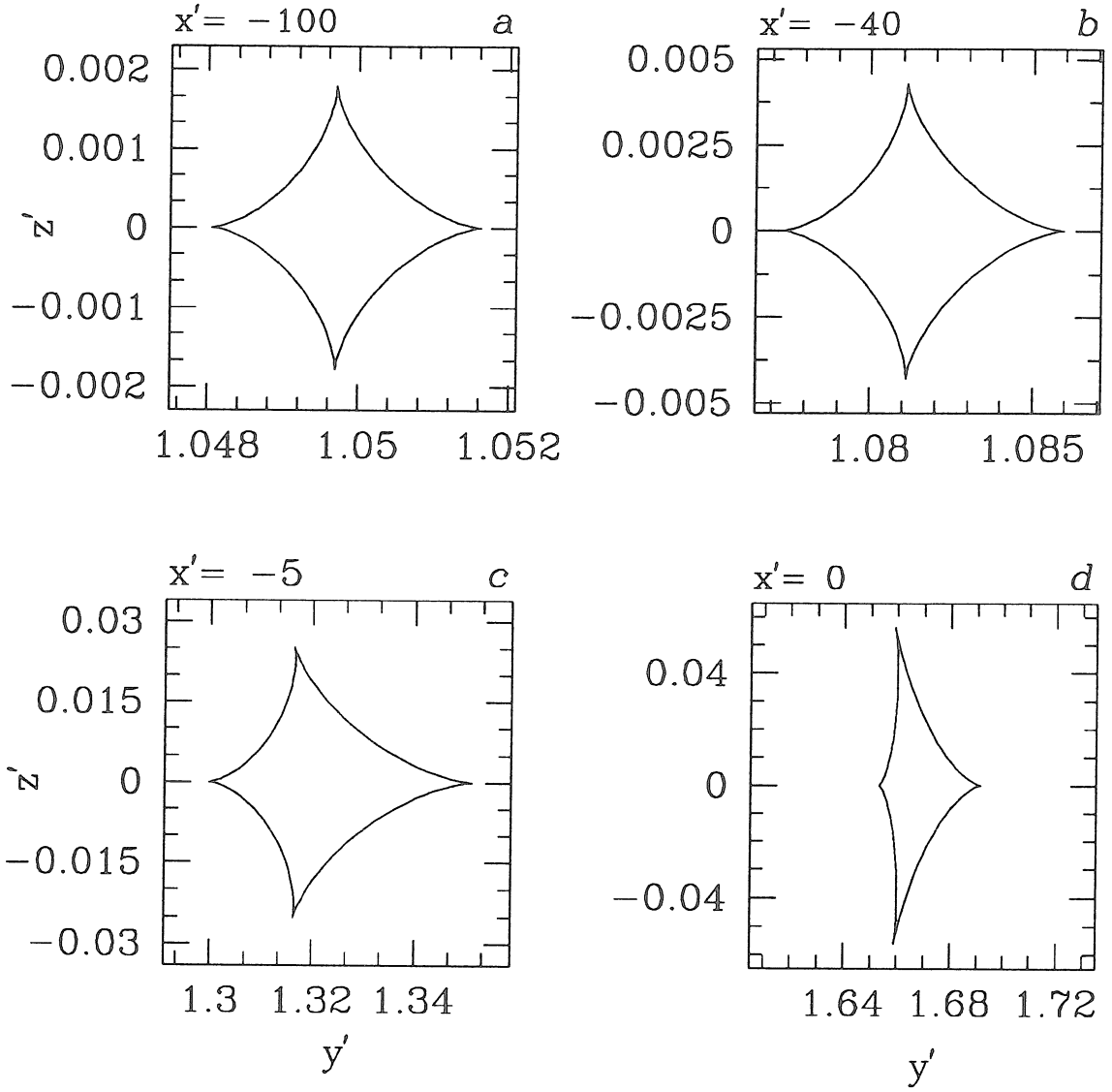

Fig. 3.3: Shown in each figure is a cross-sectional slice, at a fixed distance behind the hole ($x' = -100, -40, -5, 0$ for Figs. 3.3*a*, 3.3*b*, 3.3*c*, and 3.3*d*, respectively; all lengths are in units of the hole mass), of the caustic surface of a maximally rotating black hole and distant equatorial observer. The relative scales on each graph are equal ($dy'/dz' = 1$) to show proper proportions; note how the caustic becomes thin as it approaches the horizon (Fig. 3.3*d*; cf. Fig. 3.4).

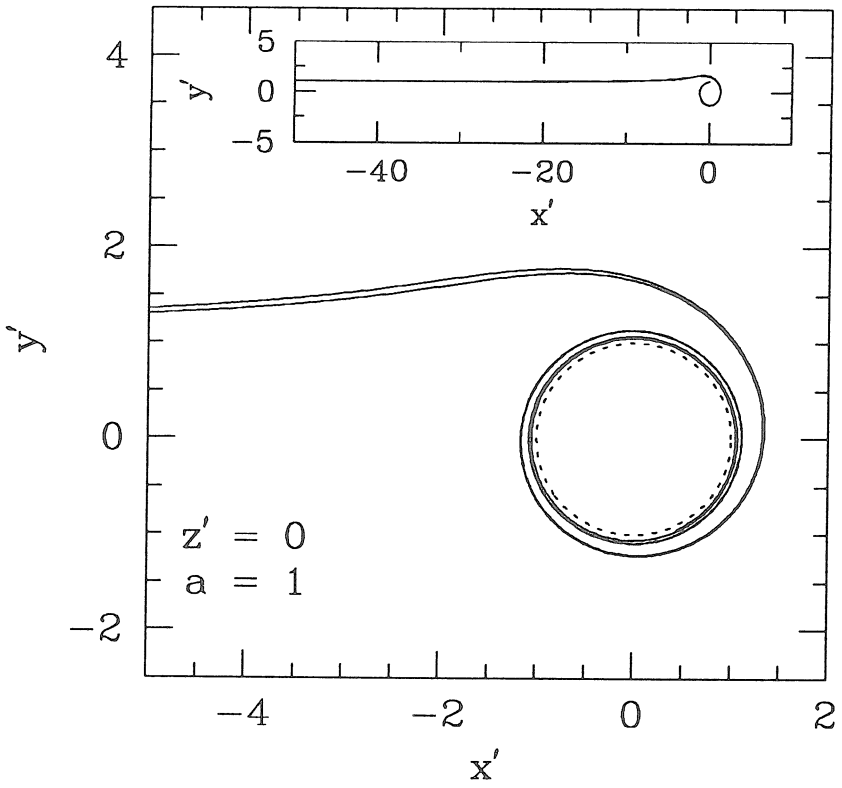

Fig. 3.4: A plan view of the same caustic surface plotted in Figs. 3.3*a*-*d*, showing the behavior of the caustic near the horizon (represented by the dotted circle), where the caustic becomes ribbon-like as it winds slowly towards the horizon (only a few revolutions are shown). The inset panel shows a macroscopic view of the caustic tube.

from the pole ($\theta_0 = 30°, 60°$, and $90°$), each figure having a fixed value of a ($= 0.1, 0.5, 0.866$, and 1, respectively). The noise at large distances is numerical error. We find that the magnification scales asymptotically as

$$\mathcal{M}_0 \sim 22\, r^{3/2}/(a^2 \sin^2\theta_0); \quad r \gg 1. \tag{3.5}$$

Also note (in Figs. 3.5*a*-*c*) that *outside* the caustic, the magnification is very high only near the cusps, which we justify in Appendix B.

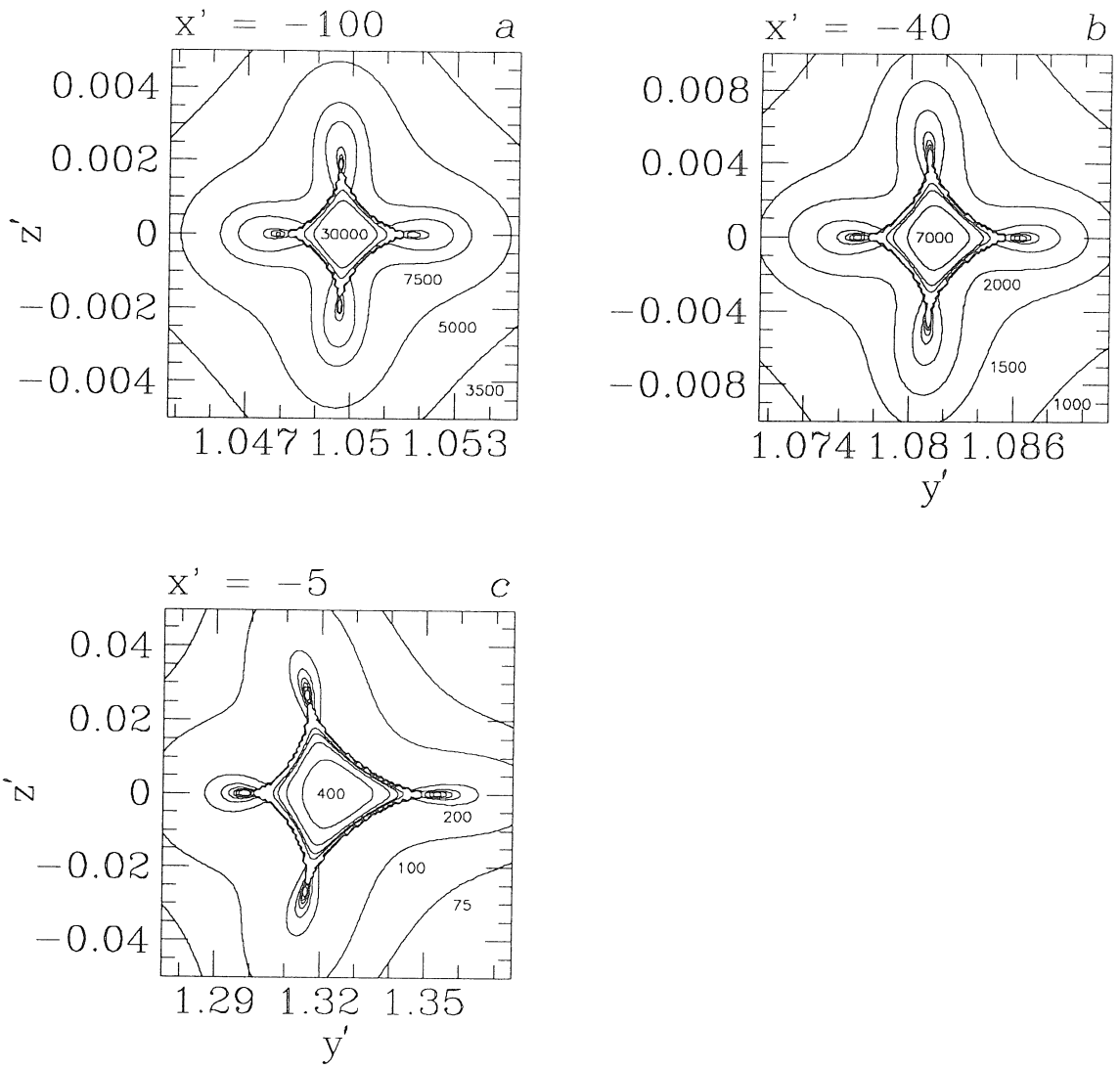

Fig. 3.5: Figs. 3.5*a*-*c* show the same caustic slices (although on slightly different scales) as Figs. 3.3*a*-*c*, with $x' = -100, -40, -5$, respectively, now superimposed with contours of constant angular magnification (see §3.4). In Fig. 3.5*a* the contour levels are $\mathcal{M} = 3500$, 5000, 7500, 10000 and by 10000 thereafter. In Fig. 3.5*b* the levels are $\mathcal{M} = 1000$, 1500, 2000 and by 1000 thereafter; in Fig. 3.5*c* the levels are $\mathcal{M} = 75$, 100 and by 100 thereafter. The magnification is high (> 100) everywhere inside these caustic sections (cf. Fig. 3.6*d*); outside them it is *very* high only near the cusps, although still large over an extended region.

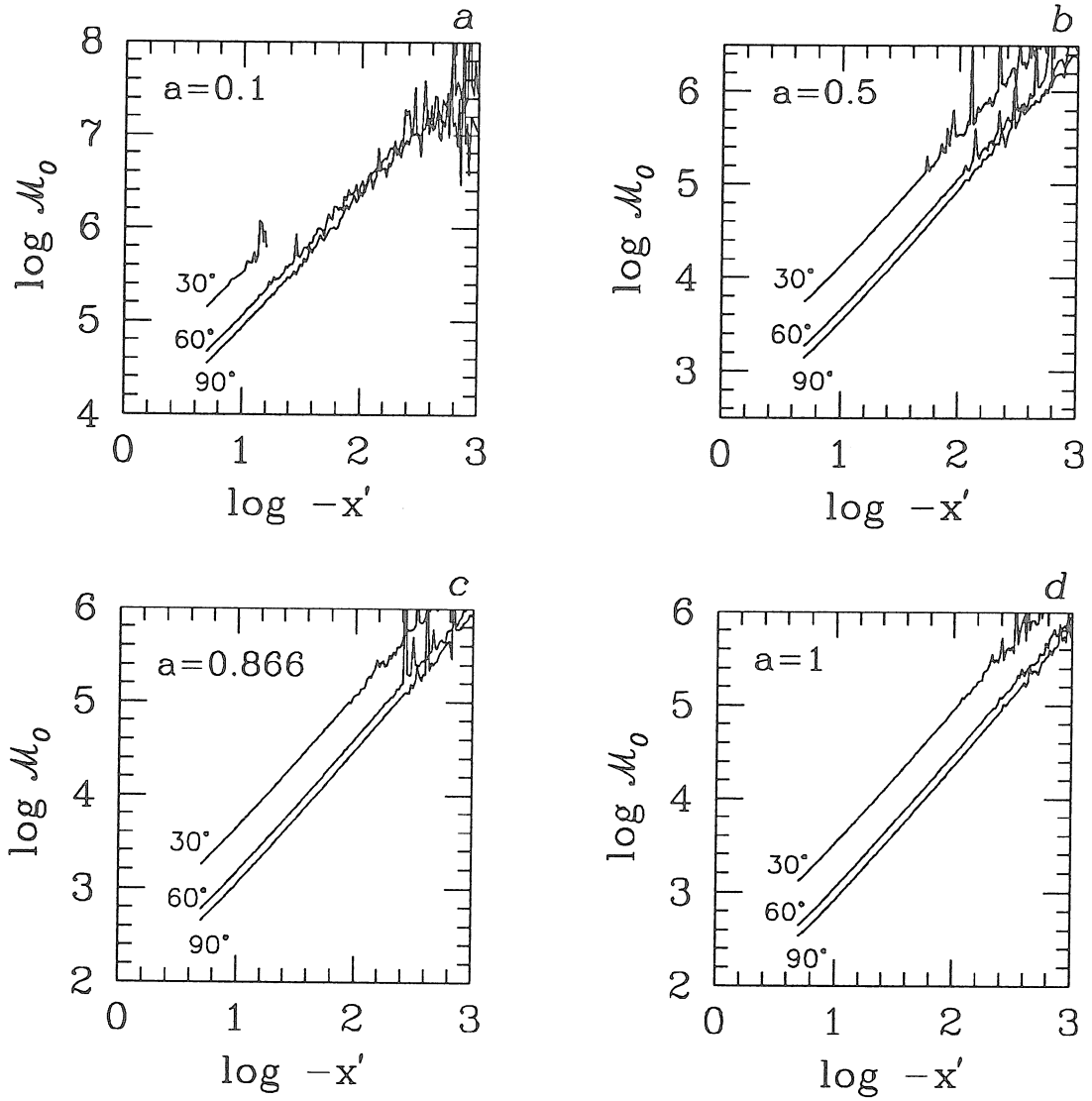

Fig. 3.6: Each figure plots the *minimum* magnification inside the caustic as a function of distance x' behind the hole for several values of observer inclination from the pole ($\theta_0 = 30, 60, 90$ degrees) for a fixed value of the spin of the hole ($a = 0.1, 0.5, 0.866, 1$, respectively). Asymptotically the magnifications vary as $\mathcal{M} \sim 22M^{1/2}r^{3/2}/(a^2 \sin^2\theta_0)$ (the fluctuations at large radii are numerical noise.)

3.3.5. *Displacement from the Optic Axis and Transverse Sizes*

Figs. 3.7*a-d* are analogous to Figs. 3.6*a-d*; each plots the displacement of the caustic center, (y_0', z_0'), for four values of observer inclination from the pole ($\theta_0 = 1°, 30°, 60°$, and 90°), each figure having a fixed value of a ($= 0.1, 0.5, 0.866$, and 1, respectively). Clearly z_0' is small in all cases, i.e., the caustics are nearly centered on the y'-axis. The asymptotic behavior of y_0' is found to be $y_0' \sim a \sin\theta_0$ for $|x'| \gg 1$; the behavior of y_0' near the horizon can be read off of Fig. 3.4.

Figs. 3.8*a-d* are analogous to Figs. 3.7*a-d*, only now showing the cross-sectional sizes $\Delta y'$ and $\Delta z'$ instead of the displacements. The numerical noise at large radii, also present in Figs. 3.6*a-d*, is caused both by the caustics becoming very small (worsening the relative error) and by the cumulative buildup of error in the integration of the deviation equations. Note that $\Delta y' \approx \Delta z'$ at all plotted radii (although near the horizon $\Delta y' \ll \Delta z'$; cf. Figs. 3.3*d* and 3.4). Asymptotically the total widths satisfy

$$s \sim \Delta y' \sim \Delta z' \sim \frac{0.34\, a^2 \sin^2\theta_0}{r}; \quad r \gg 1. \tag{3.6}$$

Note that since $s \propto r^{-1}$ for large r, the caustic surface encloses a finite volume; this expresses the rapid fall-off with distance of the influence of the hole's spin.

The uncertainty in the numerical coefficients in equations (3.5) and (3.6) is a few percent.

3.3.6. *Higher Order Caustic Surfaces*

As discussed in §3.2, there exist multiple caustic surfaces for any given a and μ_0, which are associated with rays traveling multiple times around the hole (near the photon orbit) before being captured or escaping to infinity. In the Schwarzschild case, these higher order caustics are located on the optic axis (both behind *and* in front of the hole), whereas for general a they are found at all azimuthal angles. They appear to have the same astroid cross-section as primary caustics. We have not attempted to map their locations extensively because, unlike

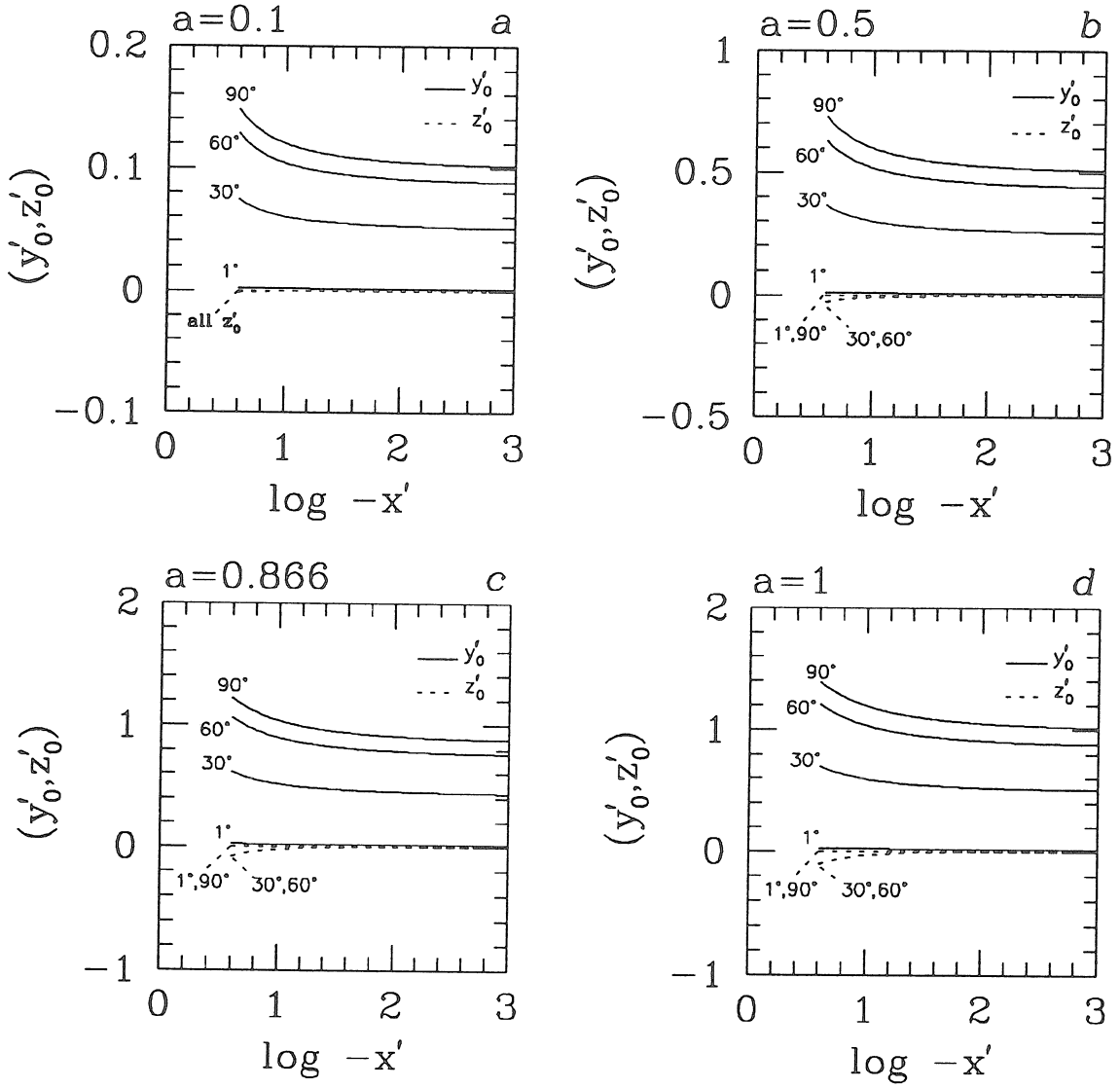

Fig. 3.7: Analogous to Figs. 3.6*a-d*, now plotting the coordinates of the caustic center (relative to the optic axis). The coordinate z_0' of the center is always small. The derived asymptotic value for y_0' is $y_0' \sim a \sin\theta_0$, $-x' \gg M$, which agrees with the analytic asymptotic results derived in Appendix B.

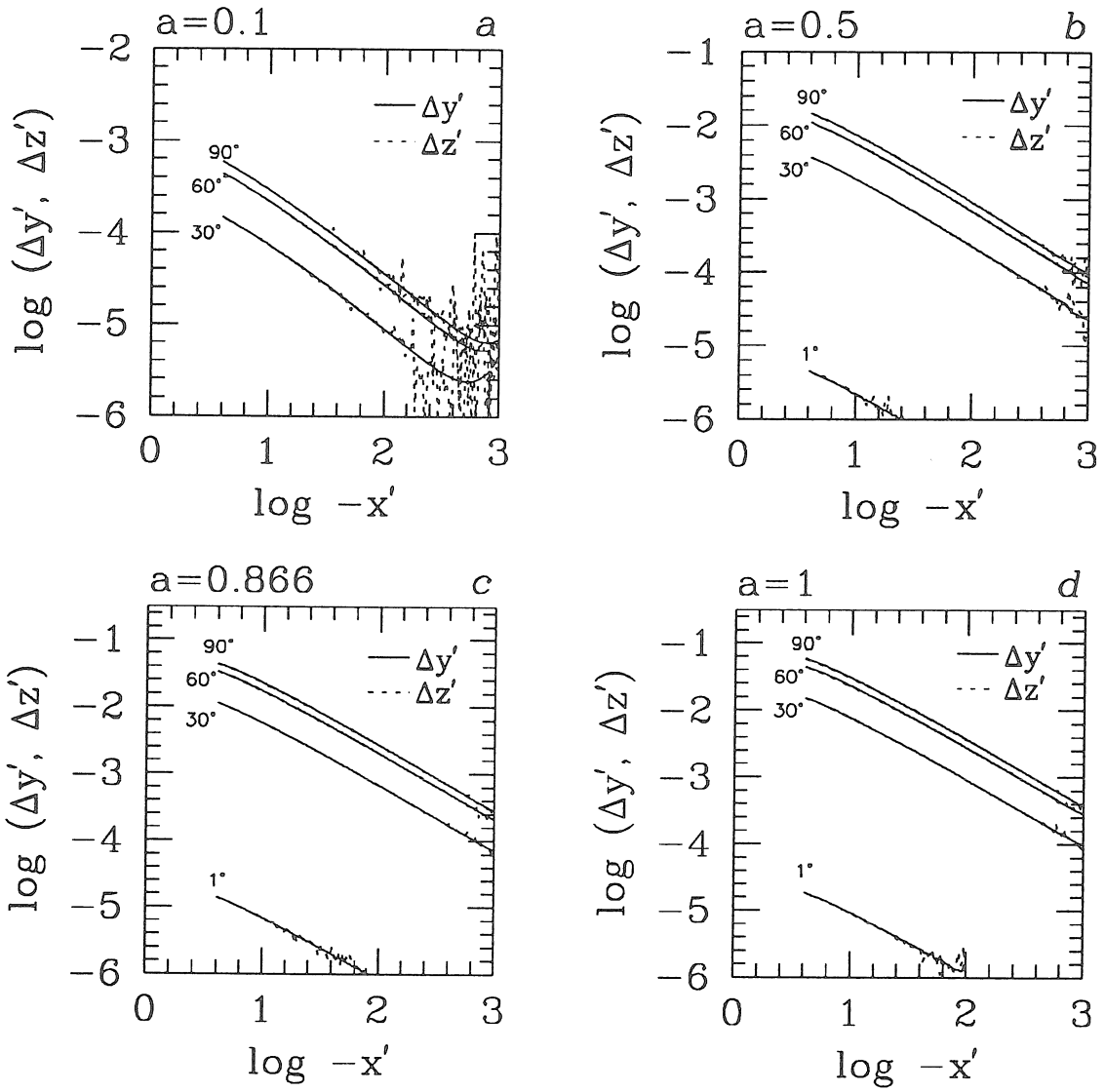

Fig. 3.8: Analogous to Figs. 3.6*a-d*, now showing the cross-sectional widths $\Delta y'$ and $\Delta z'$ of the caustics. Asymptotically the widths become $s \sim \Delta y' \sim \Delta z' \sim 0.34(a^2 \sin^2 \theta_0)\, r^{-1}$. The deviations from a power law at large radii are due to numerical noise.

the primary caustic, the image magnifications are *not* large everywhere inside the secondary caustics. This is due to the fact that near the photon orbits, small changes in impact parameter produce large changes in deflection angle, causing ray congruences to defocus very quickly—the images are highly demagnified except for sources *very* close to such caustic surfaces. The volume within which sources close to higher order caustics suffer greater than unit magnification was found to be negligibly small.

3.4. LIGHT CURVES OF ORBITING POINT SOURCES

To examine the impact of the caustics on X-ray variability, we have calculated a series of light curves of point sources following geodesic orbits around the hole, some passing inside or near the caustic and others well outside it. The general magnification of orbiting sources due to lensing has been examined many times (e.g., Cunningham & Bardeen 1973; Karas, Vokrouhlický, & Polnarev 1992 and references therein); however, since these authors did not calculate the detailed caustic structure, they did not study the highest magnification regions, nor did they compute the probability of encountering these high magnifications. Determining the possible influence of these regions on the observed total emission (from AGN in particular) is our motivation for doing the calculations described in this section.

Computing observed light curves is straightforward in principle. First, one must find all images (of which there are infinitely many, strictly-speaking) of the source at a given location; then for each one, compute both its angular magnification and overall relativistic (Doppler plus gravitational) red- or blueshift. Since each image has a different time delay, the light curves for each individual image must be combined asynchronously to produce the light curve as a function of observer time. Inside the caustic there are four bright images; outside it at most two are bright. In computing the light curves we tracked these bright images plus the two next brightest (those traveling once around the hole in either direction) to

estimate the total error due to neglected images, which was found to be less than 1 percent.

The images for the initial source position were found by searching on small grids, centered around the locations the images would have in a Schwarzschild geometry or around the relevant critical curve, as appropriate. (Since it is much easier to follow a known merging image pair for sources passing out of caustics than to find a new one when crossing into a caustic, the initial source position was normally taken to be *inside* the caustic, which was then followed forwards and backwards in time to create the total light curve.) Subsequent image positions were estimated by extrapolation of the previous positions and Newton-Raphson techniques were then used to converge on the new positions. The frequency shift $g \equiv \nu_{\rm obs}/\nu_{em}$ was computed from $g = (\mathbf{u}_{\rm obs} \cdot \mathbf{p}_{\rm obs})/(\mathbf{u}_{em} \cdot \mathbf{p}_{em})$, where $\mathbf{u}_i$ is the 4-velocity of the observer or emitter and $\mathbf{p}_i$ is the 4-momentum of the observed or emitted photon. The light curves produced assumed the emitter to have a power-law spectrum with index $\alpha = 0.5$ ($I_\nu \propto \nu^{-\alpha}$), typical of AGN X-ray spectra. The intensity at fixed observation frequency transforms as $g^{3+\alpha}$. In total the observed light curve is computed from

$$S_{\nu_{\rm obs}}(t_{\rm obs}) \propto \Big[\sum_{\text{images } i} |\mathcal{M}_i(t_{ret,i})| g_i^{3+\alpha}(t_{ret,i}) \Big] \, \nu_{\rm obs}^{-\alpha}, \tag{3.7}$$

where $t_{ret,i}(\vec{r}_{\rm obs})$ is the retarded time and $\mathcal{M}_i(t_{ret,i})$ is the associated magnification for image i.

The general features of the light curves for point sources that cross the caustic surface can be deduced from the preceding results. (We reinstate the specific angular momentum of the hole a and its mass M as physical lengths and use M_6 to denote the actual hole mass in units of 10^6 $M_\odot$.) From equation (3.6) we can write the width of the caustic surface at radius r as

$$s \sim 5 \times 10^{10} \left(\frac{a \sin\theta}{M} \right)^2 M_6 \left(\frac{r}{M} \right)^{-1} \quad \text{cm} \tag{3.8}$$

for $r \gg M$. A point source approaching the caustic surface well away from the cusp lines will produce two images on the Einstein ring, each with magnification

$\sim (Mr)^{1/2}/\zeta$ where $\zeta \gg s$ is the distance of the source from the center of the caustic. On ingress at the caustic, two extra images of the source will be created on the Einstein ring with a magnification that is initially infinite and subsequently declines in inverse proportion to the square root of the normal displacement of the source from the caustic, which we denote by $x = 10^6 x_6$ cm (cf. Schneider et al. 1992). This is characteristic of a fold catastrophe. When the source is close to the center of the caustic, there will be four images, spaced roughly equally around the Einstein ring with combined magnification given by equation (3.5). Closer to the caustic, the two brighter subimages will have a combined magnification of $\mathcal{M} \sim (\mathcal{M}_0/2)(2x/s)^{-1/2} \sim 1700(r/a\sin\theta)(x_6/M_6)^{-1/2}$. On egress, another pair of subimages will brighten, coalesce when the fold caustic is crossed and then vanish.

A point source will therefore be momentarily infinitely magnified after it crosses a caustic. The timescales involved are, however, extremely short. If we assume that the source moves with the virial velocity $v \sim 300,000(r/M)^{-1/2}$ km s^{-1}, then the elapsed interval until or since the closest caustic crossing is $\Delta t \sim 30 x_6 (r/M)^{1/2}$ μs, or in terms of the magnification,

$$\Delta t \sim 90 \left(\frac{r}{M}\right)^{5/2} \left(\frac{a\sin\theta}{M}\right)^{-2} \frac{M_6}{\mathcal{M}^2} \quad \text{s} \tag{3.9}$$

for $\Delta t \ll (a\sin\theta/M)^2 M_6 (r/M)^{-1/2}$ s.

All of these features are exhibited in the accurate light curves that have been calculated. We first consider a nearly circular orbit with radius $r = 50\,M$ in the equatorial plane of a maximal Kerr hole of mass 10^7 $M_\odot$ (Fig. 3.9). The observer is supposed to reside in the equatorial plane. Figs. 3.10*c*-*d* show the resulting light curve. A minimum magnification $\mathcal{M}_0 \sim 7500$ is produced within the caustic for ~ 2 s, consistent with the above estimates. (Eq. (3.9) actually underestimates the true Δt in this case because it assumes two dominant images, true close to a fold, but not true near the center of the caustic, where all four images are of comparable brightness.) The two sharp spikes corresponding to ingress and egress are clearly visible.

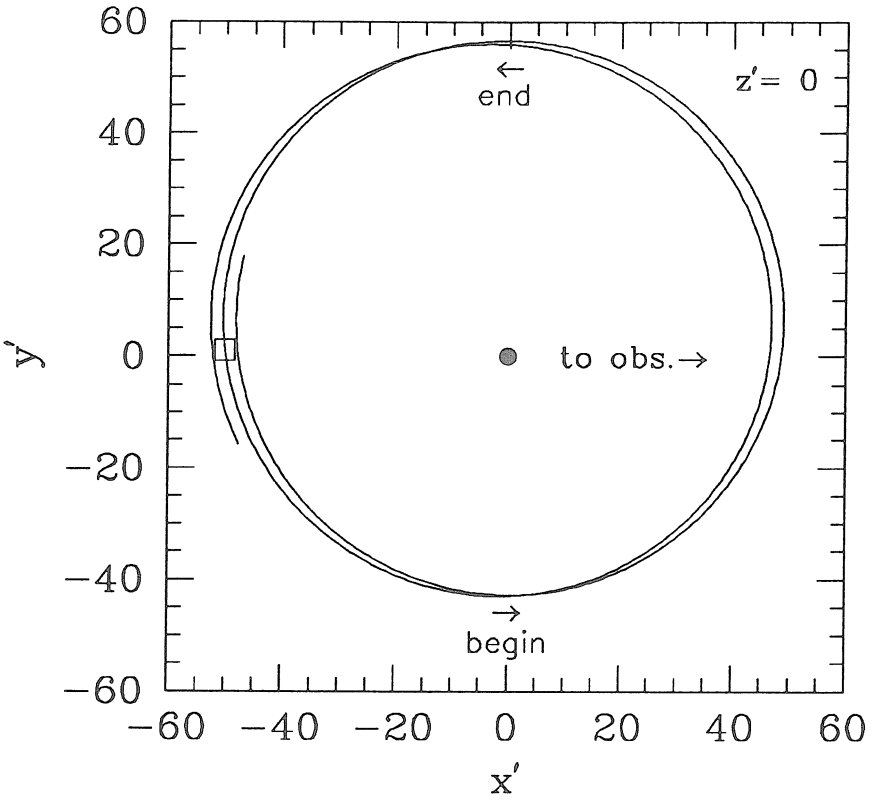

Fig. 3.9: The geodesic orbit of the point source whose light curves are plotted in Figs. 3.10*a*-*h*; it is a nearly circular orbit at a radius $r = 50$ from the hole. The observer lies in the plane of the figure at $x' = \infty$. Arrows labeled 'begin' and 'end' mark the orbital positions corresponding to the beginning and ending of the plotted light curves and show the direction of motion. The small square shows (in projection) where the orbit crosses the caustic surface.

When the source trajectory passes closer to the cusps, a slightly more complex light curve ensues. This is because one of the two original subimages will be highly magnified before the caustic is crossed (cf. Figs. 3.3*a*-*c*) and the two additional subimages are formed. This effect can be clearly seen in Figs. 3.10*a*-*b*, where the source path just misses the cusp on the caustic and both spikes develop secondary spikes. These secondary spikes are in evidence even when the path passes outside the caustic and one of the two subimages is highly magnified (Figs. 3.10*e*-*f*).

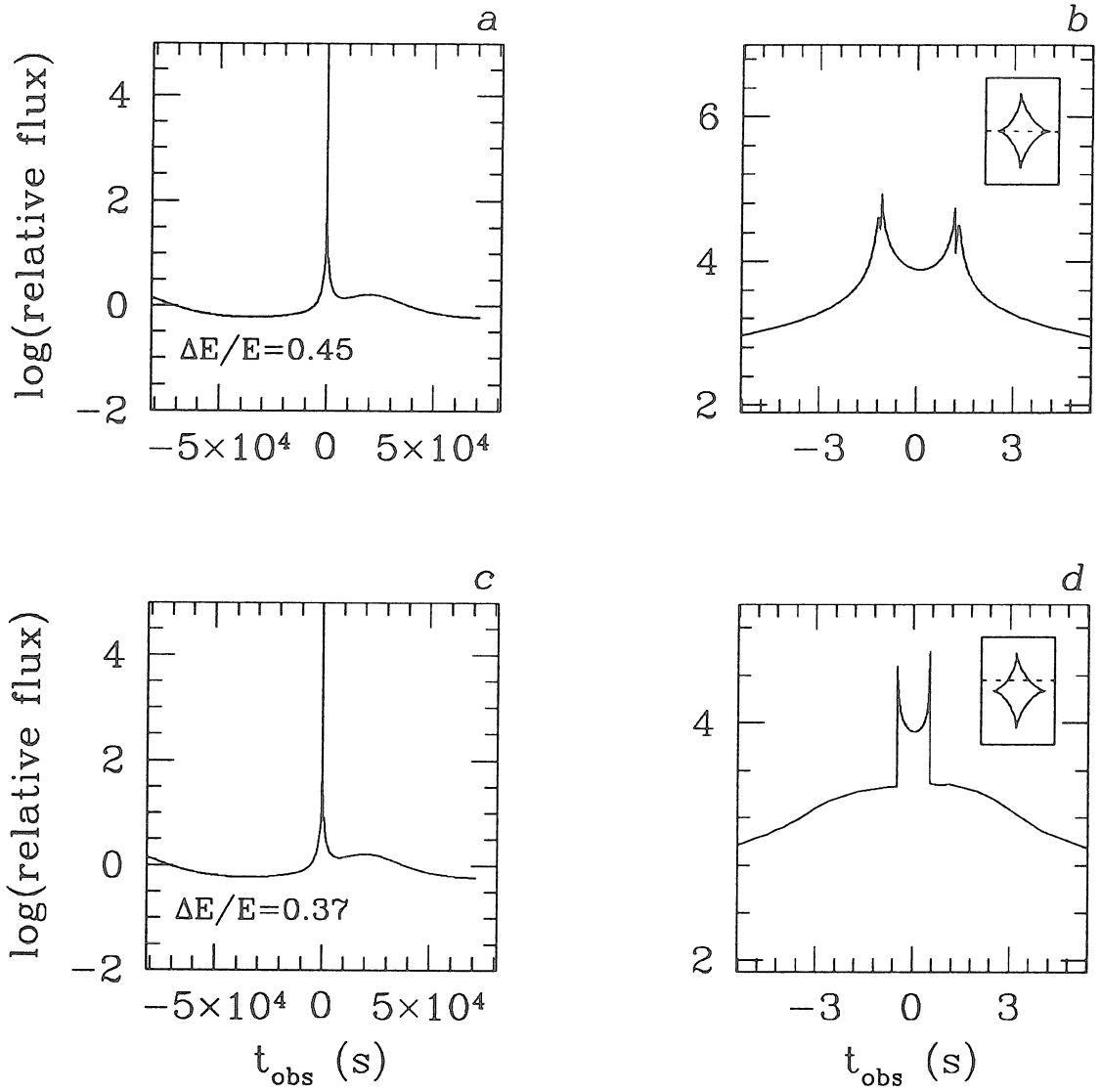

Fig. 3.10: Light curves for a point source moving on the orbit shown in Fig. 3.9. The observer time $t_{\rm obs}$ assumes a hole mass $M = 10^7$ $M_\odot$ and scales $\propto M$. Figs. 3.10*a*-*b* are for an orbit crossing nearly through the center of the caustic; Fig. 3.10*a* is the complete light curve and Fig. 3.10*b* is a detailed plot of the caustic crossing (the spike in Fig. 3.10*a*) showing the very rapid variability in the flux as the source crosses the caustic. The double-peaked structure in Fig. 3.10*b* is due to the source passing close to, but outside the cusp (smaller peaks) before crossing into and then out of the caustic itself (higher peaks); the path taken by the source through the caustic is shown in the inset box in Fig. 3.10*b*. Figs. 3.10*c*-*h* are as Figs. 3.10*a*-*b* only with the source's orbit tilted slightly so as to cross the caustic

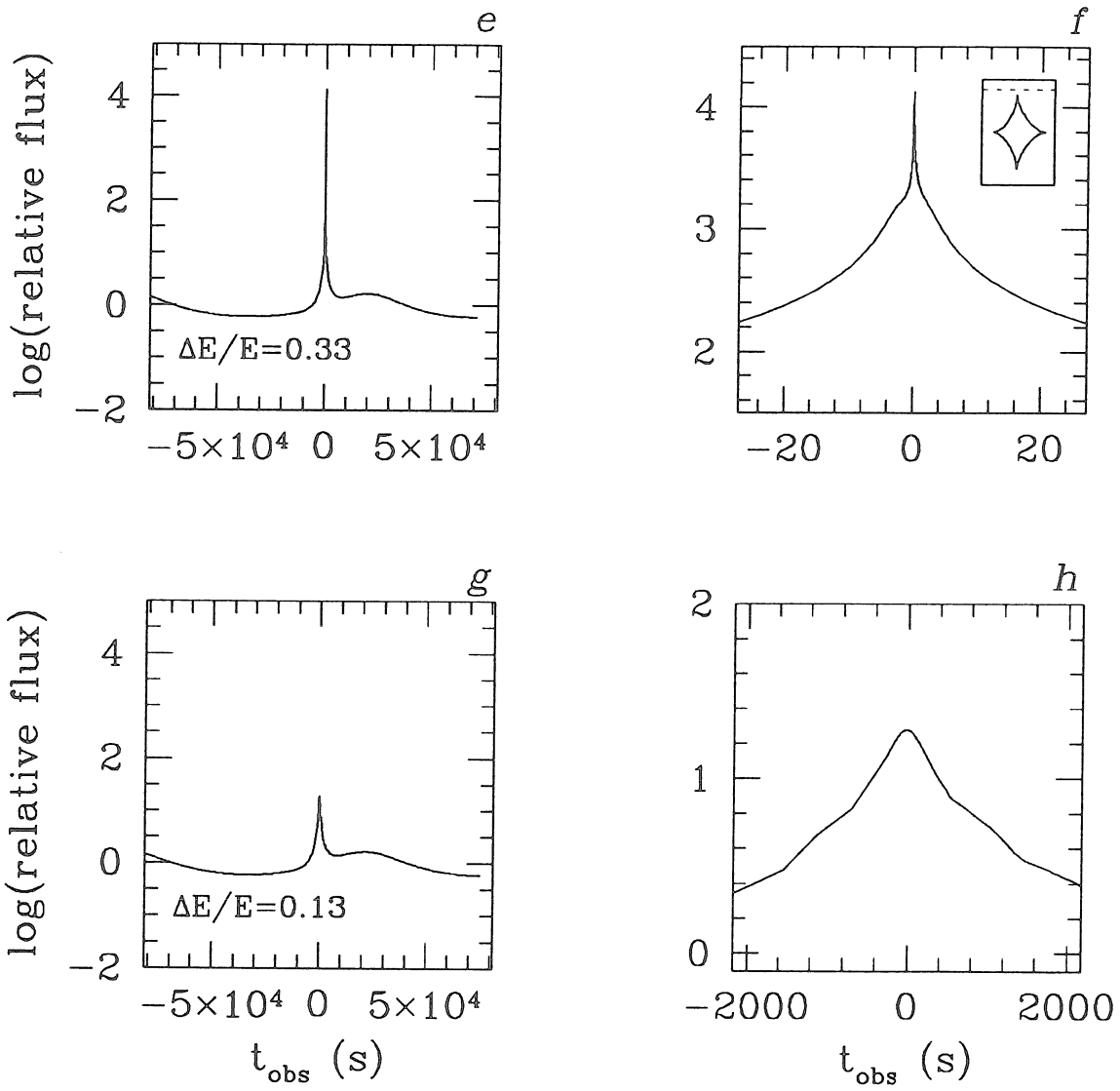

Fig. 10 (cont.): in different places (or outside it), shown schematically by their respective inset diagrams. Figs. 3.10*c*-*d* are for an orbit crossing just inside the caustic; Figs. 3.10*e*-*f* for an orbit passing just outside of it; and Figs. 3.10*g*-*h*, for an orbit passing 100 caustic widths outside of it. In all cases the lensing spike is the dominant feature in the light curves. The values of $\Delta E/E$ in the figures give the fraction of the total observed energy (over one orbit) received while the source suffered magnification $\mathcal{M} > 3$; for sources crossing the caustic, a large fraction of the received energy is the result of passage through the high magnification region.

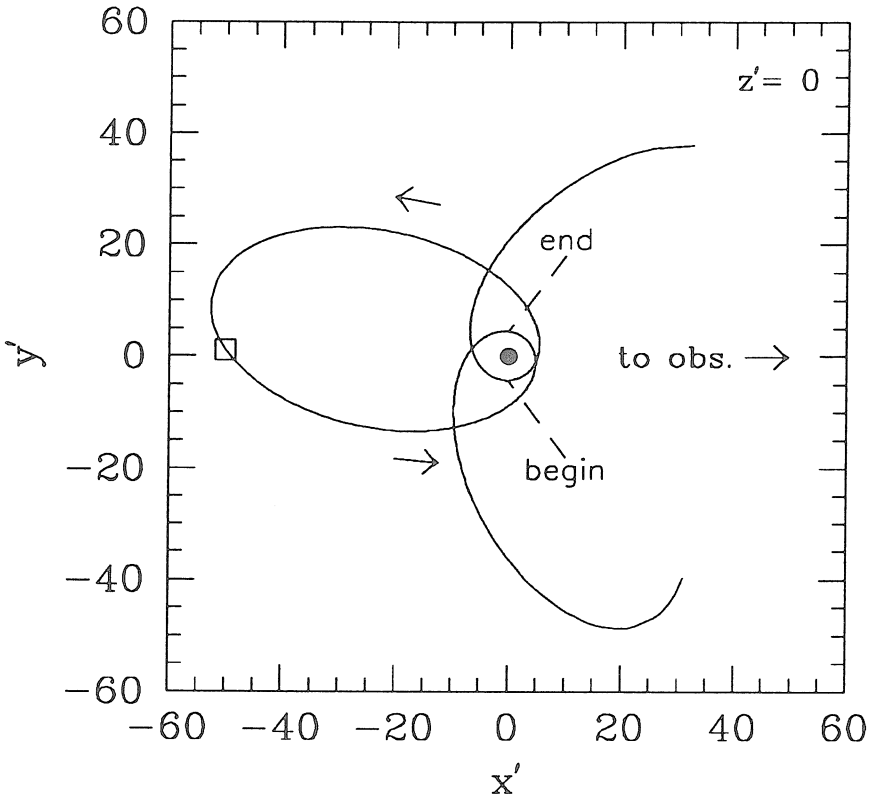

Fig. 3.11: As Fig. 3.9 except the point source orbit is a highly elliptical one passing close to the hole, with periapse at $r = 4$ and apoapse at $r = 60$.

When the distance of closest approach to the caustic greatly exceeds the caustic width s, the magnification contours are roughly circular and the magnification is given by equation (3.41). This is exhibited in Figs. 3.10g-h.

These figures are all quite symmetrical. Less symmetrical results can be seen when the source orbit is non-circular (Fig. 3.11). In Figs. 3.12a-h we display the corresponding light curves, analogous to Figs. 3.10a-h for the nearly circular source orbit. In this example, periapse and apoapse are $4M$ and $60M$ respectively. The magnification is dominated by the central lensing spike. The most extreme manifestation of the orbital asymmetry is the strong Doppler de-boosting that can occur when the source motion at periapse is directed away from the observer. Note that in Figs. 3.12g-h, where the source passes no closer than $\sim 100s$ to the caustic, the Doppler demagnification is as large as the magnification due to the

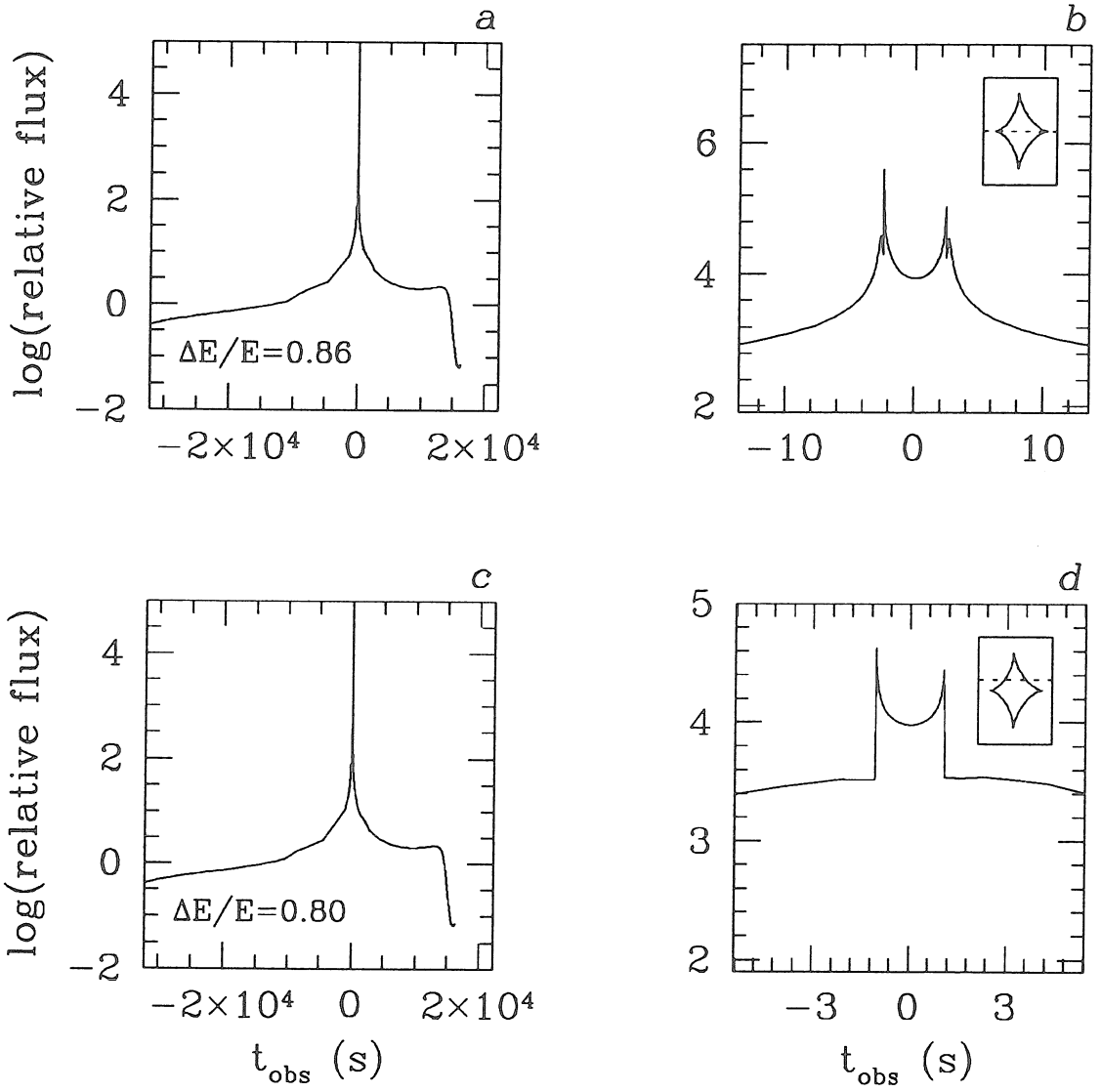

Fig. 3.12: Exactly analogous to Figs. 3.10*a-h* except the source follows the elliptical orbit of Fig. 3.11. In this case, the lensing spike is still the dominant feature except for Fig. 3.12*g* (where the orbit passes 100 caustic widths away from caustic), where it is comparable to the final dip in the light curve, which is caused by severe Doppler de-boosting as the source passes periapse while traveling almost directly away from the observer.

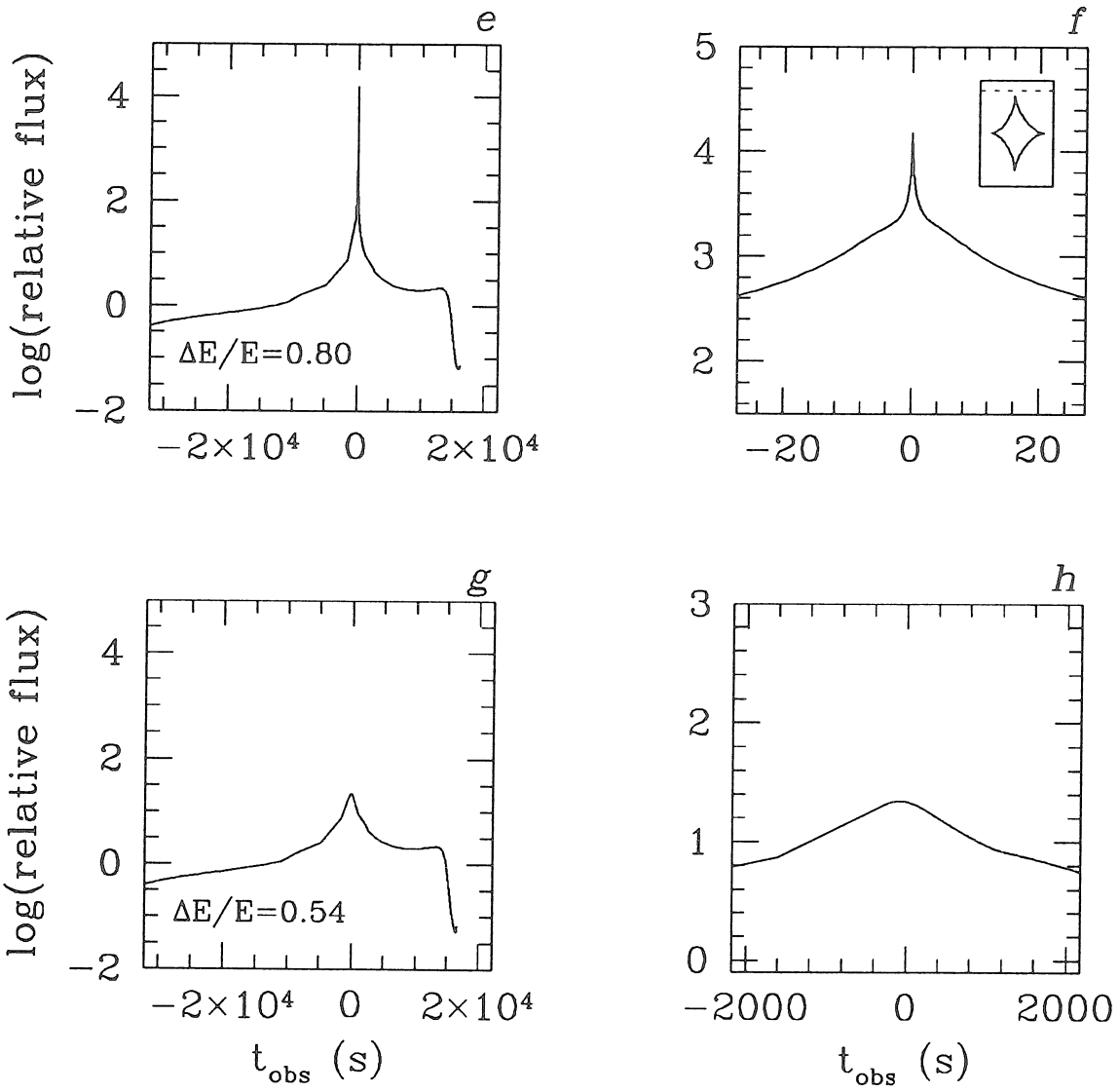

Fig. 3.12: *Continued*

gravitational lens action.

So far, we have confined our attention to the case of a point source. Real sources necessarily have finite size, $h = 10^6 h_6$ cm. When $h \lesssim s$, only those trajectories that pass close to or cross the caustic are significantly magnified. The high magnification spikes will be rounded off, relative to those of a point source, when the distance of the center of the source from the caustic $x \lesssim h$. The maximum magnification will therefore be

$$\mathcal{M}_{\max} \sim 1700(r/a\sin\theta)(h_6/M_6)^{-1/2}; \quad h \lesssim s. \tag{3.10}$$

If $h \gtrsim s$, then the caustic will be unresolved and the maximum magnification for a source crossing the caustic can be found from equation (3.41),

$$\mathcal{M}_{\max} \sim 3 \times 10^5 \left(\frac{M_6}{h_6}\right)\left(\frac{r}{M}\right)^{1/2}; \quad h \gtrsim s, \tag{3.11}$$

as long as $\mathcal{M}_{\max} \gg 1$.

3.5. OBSERVATIONS OF THICK ACCRETION DISKS

For a second application of the ray-tracing routine, we consider accretion disks in orbit around a massive black hole in an AGN. The appearance of thin accretion disks has already been considered by several authors including Luminet (1979), Sun & Malkan (1989), Laor, Netzer, & Piran (1990), Fukue (1988), Fukue & Yokoyama (1988), Rauch & Blandford (1991) and Jaroszyński, Wambsganss, & Paczyński (1992), who have treated the disk as an infinitely thin, black body emitter. Both integrated fluxes and the appearance of the disk have been computed for a variety of assumptions. The influence of gravitational lensing can be seen in an enhanced flux from the far side of the disk and that of the Doppler-gravitational blueshift in the creation of a hot-spot at the approaching ansa.

Real accretion disks are likely to be quite thick, particularly in AGN accreting at rates approaching the Eddington limit. The finite disk thickness is responsible for moderating both of these flux-enhancing effects because the front side of

the disk obscures the backside when the observer is located near the equatorial plane, although the details of the eclipse are hard to model (see Madau 1988 for a Newtonian treatment).

We consider a simple family of "slender" accretion disk models; cf. Biehle & Blandford (1993). These disks are supposed to be barotropic and to have specific angular momentum $\ell \equiv -\mathrm{u}_\phi/\mathrm{u}_t$ specified in the midplane through the relation $\ell(r,\pi/2) = (\ell_0 + \ell_1\, r)^{1/2}$, where ℓ_0 and ℓ_1 are constants. It is further supposed that the photosphere coincides with an isobar. The only non-zero components of the acceleration, $a_\beta = u^\alpha u_{\beta;\alpha}$, are a_r and a_θ and these can be expressed as the gradient of a scalar potential h,

$$h = \ln e - \int d\ell\, \Omega/(1-\Omega\ell), \quad \text{i.e.,} \quad \mathbf{a} = -\nabla P/w = \nabla h, \tag{3.12}$$

where w is the relativistic enthalpy density, $\Omega(\ell)$ is the angular velocity, and e is the specific energy, given by $e = (2g^{t\phi}\ell - g^{tt} - g^{\phi\phi}\ell^2)^{-1/2}$. We can use equation (3.12) to determine where a particular ray hits the disk surface.

In order to compute the observed radiation, we assume that the disk is radiation-dominated and that electron scattering dominates the opacity, so that the emergent flux $\mathbf{F}$ satisfies the local Eddington condition

$$\mathbf{F} = \left(\frac{c}{\kappa_T}\right) \mathbf{a}, \tag{3.13}$$

where κ_T is the Thomson opacity.

The radiation flux is evaluated in the rest frame of the fluid and then propagated to infinity. (We ignore, for simplicity, the influence of external irradiation.) We deal with the total flux, rather than the spectral flux, as the latter quantity is more model-dependent, cf. Rauch & Blandford 1991; however, qualitatively similar effects would be seen if we were to use the spectral flux as long as the observation frequency is high enough that the flux is dominated by the inner regions. At lower frequencies the effective disk size is larger and relativistic effects are less important. We have computed the observed flux, allowing for the Doppler shift from

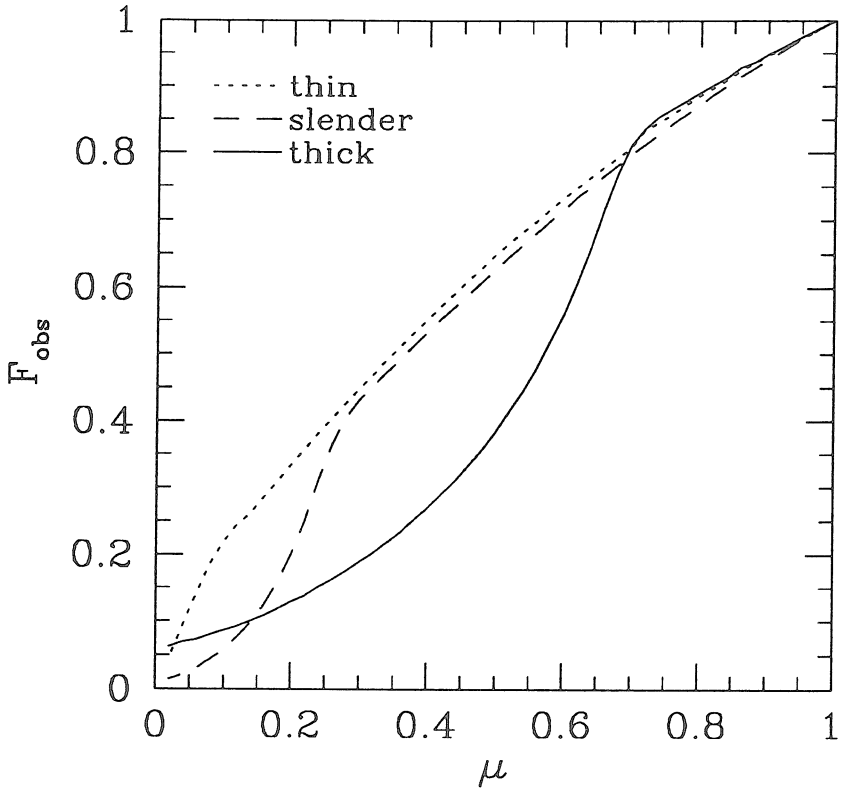

Fig. 3.13: Total (frequency-integrated) flux as a function of inclination angle, $\cos^{-1}\mu$, from three different accretion disk models around an extreme Kerr black hole. The thin disk has an asymptotic full width of $\sim 25M$, the slender disk of $\sim 120M$, and the thick disk reaches a maximum width of $\sim 230M$ at a radial distance (cylindrical radius) of $\sim 180M$. Fluxes have been normalized to agree when the disks are viewed face-on. Note that the slender and especially the thick disk are less luminous than the thin disk when viewed at large inclination angles due to obscuration of the bright, innermost regions. See §3.5.

the surface as in §3.4, as a function of μ_0 for three disk models: a thin disk, a slender disk, and a thick disk. The results are displayed in Fig. 3.13. Note that a modest flux enhancement is possible when a thin disk is observed from the equatorial plane, but that this enhancement is not present for a thick disk on account of obscuration. The high brightness hot-spot associated with those parts of the

inner disk moving towards us with relativistic speed can be seen in thin disks, but not in slender and thick disks. This is relevant to discussions of microlensing in Q2237+0305 (Jaroszyński et al. 1992; Rauch & Blandford 1991). The hot-spots associated with those parts of the disk intersecting the caustic are also obscured in these models.

3.6. DISCUSSION

In this paper, we have described a general routine to trace geodesics near a Kerr black hole and have applied it in two idealized situations. Many of the manifestations of focusing by a Kerr hole are quite distinctive in theory and an unambiguous observation of them would provide incontrovertible proof that spinning black holes really can be found in AGN (or, alternatively, binary X-ray sources).

How likely is this to happen in practice? There are four conditions that must be satisfied. First, the source must be sufficiently compact that it will fit inside the caustic at the radius where the source crosses it. Second, the magnification must be great enough that the flux from a single, highly magnified source dominate that from the remainder of the AGN. Third, the observations must be made with large enough temporal resolution that the caustic crossing can be detected. Finally, there must be enough discrete emitting elements satisfying these three requirements that caustic crossing happens with sufficient frequency to make monitoring individual AGN for it a practical proposition. Satisfying all of these conditions simultaneously is very difficult in practice, though not impossible.

3.6.1. *Magnification of Optical-UV Emission Line Clouds*

The nature of the gas flow close to an AGN black hole is still a matter of conjecture. In particular, it is not clear whether or not a disk (for which there is some evidence at large radius) is present all the way down to a few gravitational radii or if the flow becomes quasi-spherical. In addition, the radii at which the optical-UV continuum and perhaps the highest ionization emission lines are

formed is uncertain. One possibility (e.g., Ferland & Rees 1988, Celotti, Fabian & Rees 1992) is that radiation emitted non-thermally very close to the hole's event horizon is re-processed by a blanket of small gaseous clouds emitting close to their black body limits at effective temperatures $T \sim 3 \times 10^4$ K. If this were the case, then the absorption cross sections per ion, σ, would be so large for this gas that the effective Eddington limit would be reduced by factors ~ 1000 from the Thomson-Eddington limit. In order to confine such clouds, it would be necessary to invoke strong $\gtrsim 10^4$ G magnetic fields, so that the gas would have a filamentary form with characteristic thickness that might be as small as $h \sim kT/U_{\rm rad}\sigma \sim (U_{rad}/10^5\,{\rm erg\,cm^{-3}})^{-1}$ km.

The observed flux from an AGN should fluctuate as these filaments cross the caustic with speeds determined by the motion of the magnetic footprints. The maximum expected relative variation is given by

$$\frac{\delta F}{F} \sim \frac{(sh)^{1/2} h \mathcal{M}_{\rm max}(h)}{4\pi r^2 C} \sim \frac{1}{5}\left(\frac{h}{r}\right)\left(\frac{r}{M}\right)^{-1/2} C^{-1}; \quad h \lesssim s \tag{3.14}$$

where C is the covering factor and $\mathcal{M}_{\rm max}(h)$ is given by equation (3.10). The maximum fluctuation expected is therefore $\delta F/F \sim 0.1(a\sin\theta/M)^2(r/M)^{-5/2}C^{-1}$ which is always small unless $C \ll 1$, in which case the clouds will not reprocess much of the nonthermal radiation. Therefore, we do not expect to see rapid variations under such a model.

3.6.2. *Magnification of X-ray Sources*

Another possibility is that the central black hole in an AGN is surrounded by a cluster of stars containing X-ray luminous, accreting compact objects. It is a little easier in this case to imagine that one of these might create a transient flash when it moves behind a black hole as the background emission is either optically thin or has a small covering factor. There are many possibilities; however, we will only consider one specific example. Suppose that there are N individual black holes of

mass $\sim 10\ M_\odot$ at a characteristic distance r from the black hole accreting gas from their surroundings. (Some of them may also have binary companions, acquired through physical collisions.) Their intrinsic luminosities will be proportional to the mean gas density through which they move. If we denote the satellite holes with the subscript $*$, and assume that they emit X-rays with radiative efficiency per unit mass of accreted gas of $\epsilon_* \sim 0.1$, and that the central black hole emits X-rays with a radiative efficiency of $\epsilon_X \sim 0.003$, then their individual luminosities will be related to the mean luminosity L by

$$\frac{L_*}{L} \sim \left(\frac{\epsilon_*}{\epsilon_X}\right)\left(\frac{rv_r^2}{GM}\right)^{-1/2}\left(\frac{\Omega}{4\pi}\right)^{-1}\left(\frac{M_*}{M}\right)^2. \tag{3.15}$$

If we assume that the mean radial velocity of the gas v_r is $\sim 10^{-3}$ of the local Keplerian speed, and that the stars are concentrated in an equatorial band subtending a solid angle $\Omega \sim 1$ sr at the central black hole (and assumed also to contain the observer), then $L_* \sim 3 \times 10^{-5} M_6^{-2} L$, and an individual accreting star will outshine the remainder of the AGN when the magnification $\mathcal{M} \geq 3 \times 10^4 M_6^2$.

Now suppose that the emitting surface of a compact X-ray source has linear size $h \sim 30$ km. Using equation (3.10) and assuming that $a \sin\theta \sim M$, we find this requires that the star make a caustic crossing at a distance satisfying

$$30 M_6^{3/2} \leq \frac{r}{M} \leq 2 \times 10^4 M_6. \tag{3.16}$$

The mean interval between caustic crossings is given by

$$\begin{aligned} T \sim \frac{\Omega r^{5/2}}{Ns(GM)^{1/2}} &\sim 15\left(\frac{r}{M}\right)^{7/2}\frac{M_6}{N} \quad \text{s} \\ &\geq \frac{2 \times 10^6}{N} M_6^{25/4} \quad \text{s}. \end{aligned} \tag{3.17}$$

Therefore, a cluster of a few hundred $\sim 10\ M_\odot$ black holes in a low luminosity AGN could conceivably produce observable flashes of duration given by equation (3.9), $\Delta t \sim (r/100M)^{1/2}$ ms. Accreting neutron stars can produce more highly magnified sources though it will be even harder to resolve them temporally. Main sequence stars on tightly bound orbits which fill their Roche lobes

near periapse can suffer tidal stripping of gas, which could also modulate the observed continuum emission. The characteristic period of such modulations would be $P \sim \pi(G\rho_\star)^{-1/2} \sim 3\,(\rho_\star/\rho_\odot)^{-1/2}$ hr; however, because of Lense-Thirring precession and advance of periapse, the modulation would be only quasi-periodic.

3.6.3. *X-ray Blazars*

For our final example, we turn to X-ray emitting BL Lac objects and violently variable radio loud quasars, known collectively as Blazars. In these objects, it is widely supposed that the X-rays originate in outflowing jets directed towards the observer. These jets are believed to have relativistic outflow speeds so that we only see the approaching jet, amplified by relativistic beaming. However, if the speed is no more than mildly relativistic, inhomogeneities in the counter-jet might occasionally be subject to strong magnification by the intervening black hole and outshine the approaching jet. One way in which this might happen is if the jet encounters an obstacle.

In this case, the emitting region is likely to be too large to actually resolve the caustic. The maximum magnification is therefore given by equation (3.41) by $\mathcal{M} \sim b_E/\zeta$, where ζ is the distance of the source region from the caustic. For example, a source of size $\sim 10^{10}$ cm at a distance $r \sim 10^{16}$ cm from a $\sim 10^6$ $M_\odot$ black hole and at a distance $\zeta \sim 10^{11}$ cm from the caustic can be magnified by $\mathcal{M} \sim 10^3$, which may, in some sources, exceed the magnification of the approaching jet due to relativistic beaming. As a receding source is likely to be moving nearly parallel to the caustic, it may be magnified for quite a long time. Unfortunately, as with the MACHO project (e.g., Griest & Hu 1992), it will be far more difficult to recognize the variation as being due to gravitational lensing as it will not have a particularly distinctive form. (Compare, for example, Fig. 3.10*b* with Fig. 3.10*f*.) Nevertheless, it may be worth scrutinizing the existing observational database for any evidence of gravitational lensing events.

In conclusion, although the probability of any of these effects being observable is not high, the importance of a positive identification of any one of them is so

large that it is worth scrutinizing existing X-ray observations of AGN to search for high amplification events.

Acknowledgements

We thank the referee, Peter Schneider, for his helpful suggestions. Some of the metric calculations needed for this work were performed using the MathTensor symbolic tensor algebra package which runs under Mathematica. This research was supported by NSF grant AST89-17765 and by an NSF graduate student fellowship.

APPENDIX A

NUMERICAL IMPLEMENTATION OF THE GEODESIC EQUATIONS

A1. COORDINATE SYSTEM

We use units where $G = c = M = 1$. All calculations were done using a modified Boyer-Lindquist coordinate system with coordinates (t, u, μ, ϕ), related to the usual Boyer-Lindquist (t, r, θ, ϕ) by $u \equiv 1/r$ and $\mu \equiv \cos\theta$, with t and ϕ unchanged. This has the advantage of bringing infinity into finite spatial coordinates and of replacing all trigonometric operations with algebraic ones. In addition, this means that certain special observer positions can be exactly represented numerically, such as $\mu = 0$ instead of $\theta \approx 3.14159\ldots$ for the equatorial plane, so that the code can straightforwardly detect and treat these special cases. In terms of the modified coordinates, the non-zero covariant metric components are:

$$g_{tt} = -1 + \frac{2u}{\tilde{\rho}^2}; \quad g_{uu} = \frac{\tilde{\rho}^2}{u^4(1 - 2u + a^2u^2)}; \quad g_{\mu\mu} = \frac{\tilde{\rho}^2}{u^2(1-\mu^2)}$$

$$g_{\phi\phi} = \frac{(1-\mu^2)\left[u^{-2} + a^2\left((\tilde{\rho}^2 + \mu^2) + 2u(1-\mu^2)\right)\right]}{\tilde{\rho}^2}; \quad g_{t\phi} = g_{\phi t} = \frac{-2au(1-\mu^2)}{\tilde{\rho}^2},$$

where $\tilde{\rho}^2 \equiv 1 + a^2\mu^2u^2$.

A2. EQUATIONS OF MOTION

In standard Boyer-Lindquist coordinates the contravariant components of the momentum of timelike or null geodesics around a Kerr black hole are (e.g., MTW)

$$p^t \equiv \frac{dt}{d\lambda} = \rho^{-2}\left\{-a\left(aE\sin^2\theta - L_z\right) + \frac{(r^2+a^2)}{\Delta}P\right\}$$

$$p^\theta \equiv \frac{d\theta}{d\lambda} = \rho^{-2}\theta_{sgn}\left\{\mathcal{Q} - \cos^2\theta\left[a^2(m^2 - E^2) + L_z^2\csc^2\theta\right]\right\}^{\frac{1}{2}}$$

$$p^r \equiv \frac{dr}{d\lambda} = \rho^{-2}r_{sgn}\left\{P^2 - \Delta\left[m^2r^2 + (L_z - aE)^2 + \mathcal{Q}\right]\right\}^{\frac{1}{2}}$$

$$p^\phi \equiv \frac{d\phi}{d\lambda} = \rho^{-2}\left\{-aE + L_z\csc^2\theta + \frac{a}{\Delta}P\right\},$$

where $\rho^2 = r^2 + a^2\cos^2\theta$ ($\equiv u^{-2}\tilde{\rho}^2$ above), $P \equiv E(r^2+a^2) - L_z a$, and $\Delta \equiv r^2 - 2r + a^2$. The constants of motion are the angular momentum, L_z; the energy at infinity, E; the rest mass, m; and Carter's constant, $\mathcal{Q} = p_\theta^2 + \cos^2\theta\left[a^2(m^2 - E^2) + L_z^2\csc^2\theta\right]$ (Carter 1968); r_{sgn} and θ_{sgn} ($= \pm 1$) are two independent, arbitrary signs (although fixed for any particular geodesic). For computation it is convenient to use dimensionless quantities throughout. To this end, define $\ell \equiv -p_\phi/p_t = L_z/E$, $q^2 \equiv \mathcal{Q}/E^2$, $\gamma \equiv E/m$, and a new affine parameter $\lambda' \equiv E\lambda$ (note that it is possible for q^2 to be negative, but this causes no problems). After converting to the modified coordinates, the momentum components become

$$
\begin{aligned}
p^t &\equiv \frac{dt}{d\lambda'} = \rho^{-2}\left\{-a\left[a(1-\mu^2) - \ell\right] + \frac{(u^{-2}+a^2)}{\Delta}(u^{-2} + a^2 - a\ell)\right\} \\
p^\mu &\equiv \frac{d\mu}{d\lambda'} = \mu_{sgn}\rho^{-2}\sqrt{M} \\
p^u &\equiv \frac{du}{d\lambda'} = u_{sgn}\rho^{-2}\sqrt{U} \\
p^\phi &\equiv \frac{d\phi}{d\lambda'} = \rho^{-2}\left\{-a + \frac{\ell}{1-\mu^2} + \frac{a}{\Delta}(u^{-2} + a^2 - a\ell)\right\},
\end{aligned}
$$

where $U \equiv (1-\gamma^{-2}) + 2\gamma^{-2}u + \left[a^2(1-\gamma^{-2}) - q^2 - \ell^2\right]u^2 + 2\left[(a-\ell)^2 + q^2\right]u^3 - a^2q^2u^4$, $M \equiv q^2 + (\tilde{a}^2 - q^2 - \ell^2)\mu^2 - \tilde{a}^2\mu^4$, $\tilde{a}^2 = (1-\gamma^{-2})a^2$ (for photons, $\tilde{a}^2 = a^2$), $u_{sgn}, \mu_{sgn} = \pm 1$ are the arbitrary signs, and ρ and Δ are as above (with $r \to u^{-1}$). Note that this new normalization implies $p_t = -1$ and $p_\nu p^\nu = -\gamma^{-2}$ for all particles.

For observers far from the black hole, the constants of motion ℓ and q^2 are easily expressed in terms of impact parameters. Following Cunningham and Bardeen (1973), define α and β as the observed impact parameters of a ray measured perpendicular and parallel, respectively, to the spin axis of the black hole projected onto the sky. Then for an observer at polar coordinate $\mu_0 = \cos\theta_0$ from the spin axis of the hole, one finds $\ell = -\alpha(1-\mu_0^2)^{\frac{1}{2}}$ and $q^2 = \beta^2 + \mu_0^2(\alpha^2 - \tilde{a}^2)$. The inverse transformation is $\alpha = -\ell(1-\mu_0^2)^{-\frac{1}{2}}$, $\beta = +\left(q^2 - \mu_0^2\left[\ell^2/(1-\mu_0^2) - \tilde{a}^2\right]\right)^{\frac{1}{2}}$; for ingoing instead of outgoing rays, the signs of α and β should be reversed.

The roots of the quartic U determine the radial turning points of the particle trajectory. For massless particles such as photons ($\gamma^{-2} = 0$), U does not

admit quasi-periodic, elliptical-like orbits (except for the *circular* photon orbits; see Bardeen 1973), so that all null geodesics either "emerge" from the horizon at $u = u_+ \equiv 1/\left(1+(1-a^2)^{\frac{1}{2}}\right)$, reach a minimum value of u (a maximum radius), and then plunge back into the horizon (assuming a finite turning point, i.e., $u_{min} > 0$), or start from infinity, reach a maximum value of u (for $u_{max} < u_+$), and then return to infinity. For massive particles quasi-elliptical orbits are of course possible; for any particular U, the allowed region(s) is that for which U is non-negative.

The roots of the quadratic (in μ^2) M determine the polar turning points and allowed region (for the given constants of motion) in like fashion. For null particles, M always has two roots less than unity, and at least one non-negative root; call them $M_\pm$, with $M_- \leq M_+ \leq 1$. If $M_- < 0$, then the turning points of μ are symmetric about the equatorial plane and equal to $\pm\mu_+ \equiv \pm\sqrt{M_+}$; for $M_- \geq 0$, the particle is trapped between $M_- \leq \mu^2 \leq M_+$. For massive particles, it is additionally possible to have $0 \leq M_+ \leq 1 \leq M_-$ (these are the quasi-periodic orbits), where $-\mu_+ \leq \mu \leq \mu_+$ again holds.

A3. INTEGRATIONS AND REDUCTIONS TO ELLIPTIC INTEGRALS

In integrating the equations of motion, we found it most convenient to use u as the independent variable. Although this may seem unnatural from a theoretical point of view, the "obvious" choices such as the affine parameter λ' or the time t are not as useful computationally, since they themselves are expressed as integrals over du (and $d\mu$) and require prior knowledge of u and μ to compute. Since u is not in general single-valued along the geodesic, we defined a single-valued, monotonic (with respect to λ'), pseudo-radial variable $\tilde{u}$ as follows. Given an initial starting position u_0, and turning points $u_{min} \leq u_0 \leq u_{max}$, let $\tilde{u} = u$ until a turning point is reached. Just after u_{max}, where u starts decreasing again, let $\tilde{u} = u_{max} + (u_{max} - u)$, and just past the next turning point, u_{min}, $\tilde{u} = u_{max} + (u_{max} - u_{min}) + (u - u_{min})$, and so on; this defines $\tilde{u}$ for all $\tilde{u} \geq u_0$. Values $\tilde{u} < u_0$ are defined similarly: $\tilde{u} = u_{min} - (u - u_{min})$ after the first turning point (u_{min}) but before the second, $\tilde{u} = u_{min} - (u_{max} - u_{min}) - (u_{max} - u)$ after

the second but before the third, and so on. The mapping $u \to \tilde{u}$ can be written compactly as $\tilde{u}(u) = u_0 + \oint_{u_0}^{u} |du'|$; note that $|du/d\tilde{u}| = 1$ always. Replacing the affine parameterization with $\tilde{u}$ greatly simplifies the calculations.

Given $\tilde{u}$, the polar variable $\mu(\tilde{u})$ is then uniquely, implicitly defined by

$$\int_{\mu_0}^{\mu(\tilde{u})} \frac{d\mu'}{\sqrt{M(\mu')}} = u_{sgn}\mu_{sgn} \int_{u_0}^{u(\tilde{u})} \frac{du'}{\sqrt{U(u')}},$$

where it is understood that whenever a $\mu-$ or $u-$turning point is crossed, the corresponding integrand changes sign (making each integral a monotonic function of $\tilde{u}$). The amount of computing involved in the caustic calculations required a ray-tracing code as efficient as readily practical. For this reason the preceding equation was inverted analytically using standard elliptic functions and integrals. For reference, these reductions are listed in Tables 3.1 and 3.2, along with any special parameter values they correspond to.

Tables 3.1 and 3.2, inversion of $\int_{\mu_1}^{\mu} d\mu'/\sqrt{M} = I$ for $\mu(\mu_1, I)$, is the complete list of cases occurring for timelike and null geodesics. The tables assume that range reduction has already been done, i.e., that there is a solution μ between μ_{min} and μ_{max} for the given μ_1 and I (in particular, this requires $|I| \leq I_{max}$, with equality possible only for $\mu_1 = \mu_{min}$ or μ_{max}). As an example, we give the computation for the case $\tilde{a}^2 > 0$, $q^2 \neq 0$ ($M_- < 0$); for this case, $M(\mu) = \tilde{a}^2(\mu^2 + A^2)(B^2 - \mu^2)$, where $A^2 \equiv -M_- > 0$ and $B^2 \equiv M_+ \equiv \mu_+^2$. First rewrite the μ-integration as

$$\int_{\mu_1}^{\mu} \frac{dt}{\sqrt{M(t)}} = \int_{\mu_1}^{\mu_+} \frac{dt}{\sqrt{M(t)}} - \int_{\mu}^{\mu_+} \frac{dt}{\sqrt{M(t)}} \equiv I.$$

Substituting the form of M for this case and rearranging gives

$$(A^2 + B^2)^{\frac{1}{2}} \int_{\mu}^{\mu_+} \frac{dt}{\left[(t^2 + A^2)(B^2 - t^2)\right]^{\frac{1}{2}}} =$$
$$(A^2 + B^2)^{\frac{1}{2}} \int_{\mu_1}^{\mu_+} \frac{dt}{\left[(t^2 + A^2)(B^2 - t^2)\right]^{\frac{1}{2}}} - |\tilde{a}|(A^2 + B^2)^{\frac{1}{2}} I.$$

Our notation for the elliptic functions and integrals is that of Abramowitz and Stegun (1964), with the addition that we define $x = \tan\phi$ and write the elliptic

integrals $\mathrm{F}(\phi\,|\,m_1)$ as $\mathrm{F}(x\,|\,m_1)$, etc. Note that we are using the complementary parameter $m_1 = 1 - m$ instead of m, which has certain numerical advantages. From Abramowitz and Stegun (Chapter 17),

$$(c^2+d^2)^{\frac{1}{2}} \int_y^d \frac{dt}{\left[(t^2+c^2)(d^2-t^2)\right]^{\frac{1}{2}}} = \mathrm{F}(\phi\,|\,m_1) = \mathrm{cn}^{-1}\Big(\frac{y}{d}\,\Big|\,m_1\Big),$$

with $\tan^2\phi = (d/y)^2 - 1$ and $m_1 = c^2/(c^2+d^2)$. This gives

$$\mathrm{cn}^{-1}\Big(\frac{\mu}{\mu_+}\,\Big|\,m_1\Big) = \mathrm{F}\Big(\Big[\Big(\frac{\mu_+}{\mu_1}\Big)^2 - 1\Big]^{\frac{1}{2}}\,\Big|\,m_1\Big) - |\tilde{a}|(A^2+B^2)^{\frac{1}{2}} I,$$

from which the entry in the table immediately follows (except that for $\mu_1 < 0$ the sign s_1 becomes relevant).

Tables 3.3-3.6, reduction of $\int_{u_1}^{u} du'/\sqrt{U} = I(u)$ for $I(u)$ is a complete list of cases for the specialized problem of null geodesics (the program for computing timelike trajectories currently does this integral numerically), except that the possible case $a \neq 0$, $q^2 < 0$ (2 equal) is not included because it was not encountered in practice and is of no particular physical interest, since the merging roots are negative. The cases with equal roots listed in the tables all have positive merging roots; these correspond to the limiting circular photon orbits.

Besides greatly speeding up the computation of $\mu(\tilde{u})$ (by a factor of ≈ 40, it was found), using elliptic functions and integrals can deliver nearly machine accuracy over a very broad range of parameter values. In practice, the limiting factor is the accuracy to which the roots of the quartic U are computed; in particular, rays which travel many times around the hole before escaping back to infinity or being captured, for which U has nearly a multiple root, can be especially taxing numerically. In addition, the $u-$integrations for ϕ and t (see below), which have the form $\int_{u_1}^{u_2} h(u)du/\sqrt{U}$, suffer a logarithmic divergence (like $\int^{\sim c} du/(u-c)$) for nearly coincident roots and require special treatment to maintain high accuracy in those cases. The program (for null geodesics only, at present) implements two changes of variables for these cases, one for a conjugate pair of imaginary roots with small imaginary part (and with real part between the limits of integration u_1

and u_2), and one for nearly equal real roots (in this case the smaller root will be the radial turning point u_{max}). In the former case, U contains a factor $(u-c)^2+d^2$, $|d| \ll 1$, from the pair of roots $c \pm di$; to remove the divergence, the program uses the following identity:

$$\int_{u_1}^{u_2} \frac{h(u)du}{\left[(u-c)^2+d^2\right]^{\frac{1}{2}}} = \int_{y_1}^{y_2} h(c+d\sinh y)dy\,,$$

where $y_i = \sinh^{-1}\big((u_i-c)/d\big)$. In the latter case we have $U \propto (c-u)(d-u)$, with $d > c$ and $d-c \ll 1$. In this situation the identity used is

$$\int_{u_1}^{u_2} \frac{h(u)du}{\left[(c-u)(d-u)\right]^{\frac{1}{2}}} = \int_{y_1}^{y_2} h\big(c-\tfrac{1}{2}(d-c)(\cosh y-1)\big)dy\,,$$

with $y_i = -\cosh^{-1}\big(1+2(c-u_i)/(d-c)\big)$. These transformations are used when necessary (as determined by prior testing) to preserve accuracy.

The integral for $\phi(\tilde{u})$ can be written as (cf. Cunningham & Bardeen 1973)

$$\phi(\tilde{u}) = \phi_0 + \mu_{sgn}\int_{\mu_0}^{\mu(\tilde{u})} \frac{\ell\mu^2}{1-\mu^2}\frac{d\mu}{\sqrt{M}} + u_{sgn}\int_{u_0}^{\tilde{u}} \frac{2(a-\ell)u+\ell}{(u/u_+-1)(u/u_--1)}\frac{du}{\sqrt{U}}\,,$$

where $u_\pm = \left[1\pm(1-a^2)^{\frac{1}{2}}\right]^{-1}$. This slight rearranging of $\frac{d\phi}{d\lambda'}d\lambda'$ is especially useful for geodesics confined to the equatorial plane, where the first term vanishes. In principle, both integrals can be decomposed into a combination of elliptic integrals and elementary functions, but it is much easier (as well as computationally faster) to do them numerically. The program uses a Gauss-Kronrod integration scheme to remove the $(x-x_0)^{-\frac{1}{2}}$ type singularities that occur at the turning points of u and μ. The μ-integral suffers an additional singularity when $|\mu| \to 1$; this corresponds to trajectories which pass very close to the pole, where ϕ changes rapidly. To alleviate this problem, the change of variables

$$\int_{\mu_1}^{\mu_2} \frac{h(\mu)d\mu}{(1-\mu)\sqrt{\mu_+-\mu}} = 2\int_{y_1}^{y_2} h\big(\mu_+-(1-\mu_+)\tan^2[y(1-\mu_+)^{\frac{1}{2}}]\big)dy\,,$$

where $y_i = -(1-\mu_+)^{-\frac{1}{2}}\tan^{-1}\left[(\mu_+-\mu_i)/(1-\mu_+)\right]^{\frac{1}{2}}$, is performed by the program when needed, which was found to maintain accuracy until at least $1-\mu_+ \lesssim 10^{-8}$ (approximately one arcmin from the axis).

After similar rearranging, the coordinate-time integral becomes

$$t(\tilde{u}) = t_0 + \mu_{sgn}\int_{\mu_0}^{\mu(\tilde{u})} a^2\mu^2\frac{d\mu}{\sqrt{M}} + u_{sgn}\int_{u_0}^{\tilde{u}}\left[\frac{2a(a-\ell)u^3 + a^2u^2 + 1}{u^2(u/u_+ - 1)(u/u_- - 1)}\right]\frac{du}{\sqrt{U}}\,.$$

The μ-integration is readily written in terms of elliptic integrals of the second kind, $\mathrm{E}(x\,|\,m_1)$, and this was done in the program. For the case $\tilde{a}^2 > 0$, $q^2 \neq 0$ ($M_- < 0$) examined above, for example, the reduction proceeds as follows:

$$\mu_{sgn}\int_{\mu_0}^{\mu(\tilde{u})} a^2\mu^2\frac{d\mu}{\sqrt{M}} =$$

$$a^2\mu_{sgn}\left\{\int_{\mu_0}^{\mu}\frac{(t^2+A^2)dt}{|\tilde{a}|\left[(t^2+A^2)(B^2-t^2)\right]^{\frac{1}{2}}} - A^2\int_{\mu_0}^{\mu}\frac{d\mu'}{\sqrt{M(\mu')}}\right\} =$$

$$a^2\mu_{sgn}\left\{\left[(A^2+B^2)^{\frac{1}{2}}\,\mathrm{E}\big(x(\mu')\,|\,m_1\big)\right]\Big|_{\mu_0}^{\mu} - A^2\mu_{sgn}u_{sgn}\int_{u_0}^{u}\frac{du'}{\sqrt{U(u')}}\right\},$$

where as in the previous discussion $x^2(\mu') = \tan^2\phi = (\mu_+/\mu')^2 - 1$, $A^2 = -M_-$, $B^2 = M_+ = \mu_+^2$, and $m_1 = -M_-/(M_+ - M_-)$. The U integral is already available from the calculation of $\mu(\tilde{u})$ and is not recomputed. (The above derivation ignores the range reduction and sign issues mentioned in the discussion of Tables 3.1-3.2). There is no divergence as $|\mu| \to 1$ here. The u-integration was done numerically, in the same manner as for $\phi(\tilde{u})$.

TABLE 3.1

Inversion of $\int_{\mu_1}^{\mu} d\mu'/\sqrt{M} = I$

Case	$M_+ \equiv \mu_+^2$	$M_- \equiv \mu_-^2$	$\mu(\mu_1, I)$[a]
$\tilde{a}^2 = 0$	$[0,1]$	$\cdots$	$\mu_+ \sin x$
$\tilde{a}^2 \neq 0,\ q^2 = 0$	$[0,1]$	0	$s_1\mu_+ \operatorname{sech} x$
$\tilde{a}^2 > 0,\ q^2 \neq 0\ (M_- < 0)$	$[0,1]$	< 0	$s_1\mu_+ \operatorname{cn}(x \mid m_1)$
$\tilde{a}^2 > 0,\ q^2 \neq 0\ (M_- > 0)$	$(M_-, 1]$	> 0	$s_1\mu_+ \operatorname{dn}(x \mid m_1)$
$\tilde{a}^2 < 0$	$[0,1]$	> 1	$s_1\mu_+ \operatorname{cd}(x \mid m_1)$

[a] In Tables 3.1 and 3.2, $s_1 = \operatorname{sign}(\mu_1)$ (if $\mu_1 = 0$, $s1 = \operatorname{sign}(I)$).

TABLE 3.1 (*continued*)

Case	I_{max}[b]
$\tilde{a}^2 = 0$	$\pi(\ell^2 + q^2)^{-\frac{1}{2}}$
$\tilde{a}^2 \neq 0,\ q^2 = 0$	∞
$\tilde{a}^2 > 0,\ q^2 \neq 0\ (M_- < 0)$	$2K(m_1)/\left[\lvert\tilde{a}\rvert(M_+ - M_-)^{\frac{1}{2}}\right]$
$\tilde{a}^2 > 0,\ q^2 \neq 0\ (M_- > 0)$	$K(m_1)/(\lvert\tilde{a}\rvert\mu_+)$
$\tilde{a}^2 < 0$	$2\,\mathrm{K}(m_1)/(-\tilde{a}^2 M_-)^{\frac{1}{2}}$

[b] $I_{max} \equiv \left| \int_{\mu_{min}}^{\mu_{max}} d\mu'/\sqrt{M} \right|$.

TABLE 3.2

Definitions for Table 3.1

Case	$x(\mu_1, I)$
$\tilde{a}^2 = 0$	$\sin^{-1}\left(\frac{\mu_1}{\mu_+}\right) + (q^2+\ell^2)^{\frac{1}{2}} I$
$\tilde{a}^2 \neq 0,\ q^2 = 0$	$\operatorname{sech}^{-1}\left\lvert\frac{\mu_1}{\mu_+}\right\rvert - s_1\lvert\tilde{a}\rvert\mu_+ I$
$\tilde{a}^2 > 0,\ q^2 \neq 0\ (M_- < 0)$	$\mathrm{F}\left(\left[(\frac{\mu_+}{\mu_1})^2 - 1\right]^{\frac{1}{2}} \middle\vert m_1\right) - s_1\lvert\tilde{a}\rvert(M_+ - M_-)^{\frac{1}{2}} I$
$\tilde{a}^2 > 0,\ q^2 \neq 0\ (M_- > 0)$	$\mathrm{F}\left(\left[\frac{M_+ - \mu_1^2}{\mu_1^2 - M_-}\right]^{\frac{1}{2}} \middle\vert m_1\right) - s_1\lvert\tilde{a}\rvert\mu_+ I$
$\tilde{a}^2 < 0$	$\mathrm{F}\left(\left[\frac{(M_+ - \mu_1^2)M_-}{(M_- - M_+)\mu_1^2}\right]^{\frac{1}{2}} \middle\vert m_1\right) - s_1(-\tilde{a}^2)^{\frac{1}{2}}\mu_- I$

TABLE 3.2 (*continued*)

Case	m_1	μ^{a}_{min}	μ^{a}_{max}
$\tilde{a}^2 = 0$	$\cdots$	$-\mu_+$	μ_+
$\tilde{a}^2 \neq 0,\ q^2 = 0$	$\cdots$	0	$s_1\mu_+$
$\tilde{a}^2 > 0,\ q^2 \neq 0\ (M_- < 0)$	$\frac{-M_-}{M_+ - M_-}$	$-\mu_+$	μ_+
$\tilde{a}^2 > 0,\ q^2 \neq 0\ (M_- > 0)$	$\frac{M_-}{M_+}$	$s_1\mu_-$	$s_1\mu_+$
$\tilde{a}^2 < 0$	$1 - \frac{M_+}{M_-}$	$-\mu_+$	μ_+

[a] μ must lie between μ_{min} and μ_{max}.

TABLE 3.3

Reduction of $\int_{u_1}^{u} du'/\sqrt{U} = \mathrm{I}(x, m_1)$[a]

Case	β_1[b]	β_2	β_3
$a = 0,\ \ell^2 + q^2 > 27$	$(\frac{1}{3}, \frac{1}{2})$	$(0, \frac{1}{3})$	< 0
$a = 0,\ \ell^2 + q^2 > 27$	$(\frac{1}{3}, \frac{1}{2})$	$(0, \frac{1}{3})$	< 0
$a = 0,\ \ell^2 + q^2 = 27$	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{6}$
$a = 0,\ \ell^2 + q^2 = 27$	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{6}$
$a = 0,\ \ell^2 + q^2 < 27$	complex	$\bar{\beta}_1$	< 0
$a \neq 0,\ q^2 = 0,\ \ell = a$	$\cdots$	$\cdots$	$\cdots$
$a \neq 0,\ q^2 = 0,\ \ell = -a$	complex	$\bar{\beta}_1$	$-\frac{1}{2}a^{-\frac{2}{3}}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (3 real)	$> \beta_2$	> 0	< 0
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (3 real)	$> \beta_2$	> 0	< 0
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (2 equal)	$= \beta_2$	$[\frac{1}{4}, 1]$	< 0
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (2 equal)	$= \beta_2$	$[\frac{1}{4}, 1]$	< 0
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (1 real)	complex	$\bar{\beta}_1$	< 0
$a \neq 0,\ q^2 < 0$ (2 real)	complex	$\bar{\beta}_1$	< 0
$a \neq 0,\ q^2 < 0$ (0 real)	complex[c]	$\bar{\beta}_1$	complex[c]
$a \neq 0,\ q^2 > 0$ (4 real, distinct)	$> \beta_2$	$> \beta_3$	> 0
$a \neq 0,\ q^2 > 0$ (4 real, distinct)	$> \beta_2$	$> \beta_3$	> 0
$a \neq 0,\ q^2 > 0$ (4 real, 2 equal)	$> \beta_2$	$= \beta_3$	$(\frac{1}{4}, 1)$
$a \neq 0,\ q^2 > 0$ (4 real, 2 equal)	$> \beta_2$	$= \beta_3$	$(\frac{1}{4}, 1)$
$a \neq 0,\ q^2 > 0$ (2 real)	> 0	complex	$\bar{\beta}_2$

[a] Listed are the cases for photons (constant $\gamma^{-2} = 0$); extension to timelike geodesics is straightforward in principal.

[b] The β_i are the roots of U.

[c] Roots ordered so that $\Re(\beta_1) \geq \Re(\beta_3)$ and $\Im(\beta_1) > 0$.

TABLE 3.3 (*continued*)

Case	β_4	u_1	$u \in [,]$[d]
$a = 0,\ \ell^2 + q^2 > 27$	$\cdots$	β_3	$[0, \beta_2]$
$a = 0,\ \ell^2 + q^2 > 27$	$\cdots$	∞	$[\beta_1, u_+)$
$a = 0,\ \ell^2 + q^2 = 27$	$\cdots$	β_3	$[0, \beta_2)$
$a = 0,\ \ell^2 + q^2 = 27$	$\cdots$	∞	(β_1, u_+)
$a = 0,\ \ell^2 + q^2 < 27$	$\cdots$	β_3	$[0, u_+)$
$a \neq 0,\ q^2 = 0,\ \ell = a$	$\cdots$	0	$[0, u_+)$
$a \neq 0,\ q^2 = 0,\ \ell = -a$	$\cdots$	β_3	$[0, u_+)$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (3 real)	$\cdots$	β_3	$[0, \beta_2]$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (3 real)	$\cdots$	∞	$[\beta_1, u_+)$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (2 equal)	$\cdots$	β_3	$[0, \beta_2]$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (2 equal)	$\cdots$	∞	(β_1, u_+)
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (1 real)	$\cdots$	β_3	$[0, u_+)$
$a \neq 0,\ q^2 < 0$ (2 real)	$< \beta_3$	β_3	$[0, u_+)$
$a \neq 0,\ q^2 < 0$ (0 real)	$\bar{\beta}_3$	c_3	$[0, u_+)$
$a \neq 0,\ q^2 > 0$ (4 real, distinct)	< 0	β_4	$[0, \beta_3]$
$a \neq 0,\ q^2 > 0$ (4 real, distinct)	< 0	β_1	$[\beta_2, \beta_1)$
$a \neq 0,\ q^2 > 0$ (4 real, 2 equal)	< 0	β_4	$[0, \beta_3)$
$a \neq 0,\ q^2 > 0$ (4 real, 2 equal)	< 0	β_1	(β_2, β_1)
$a \neq 0,\ q^2 > 0$ (2 real)	< 0	β_4	$[0, \beta_1]$

[d] Range of the radial coordinate is $u \in [u_{min}, u_{max}]$; if $u_{max} > u_+$ occurs $u_{max} \to u_+$ is implied. If u_{max} lies outside the horizon ($u_{max} < u_+$) the trajectory is hyperbolic-like, otherwise the ray is captured.

TABLE 3.4

Definitions for Table 3.3

Case	$\mathrm{I}(x, m_1)$	m_1
$a = 0,\ \ell^2 + q^2 > 27\ (u < \beta_2)$	$c_1 \mathrm{F}(x \mid m_1)$	$\frac{(\beta_1-\beta_2)}{(\beta_1-\beta_3)}$
$a = 0,\ \ell^2 + q^2 > 27\ (u > \beta_1)$	$c_1 \mathrm{F}(x \mid m_1)$	$\frac{(\beta_1-\beta_2)}{(\beta_1-\beta_3)}$
$a = 0,\ \ell^2 + q^2 = 27\ (u < \beta_2)$	$\frac{2}{\sqrt{27}} \tanh^{-1} x$	$\cdots$
$a = 0,\ \ell^2 + q^2 = 27\ (u > \beta_1)$	$-\frac{2}{\sqrt{27}} \tanh^{-1} x$	$\cdots$
$a = 0,\ \ell^2 + q^2 < 27$	$c_1 \mathrm{F}(x \mid m_1)$[a]	$\frac{1}{2} + \frac{(6\beta_3-1)}{8c_2}$
$a \neq 0,\ q^2 = 0,\ \ell = a$	x	$\cdots$
$a \neq 0,\ q^2 = 0,\ \ell = -a$	$c_1 \mathrm{F}(x \mid m_1)$[a]	$\frac{2-\sqrt{3}}{4}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a\ (3 \text{ real}, u < \beta_2)$	$c_1 \mathrm{F}(x \mid m_1)$	$\frac{(\beta_1-\beta_2)}{(\beta_1-\beta_3)}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a\ (3 \text{ real}, u > \beta_1)$	$c_1 \mathrm{F}(x \mid m_1)$	$\frac{(\beta_1-\beta_2)}{(\beta_1-\beta_3)}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a\ (2 \text{ equal}, u < \beta_2)$	$c_1 \tanh^{-1} x$	$\cdots$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a\ (2 \text{ equal}, u > \beta_1)$	$c_1 \tanh^{-1} x$	$\cdots$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a\ (1 \text{ real})$	$c_1 \mathrm{F}(x \mid m_1)$[a]	$\frac{1}{2} + \frac{6\beta_3+c_3}{8c_2}$
$a \neq 0,\ q^2 < 0\ (2 \text{ real})$	$c_1 \mathrm{F}(x \mid m_1)$[a]	$\frac{(\beta_3-\beta_4)^2-(c_4-c_5)^2}{4c_4c_5}$
$a \neq 0,\ q^2 < 0\ (0 \text{ real})$	$c_1 \mathrm{F}(x \mid m_1)$	$\left(\frac{c_4-c_5}{c_4+c_5}\right)^2$
$a \neq 0,\ q^2 > 0\ (4 \text{ real, distinct}, u < \beta_3)$	$c_1 \mathrm{F}(x \mid m_1)$	$\frac{(\beta_1-\beta_4)(\beta_2-\beta_3)}{(\beta_1-\beta_3)(\beta_2-\beta_4)}$
$a \neq 0,\ q^2 > 0\ (4 \text{ real, distinct}, u > \beta_2)$	$c_1 \mathrm{F}(x \mid m_1)$	$\frac{(\beta_1-\beta_4)(\beta_2-\beta_3)}{(\beta_1-\beta_3)(\beta_2-\beta_4)}$
$a \neq 0,\ q^2 > 0\ (4 \text{ real, 2 equal}, u < \beta_3)$	$c_1 \ln \lvert x \rvert$	$\cdots$
$a \neq 0,\ q^2 > 0\ (4 \text{ real, 2 equal}, u > \beta_2)$	$c_1 \ln \lvert x \rvert$	$\cdots$
$a \neq 0,\ q^2 > 0\ (2 \text{ real})$	$c_1 \mathrm{F}(x \mid m_1)$[a]	$\frac{(c_4+c_5)^2-(\beta_1-\beta_4)^2}{4c_4c_5}$

[a] If the absolute-valued quantity in the corresponding definition for $x(u)$ is negative, $\mathrm{F}(x \mid m_1)$ should be replaced by $2\,\mathrm{K}(m_1) - \mathrm{F}(x \mid m_1)$, where $\mathrm{K}(m_1)$ is the complete elliptic integral of the first kind.

TABLE 3.4 (*continued*)

Case	$x(u)$
$a = 0,\ \ell^2 + q^2 > 27\ (u < \beta_2)$	$\left[\frac{u-\beta_3}{\beta_2-u}\right]^{\frac{1}{2}}$
$a = 0,\ \ell^2 + q^2 > 27\ (u > \beta_1)$	$\left[\frac{\beta_1-\beta_3}{u-\beta_1}\right]^{\frac{1}{2}}$
$a = 0,\ \ell^2 + q^2 = 27\ (u < \beta_2)$	$(2u + 1/3)^{\frac{1}{2}}$
$a = 0,\ \ell^2 + q^2 = 27\ (u > \beta_1)$	$(2u + 1/3)^{-\frac{1}{2}}$
$a = 0,\ \ell^2 + q^2 < 27$	$\frac{2[c_2(u-\beta_3)]^{\frac{1}{2}}}{\lvert c_2-(u-\beta_3)\rvert}$
$a \neq 0,\ q^2 = 0,\ \ell = a$	u
$a \neq 0,\ q^2 = 0,\ \ell = -a$	$\frac{3^{\frac{1}{4}}2(1+c_2u)^{\frac{1}{2}}}{\lvert\sqrt{3}-(1+c_2u)\rvert}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (3 real, $u < \beta_2$)	$\left[\frac{u-\beta_3}{\beta_2-u}\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (3 real, $u > \beta_1$)	$\left[\frac{\beta_1-\beta_3}{u-\beta_1}\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (2 equal, $u < \beta_2$)	$\left[\frac{u-\beta_3}{\beta_2-\beta_3}\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (2 equal, $u > \beta_1$)	$\left[\frac{\beta_2-\beta_3}{u-\beta_3}\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (1 real)	$\frac{2[c_2(u-\beta_3)]^{\frac{1}{2}}}{\lvert c_2-(u-\beta_3)\rvert}$
$a \neq 0,\ q^2 < 0$ (2 real)	$\frac{2c_2(u)}{\lvert 1-c_2^2(u)\rvert}$
$a \neq 0,\ q^2 < 0$ (0 real)	$\frac{u-c_3}{\Im(\beta_1)(1+c_2^2)+c_2(u-c_3)}$
$a \neq 0,\ q^2 > 0$ (4 real, distinct, $u < \beta_3$)	$\left[\frac{(\beta_1-\beta_3)(u-\beta_4)}{(\beta_1-\beta_4)(\beta_3-u)}\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (4 real, distinct, $u > \beta_2$)	$\left[\frac{(\beta_2-\beta_4)(\beta_1-u)}{(\beta_1-\beta_4)(u-\beta_2)}\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (4 real, 2 equal, $u < \beta_3$)	$c_2 + \frac{2\left(c_3+[c_3(\beta_1-u)(u-\beta_4)]^{\frac{1}{2}}\right)}{(\beta_1-\beta_4)(\beta_3-u)}$
$a \neq 0,\ q^2 > 0$ (4 real, 2 equal, $u > \beta_2$)	$c_2 + \frac{2\left(c_3+[c_3(\beta_1-u)(u-\beta_4)]^{\frac{1}{2}}\right)}{(\beta_1-\beta_4)(\beta_3-u)}$
$a \neq 0,\ q^2 > 0$ (2 real)	$\frac{2c_2(u)}{\lvert 1-c_2^2(u)\rvert}$

TABLE 3.5

Auxiliary definitions for Tables 3.3 and 3.4

Case	c_1
$a = 0,\ \ell^2 + q^2 > 27\ (u < \beta_2)$	$\left[\frac{2}{(\beta_1-\beta_3)(\ell^2+q^2)}\right]^{\frac{1}{2}}$
$a = 0,\ \ell^2 + q^2 > 27\ (u > \beta_1)$	$-\left[\frac{2}{(\beta_1-\beta_3)(\ell^2+q^2)}\right]^{\frac{1}{2}}$
$a = 0,\ \ell^2 + q^2 < 27$	$\left[2c_2(\ell^2+q^2)\right]^{-\frac{1}{2}}$
$a \neq 0,\ q^2 = 0,\ \ell = -a$	$3^{-\frac{1}{4}}c_2^{-1}$
$a \neq 0,\ q^2 = 0$ (3 real, $u < \beta_2$)	$\left[\frac{2}{\beta_1-\beta_3}\right]^{\frac{1}{2}}\lvert a-\ell\rvert^{-1}$
$a \neq 0,\ q^2 = 0$ (3 real, $u > \beta_1$)	$-\left[\frac{2}{\beta_1-\beta_3}\right]^{\frac{1}{2}}\lvert a-\ell\rvert^{-1}$
$a \neq 0,\ q^2 = 0$ (2 equal, $u < \beta_2$)	$\left[\frac{2}{\beta_2-\beta_3}\right]^{\frac{1}{2}}\lvert a-\ell\rvert^{-1}$
$a \neq 0,\ q^2 = 0$ (2 equal, $u > \beta_1$)	$-\left[\frac{2}{\beta_2-\beta_3}\right]^{\frac{1}{2}}\lvert a-\ell\rvert^{-1}$
$a \neq 0,\ q^2 = 0,\ \lvert\ell\rvert \neq a$ (1 real)	$(2c_2)^{-\frac{1}{2}}\lvert a-\ell\rvert^{-1}$
$a \neq 0,\ q^2 < 0$ (2 real)	$(-a^2q^2c_4c_5)^{-\frac{1}{2}}$
$a \neq 0,\ q^2 < 0$ (0 real)	$2(c_4+c_5)^{-1}(-a^2q^2)^{-\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (4 real, $u < \beta_3$)	$2\left[a^2q^2(\beta_1-\beta_3)(\beta_2-\beta_4)\right]^{-\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (4 real, $u > \beta_2$)	$-2\left[a^2q^2(\beta_1-\beta_3)(\beta_2-\beta_4)\right]^{-\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (2 equal, $u < \beta_3$)	$(a^2q^2c_3)^{-\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (2 equal, $u > \beta_2$)	$(a^2q^2c_3)^{-\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (2 real)	$(a^2q^2c_4c_5)^{-\frac{1}{2}}$

TABLE 3.5 (*continued*)

Case	c_2	c_3
$a=0,\ \ell^2+q^2>27\ (u<\beta_2)$	$\cdots$	$\cdots$
$a=0,\ \ell^2+q^2>27\ (u>\beta_1)$	$\cdots$	$\cdots$
$a=0,\ \ell^2+q^2<27$	$[\beta_3(3\beta_3-1)]^{\frac{1}{2}}$	$\cdots$
$a\neq 0,\ q^2=0,\ \ell=-a$	$2a^{\frac{2}{3}}$	$\cdots$
$a\neq 0,\ q^2=0\ (3\text{ real},\ u<\beta_2)$	$\cdots$	$\cdots$
$a\neq 0,\ q^2=0\ (3\text{ real},\ u>\beta_1)$	$\cdots$	$\cdots$
$a\neq 0,\ q^2=0\ (2\text{ equal},\ u<\beta_2)$	$\cdots$	$\cdots$
$a\neq 0,\ q^2=0\ (2\text{ equal},\ u>\beta_1)$	$\cdots$	$\cdots$
$a\neq 0,\ q^2=0,\ \lvert\ell\rvert\neq a\ (1\text{ real})$	$[\beta_3(3\beta_3+c_3)]^{\frac{1}{2}}$	$\frac{a+\ell}{a-\ell}$
$a\neq 0,\ q^2<0\ (2\text{ real})$	$\left[\frac{c_5(u-\beta_3)}{c_4(u-\beta_4)}\right]^{\frac{1}{2}}$	$\cdots$
$a\neq 0,\ q^2<0\ (0\text{ real})$	$\left[\frac{4\left(\Im(\beta_1)\right)^2-(c_4-c_5)^2}{(c_4+c_5)^2-4\left(\Im(\beta_1)\right)^2}\right]^{\frac{1}{2}}$	$\Re(\beta_1)+c_2\Im(\beta_1)$
$a\neq 0,\ q^2>0\ (4\text{ real},\ u<\beta_3)$	$\cdots$	$\cdots$
$a\neq 0,\ q^2>0\ (4\text{ real},\ u>\beta_2)$	$\cdots$	$\cdots$
$a\neq 0,\ q^2>0\ (2\text{ equal},\ u<\beta_3)$	$\frac{2\beta_3-\beta_1-\beta_4}{\beta_1-\beta_4}$	$(\beta_1-\beta_3)(\beta_3-\beta_4)$
$a\neq 0,\ q^2>0\ (2\text{ equal},\ u>\beta_2)$	$\frac{2\beta_3-\beta_1-\beta_4}{\beta_1-\beta_4}$	$(\beta_1-\beta_3)(\beta_3-\beta_4)$
$a\neq 0,\ q^2>0\ (2\text{ real})$	$\left[\frac{c_4(u-\beta_4)}{c_5(\beta_1-u)}\right]^{\frac{1}{2}}$	$\cdots$

TABLE 3.6

Auxiliary definitions for Tables 3.3, 3.4, and 3.5

Case	c_4
$a \neq 0,\ q^2 < 0$ (2 real)	$\left[\left(\Re(\beta_1) - \beta_3\right)^2 + \left(\Im(\beta_1)\right)^2\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 < 0$ (0 real)	$\left[\left(\Re(\beta_1) - \Re(\beta_3)\right)^2 + \left(\Im(\beta_1) + \Im(\beta_3)\right)^2\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (2 real)	$\left[\left(\Re(\beta_2) - \beta_1\right)^2 + \left(\Im(\beta_2)\right)^2\right]^{\frac{1}{2}}$

TABLE 3.6 (*continued*)

Case	c_5
$a \neq 0,\ q^2 < 0$ (2 real)	$\left[\left(\Re(\beta_1) - \beta_4\right)^2 + \left(\Im(\beta_1)\right)^2\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 < 0$ (0 real)	$\left[\left(\Re(\beta_1) - \Re(\beta_3)\right)^2 + \left(\Im(\beta_1) - \Im(\beta_3)\right)^2\right]^{\frac{1}{2}}$
$a \neq 0,\ q^2 > 0$ (2 real)	$\left[\left(\Re(\beta_2) - \beta_4\right)^2 + \left(\Im(\beta_2)\right)^2\right]^{\frac{1}{2}}$

APPENDIX B

WEAK DEFLECTION LIMIT

In this appendix we present a perturbative treatment of light focusing by a Kerr black hole for an observer at infinity and located in the equatorial plane $\mu_0 = 0, \phi_0 = 0$. We restrict attention to null geodesics with impact parameter $b \gg 1$. (Distances are still measured in units of the mass M.) The generalization to arbitrary observer inclination is straightforward, though algebraically involved and introduces no new principles.

In Appendix A, we give first order differential equations for the evolution of the four dependent variables t, μ, u, ϕ in terms of a modified affine parameter λ'. We perform a perturbation expansion in inverse powers of b. It is convenient to introduce an angle ψ so that the two components of the impact parameter introduced in Appendix A are $\alpha = b\cos\psi$, $\beta = b\sin\psi$. As $\mu_0 = 0$, $\ell = b\cos\psi$, and $q^2 = b^2\sin^2\psi$. It is also convenient to change the modified affine parameter λ' to χ, defined by

$$d\chi = \rho^{-2} b d\lambda' \tag{3.18}$$

First consider the equation for μ. In view of the singularity due to the vanishing of M near periapse, it is easiest to differentiate $d\mu/d\chi$ to obtain a non-singular equation,

$$\ddot{\mu} + \mu = \frac{a^2}{b^2}\mu(1 - 2\mu^2), \tag{3.19}$$

where a dot designates differentiation with respect to χ. We chose to integrate along an ingoing null geodesic starting at the observer ($\chi = 0$) and extending past the hole so that the signs of α, β should be reversed with respect to Appendix A. Equation (3.19) must be solved subject to the initial conditions $\mu(0) = 0, \dot{\mu}(0) = \sin\psi$. The right-hand side of this equation is smaller than the left-hand side by

$O(b^{-2})$, and so we can neglect it to obtain the zero'th order solution satisfying the boundary conditions.

$$\mu = \sin\psi \sin\chi \tag{3.20}$$

This solution clearly responds to rectilinear propagation in a flat space without a black hole.

We now solve equation (3.19) perturbatively, by writing $\mu = \sin\psi\sin\chi + \delta_\mu(\chi)$, where the perturbation δ_μ satisfies the differential equation

$$\ddot{\delta}_\mu + \delta_\mu = \frac{a^2}{b^2}\sin\psi\sin\chi(1 - 2\sin^2\psi\sin^2\chi) \equiv F(\chi). \tag{3.21}$$

This equation must be solved subject to the boundary conditions $\delta_\mu(0) = \dot{\delta}_\mu(0) = 0$. The solution is

$$\begin{aligned}\delta_\mu &= \int_0^\chi d\chi' \sin(\chi - \chi')F(\chi') \\ &= \frac{a^2\sin\psi}{16b^2}[8(\sin\chi - \chi\cos\chi) + \sin^2\psi(12\chi\cos\chi - 9\sin\chi - \sin 3\chi)].\end{aligned} \tag{3.22}$$

This perturbation accounts for the leading correction introduced by a Kerr hole on the light propagation.

Next consider the perturbation equation for the inverse radius coordinate, u. Again we remove the singularity near periapse by working with the second order differential equation obtained by differentiating $du/d\chi$. We obtain

$$\ddot{u} + u = 3u^2 + \frac{a^2u}{b^2}(1 - 2u^2b^2\sin^2\psi) - \frac{6a\cos\psi u^2}{b} \tag{3.23}$$

good to second order in b^{-1}. To leading order, we ignore the right-hand side to obtain

$$u = \frac{\sin\chi}{b}. \tag{3.24}$$

Again this solution corresponds to rectilinear propagation in a flat space. If we now retain only the first term on the right-hand side of equation (3.23), we can solve for the first order correction to u, $\Delta_u(\chi)$. Δ_u satisfies

$$\ddot{\Delta}_u + \Delta_u = \frac{3}{b^2}\sin^2\chi. \tag{3.25}$$

Solving this equation subject to $\dot{\Delta}_u(0) = \Delta_u(0) = 0$, we obtain

$$\Delta_u = b^{-2}(1 - \cos\chi)^2. \tag{3.26}$$

This correction accounts for the leading correction attributable to a Schwarzschild black hole. The ray, with impact parameter b is deflected by the hole onto the optic axis ($\chi = \pi$), when $u = \Delta_u = 4b^{-2}$. This verifies that, to lowest order, the deflection angle $\tilde{\alpha}$ is

$$\tilde{\alpha} = bu = 4/b. \tag{3.27}$$

The leading corrections attributable to the spin of the hole can be computed by writing $u = b^{-1}\sin\chi + b^{-2}(1-\cos\chi)^2 + \delta_u$. Retaining the remaining terms on the right-hand side of equation (3.23), we find that

$$\ddot{\delta}_u + \delta_u = \frac{6}{b^3}\sin\chi(1-\cos\chi)^2 - \frac{6a\cos\psi}{b^3}\sin^2\chi + \frac{a^2}{b^3}\sin\chi(1-2\sin^2\psi\sin^2\chi). \tag{3.28}$$

The first term on the right-hand side describes the second order correction to the Schwarzschild deflection. The second term, which is directly proportional to the specific angular momentum a, is traceable to the dragging of inertial frames and acts differentially on opposite sides of the hole. The third term $\propto a^2$ is due to the quadrupolar distortion caused by the hole's spin. Solving this equation according to equation (3.22), we obtain

$$\delta_u = \frac{5\sin\chi + 32\sin 2\chi - 3\sin 3\chi - 60\chi\cos\chi}{16b^3} - \frac{a\cos\psi(24 - 32\cos\chi + 8\cos 2\chi)}{8b^3} + \frac{a^2[7\sin\chi - 4\chi\cos\chi - \sin 3\chi + \cos 2\psi(9\sin\chi - 12\chi\cos\chi + \sin 3\chi)]}{32b^3}. \tag{3.29}$$

Finally, we expand $d\phi/d\chi$ to obtain the perturbation equation for ϕ.

$$\dot{\phi} = \frac{\cos\phi}{1-\mu^2} + \frac{au}{b}(2 - abu\cos\psi) \tag{3.30}$$

(Note that we do not need to resort to the second order differential equation in this case as there are no turning points for ϕ.) To leading order, we need only retain

the first term on the right-hand side of equation (3.30) and substitute expression (3.20). The solution for the undeflected ray behind the hole can be expressed in the form

$$\phi = \pi - \sin^{-1}\left[\frac{\sin\chi\cos\psi}{(1-\sin^2\chi\sin^2\psi)^{1/2}}\right]. \tag{3.31}$$

This can, indeed, be recognized as the equation for the flat space ray. Now, once more, we augment ϕ by a perturbation δ_ϕ, substitute $\mu = \sin\psi\sin\chi + \delta_\mu$ into the first term on the right-hand side of equation (3.29), to obtain,

$$\dot{\delta}_\phi = \frac{\sin 2\psi\sin\chi\delta_\mu}{(1-\sin^2\psi\sin^2\chi)^2} + \frac{a\sin\chi(2-a\sin\psi\sin\chi)}{b^2}. \tag{3.32}$$

The solution to equation (3.32) is obtained by simple quadrature,

$$\delta_\phi = \frac{2a}{b^2}(1-\cos\chi)+\frac{a^2\cos\psi}{8b^2}\left[\frac{2(\sin 2\chi - 2\chi)+\sin^2\psi\{6\chi-\sin 2\chi(3+2\sin^2\chi)\}}{(1-\sin^2\psi\sin^2\chi)}\right]. \tag{3.33}$$

Once again we can identify terms associated with the dragging of inertial frames ($\propto a$) and the quadrupolar distortion ($\propto a^2$).

We can now recover the equation for the caustic far behind the hole. We define a source "plane" to be the surface $u = 4/b_E$, where b_E is the Einstein ring radius for a source located a distance u^{-1} behind the hole, and introduce coordinates

$$\xi = (\pi-\phi)\frac{b_E^2}{4}, \qquad \eta = \mu\frac{b_E^2}{4} \tag{3.34}$$

on this surface, assuming that $|\xi|, |\eta| \ll b_E^2/4$. Next introduce $\chi = \pi - \delta_\chi$, where $|\delta_\chi| \ll 1$ in the vicinity of the caustic. To leading order in an expansion in b^{-1} we obtain, from equation (3.24), $u = \delta_\chi/b$. We can then iterate for δ_χ, substituting $\chi = \pi$ in equation (3.29) to obtain

$$\delta_\chi = \frac{4b}{b_E^2} - \frac{4}{b} - \frac{15\pi}{4b_E^2} + \frac{8a\cos\psi}{b_E^2} - \frac{\pi a^2(1+3\cos 2\psi)}{8b_E^2}. \tag{3.35}$$

We next substitute this expression for δ_χ into expansions for χ, η, incorporating equation (3.22), (3.26), (3.29) and (3.33) to obtain, after some algebra,

$$\begin{aligned} \xi &= B(b)\cos\psi + C(b)\cos 2\psi \\ \eta &= B(b)\sin\psi + C(b)\sin 2\psi \end{aligned} \tag{3.36}$$

where the coefficients of the third harmonic terms all vanish. The two coefficients B, C are functions of the impact parameter b and are given by

$$\begin{aligned} B(b) &= b - \frac{b_E^2}{b} - \frac{15}{16}\pi \\ C(b) &= \frac{b_E^2 a}{b^2}. \end{aligned} \tag{3.37}$$

Note that the terms $\propto a^2$ have all vanished to this order. The coefficient B must be expanded to first order about the zero'th order impact parameter b_E by writing $b = b_E + \delta_b$. As $C = O(b^{-1})B$, it is not necessary to expand $C(b)$.

In order to understand the mapping represented by equation (3.36), consider a ring of rays whose impact parameters form a circle of radius b_E and whose center is displaced from the center of the black hole by a distance $\delta_\rho - a$ along the equatorial direction $\psi = 0$. In this case, $\delta_b = (\delta_\rho - a)\cos\psi$. Substituting this particular perturbation into equation (3.37), we obtain the circle's map on the source plane given by the equation

$$(\xi + a - \delta_\rho)^2 + \eta^2 = \delta_\rho^2. \tag{3.38}$$

Therefore, these circles map onto circles in the source plane displaced by the same amount $\delta_\rho - a$ and of radius δ_ρ. In particular, for $\delta_\rho = 0$, the circle degenerates to a point located at $\xi = -a, \eta = 0$. This is the caustic and the corresponding critical curve is a circle displaced by $-a$ from the Einstein ring for a Schwarzschild black hole along the equatorial direction. Therefore, to the order of our perturbation expansion, the caustic surface for a distant observer is a line displaced by a from the optic axis.

We can also use our perturbation expansion to calculate the magnification of a point source located close to the optic axis. To do this, we must calculate the Hessian for the transformation equation (3.36). We write $\alpha = b\cos\psi, \beta = b\sin\psi$. The Hessian is

$$H(\alpha, \beta) = \frac{\partial\xi}{\partial\alpha} \cdot \frac{\partial\eta}{\partial\beta} - \frac{\partial\xi}{\partial\beta} \cdot \frac{\partial\eta}{\partial\alpha}. \tag{3.39}$$

A straightforward evaluation gives

$$H(\xi,\eta) = \frac{2}{b_E}\left[(\xi+a)^2+\eta^2\right]^{1/2}. \tag{3.40}$$

The image magnification, $\mathcal{M}$, which is the reciprocal of the Hessian, is then inversely proportional to the distance $\zeta = [(\xi+a)^2+\eta^2]^{1/2}$ of the source from the caustic line, i.e.,

$$\mathcal{M} = \frac{b_E}{2\zeta} \tag{3.41}$$

per image, giving a total magnification of $\sim b_E/\zeta$. This is the standard relation for a point mass.

In the numerical computations, the line caustic unfolds into a four-cusped surface as expected. In order to reproduce this, the perturbation expansion must be pursued to higher order. The width of the caustic cross section is $O(u) = O(b^{-2})$. In order to compute it, it is then necessary to expand δu to $O(b^{-5})$ instead of $0(b^{-3})$ and $\delta\mu, \delta\phi$ to $O(b^{-4})$ instead of $O(b^{-2})$. This is straightforward in principle, though lengthy in practice, and we have not attempted this calculation as the lower order results are sufficient to interpret and verify our numerical results.

There is one asymptotic relation which can be quickly derived and which provides a final check on the computations. Although we have not computed the combined magnification of the four images for a source lying at the center of the four-cusped caustic, flux conservation requires that it be inversely proportional to the size of the caustic. Following Blandford & Kovner (1988), we consider a nearly circular, normalized Newtonian potential of the form

$$\Phi = f(b^2/2) + \epsilon\psi_2(b)\cos 2\psi; \quad \epsilon \ll 1, \tag{3.42}$$

where the function $f(x)$ describes the circularly symmetric deflection and the function $\psi_2(b)\cos 2\psi$ describes the elliptical perturbation. In this notation, the radius of the Einstein ring satisfies $f(b_E^2/2) = 1$ and the Hessian is given by

$$\left(1-\frac{\partial^2\Phi}{\partial b^2}\right)\left(1-\frac{1}{b}\frac{\partial\Phi}{\partial b}-\frac{1}{b^2}\frac{\partial^2\Phi}{\partial\phi^2}\right)-\left[\frac{\partial}{\partial b}\left(\frac{1}{b}\frac{\partial\Phi}{\partial\psi}\right)\right]^2. \tag{3.43}$$

For a source at the center of the caustic, the four images are located close to the Einstein ring at $\psi = 0, \pi/2, \pi$, and $3\pi/2$. The Hessian for each image evaluates to

$$H_0 = \pm 4\epsilon\psi_2(b_E)\frac{d^2 f}{dx^2}(b_E^2/2). \tag{3.44}$$

However, we can also set $H = 0$ to locate the caustic and find that, asymptotically, the total width of the caustic is given by

$$s = \frac{8\epsilon\psi_2(b_E)}{b_E} \tag{3.45}$$

(Blandford & Kovner 1988). Now to lowest order, when $b \gg 1$ we can treat the black hole as a Newtonian point mass and take $f(x) = (b_E^2/2)\ln(2x)$ so that

$$H_0 = \pm\frac{8\epsilon\psi_2(b_E)}{b_E^2} = \pm\frac{s}{b_E}. \tag{3.46}$$

The total asymptotic magnification for a source located at the center of the caustic therefore satisfies

$$\mathcal{M}_0 = \frac{4}{|H_0|} = \frac{4b_E}{s}. \tag{3.47}$$

This result depends only on the existence of a small elliptical perturbation to the circular symmetry, not on the details of the gravitational field close to the black hole. It should therefore be applicable to the present problem, and equation (3.47) is in fact satisfied by our numerical results.

REFERENCES

Abramowitz, M., & Stegun, I. A. 1968, *Handbook of Mathematical Functions* (New York: Dover Publications)

Bardeen, J. M. 1973, in *Black Holes*, ed. C. DeWitt & B. S. DeWitt (New York: Gordon & Breach), p. 44

Berry, M. V., & Upstill, C. 1980, Progress in Optics, 18, 258

Biehle, G. T., & Blandford, R. D. 1993, ApJ, 411, 302

Blandford, R. D., & Kovner, I. 1988, Phys. Rev. A, 38, 4028

Blandford, R. D., & Narayan, R. 1992, ARA&A, 30, 311

Carter, B. 1968, Phys. Rev., 174, 1559

Celotti, A., Fabian, A. C., & Rees, M. J. 1992, MNRAS, 255, 419

Cunningham, C. T., & Bardeen, J. M. 1973, ApJ, 183, 237

Darwin, C. 1959, Proc. Roy. Soc. London A, 249, 180

Ferland, G. J., & Rees, M. J. 1988, ApJ, 332, 141

Fukue, J. 1988, in *Physics of Neutron Stars and Black Holes*, ed. Y. Tanaka (Tokyo: Universal Academy Press), p. 215

Fukue, J., & Yokoyama, T. 1988, PASJ, 40, 15

Griest, K., & Hu, W. 1992, ApJ, 397, 362

Jaroszyński, M., Wambsganss, J., & Paczyński, B. 1992, ApJ, 396, L65

Karas, V., Vokrouhlický, D., & Polnarev, A. G. 1992, MNRAS, 259, 569

Kormendy, J., & Richstone, D. 1992, ApJ, 393, 559

Laor, A., Netzer, H., & Piran, T. 1990, MNRAS, 242, 560

Luminet, J.-P. 1979, A&A, 75, 228

Madau, P. 1988, ApJ, 327, 116

Misner, C. W., Thorne, K. S., & Wheeler, J. A. 1973, *Gravitation* (San Francisco: W. H. Freeman & Co.) (MTW)

Mittaz, J. P. D., & Branduardi-Raymont, G. 1989, MNRAS, 238, 1029

Piessens, R., de Doncker, E., Uberhuber, C. W., & Kahaner, D. K. 1983, *QUADPACK: A Subroutine Package for Automatic Integration* (New York: Springer-Verlag)

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, *NUMERICAL RECIPES in C: The Art of Scientific Computing* (Cambridge: Cambridge Univ. Press)

Rauch, K. P., & Blandford, R. D. 1991, ApJ, 381, L39

Remillard, R. A., McClintock, J. E., & Bailyn, C. D. 1992, ApJ, 399, 445

Schneider, P., Ehlers, J., & Falco, E. E. 1992, *Gravitational Lensing* (Berlin: Springer-Verlag)

Sun, W. H., & Malkan, M. 1989, ApJ, 346, 68

4

Dynamical Evolution of a Star Cluster Around a Rotating Black Hole with an Accretion Disk†

Kevin P. Rauch

California Institute of Technology

ABSTRACT

The dynamical evolution of the core of a dense star cluster brought into the equatorial plane around a massive Kerr black hole is examined. It is assumed that the black hole dominates the potential and that the evolution of the cluster is driven by stellar interactions with an accretion disk lying in the black hole's equatorial plane, and that stellar collisions and two-body relaxation are unimportant. It is shown that there are astrophysically interesting circumstances under which these assumptions are valid. A set of diagrams exhibiting the time development of the stellar density profile and the distributions of semi-major axes, orbital eccentricities, and orbital inclinations of the model system are presented; plots of the latter three quantities against each other and as a function of time for a few illustrative individual orbits are also given. It is found that the main effect of star-disk interactions on the cluster, besides the general circularization and alignment of orbits with the disk, is to steepen the initial cusp profile towards an equilibrium $\rho_\star \propto r^{-3}$ stellar density profile and to increase the central density (by several hundred very close to the black hole). Relativistic effects were found to affect the cluster properties significantly only at very small radii ($\lesssim 10\,M$, where M is the mass of the hole and in units where $G = c = 1$); in particular, the location of the

† Submitted to Monthly Notices of the Royal Astronomical Society.

last stable orbit limits the cluster's inner extent. By significantly increasing the central stellar density, star-disk interactions could self-limit themselves by making stellar collisions important; the future evolution of the cluster in this case will probably depend on the relative balance of the collisional, alignment, and stellar evolutionary timescales.

Subject headings: stellar dynamics – galaxies: nuclei – accretion disks

4.1. INTRODUCTION

There is a promising opportunity to deepen considerably our understanding of the dynamical evolution of active galactic nuclei (AGNs), and galactic nuclei in general, in the next decade. On the observational front, the prospects of a repaired Hubble Space Telescope (HST) and a new generation of large, ground-based telescopes with adaptive optics should allow galactic nuclei to be studied both on smaller angular scales and over a larger range of frequencies than ever before. Observations by HST of the nuclei of M32 (Lauer et al. 1992b; Lugger et al. 1992) and M87 (Lauer et al. 1992a), for instance, already provide evidence for stellar cusps which continue into the center as far as the current resolution limits of the data, giving information on conditions in the nuclear star cluster within a parsec (in the case of M32) of the galaxy's center. Such observations could finally provide convincing evidence for the existence of central compact masses, commonly supposed to be massive black holes, that are often thought to reside in the centers of galaxies (for an overview of the observational evidence so far, see Kormendy 1994 and references therein), but in any event should markedly improve our knowledge of the distribution of stars and gas in the central parsec of nearby galaxies. On the numerical front, computing power is now approaching that needed to perform N-body simulations of globular clusters with realistic numbers of stars, putting analogous calculations in the context of galactic nuclei on the visible horizon. The dynamical evolution of AGNs, however, is likely to be quite complicated because of the components that they might contain in addition to dense central star clusters, such as massive black holes, accretion disks or other large scale gas flows, magnetic fields, possible star formation, etc., and the interaction of all of these elements is not well understood. The present paper investigates one piece of this evolutionary jigsaw puzzle: the dynamical evolution of the central regions of a dense star cluster around a Kerr black hole with an accretion disk, in a regime where the black hole dominates the potential and star-disk interactions dominate the evolution of the cluster.

Several groups have investigated the dynamical evolution of AGNs (e.g., Duncan & Shapiro 1983; David, Durisen, & Cohn 1987a,b; Quinlan & Shapiro 1990; Murphy, Cohn, & Durisen 1991). All of these studies performed calculations using an orbit-averaged Fokker-Planck formulation to follow the evolution of the central star cluster, in which processes such as physical collisions between stars, stellar evolution, and interactions with binaries were incorporated in a statistical manner. These studies assumed that either no central black hole, or only a small "seed" black hole, was present initially, with focus being placed on the probable time of formation or on the rate of growth of the black hole, as well as on the overall dynamical evolution of the cluster. Although the Fokker-Planck formalism provides a good approximation on larger scales where there are many stars present in each zone followed, in the inner regions, where the numbers are small, the continuum approximation becomes a poor one. In addition, if there is a black hole present in the center, general relativistic effects, which would be difficult to include properly in an orbit-averaged calculation, will become important in its vicinity. Both of these problems can be overcome by using N-body methods in this region.

The orbital evolution of a single star interacting with an accretion disk has also been examined previously (Syer, Clarke, & Rees 1991; Karas & Vokrouhlický 1993; Artymowicz, Lin, & Wampler 1993). The qualitative effects of an accretion disk on an orbit initially inclined with respect to it are to cause the orbital plane to align with the disk, to increase the orbital binding energy, and to circularize the orbit (assuming the disk material is traveling on nearly circular orbits), due to the small momentum kicks and kinetic energy losses the star suffers upon crashing into the disk twice each orbit. The detailed physics of the star-disk interaction is uncertain; in those studies (as well as this one) it was assumed that the interaction was well described as an impulse applied to the star at the point of intersection with the disk, in the form of a drag in the direction of the star-disk relative velocity. This approximation should be a good one when the disk is thin and the stellar orbital velocity is supersonic with respect to the local disk sound speed. Artymowicz et al. (1993) proposed a disk-induced 'star-trapping' process to explain metallicity

enrichment in AGNs. The work of Syer et al. (1991) explored star-disk interactions in the Newtonian regime, and also speculated on the possible evolutionary paths traversed after the star has been brought into the plane of the disk. Karas & Vokrouhlický (1993) contains a brief treatment of the process in the relativistic case. Some of the calculations presented here are complementary to those of Karas & Vokrouhlický and examine the evolution of the orbital parameters of a single star in greater detail, also using a fully general relativistic treatment. In addition, the time evolution of the density and distribution of orbital parameters for a star cluster around a (rotating) black hole is also considered. This work is part of a larger project to investigate relativistic effects of rotating black holes and the Kerr spacetime in the context of AGNs.

This paper uses an N-body approach (i.e., one in which individual stars are followed) to examine the dynamical evolution of a dense star cluster around a rotating, massive black hole possessing an accretion disk, subject to a number of restrictions (explained in § 4.2). A few illustrative results for the time development of orbital characteristics for individual stars are also given. The aim is to study not only the manner in which the cluster evolves, but also the importance of relativistic influences on that development, as judged, for instance, by its dependence on the angular momentum of the black hole. General relativistic effects on the stellar orbits are included accurately using an efficient, high precision code to follow geodesics in the Kerr spacetime that has been developed previously (see Rauch & Blandford 1994 for a description). The routines evaluate some of the equations of motion analytically in terms of elliptic functions and elliptic integrals for fast and accurate calculation of the geodesic trajectories. That code was then integrated into an N-body routine that traced the time evolution of a cluster of stars under the influence of star-disk interactions. The following section describes the assumptions used in the calculations and the range of conditions over which they are likely to be valid; results of the calculations are presented in § 4.3. Discussion is given in the final section.

4.2. METHOD AND ASSUMPTIONS

The computational model consisted of three components: a central massive black hole, an associated accretion disk lying in the equatorial plane of the hole, and a surrounding star cluster. To restrict attention to the influence of star-disk interactions on the evolution, and to keep the problem numerically tractable, the calculations assume that:

1) The central black hole dominates the gravitational potential.

2) The accretion disk is thin, and the stars impact the disk supersonically.

3) Star-disk interactions dominate the evolution of the stellar orbits.

The first constraint merely requires that the mass of the black hole be much larger than the mass of the cluster (and disk) interior to the maximum radius considered in the simulations. In this limit, each star essentially follows a geodesic orbit around the hole (except when it is interacting with the disk), which is traced in the simulations using the Kerr code mentioned previously. The simulations used a canonical black hole mass of 10^8 $M_\odot$, making this approximation a very good one, although the results can be scaled to any hole mass for which the assumptions are still valid. The second assumption allows the star-disk collisions to be treated as impulsive interactions, as noted previously—or at least until the star's orbital plane becomes nearly aligned with the disk, when the star will skim through the disk for most or all of its orbit. In this limit the drag force felt by a star passing through the disk due to both dynamical friction produced by the disk (Ostriker 1983) and the excitation of disk normal modes (Artymowicz 1993) are unimportant, each being smaller by a factor $\sim (v_{\rm esc}/v_{\rm orb})^4$ ($\lesssim 10^{-3}$ in the case considered here) than the 'ballistic' drag caused by the sweeping up of intercepted disk material (cf. eq. (2) of Artymowicz et al. 1993). Validity of the final condition requires that two-body relaxation, physical collisions between stars, and tidal disruption of stars passing close to the black hole be unimportant relative to star-disk collisions. The relative timescales for these processes in the fiducial model are given below; except perhaps for stellar collisions, all of these other processes turn out to be unimportant in the present case.

The use of the fast Kerr code allowed simulations with up to 10^4 particles to be performed. The simulations were done as follows. One specific model consisting of 10^4 stars was calculated to allow a reference model to be followed in as much detail as possible. To estimate the dependence of the results on the prescription for the surface density of the disk, the star-disk interaction, and the initial stellar density profile, a number of smaller simulations (using 1000 stars) were done, in which one of the prescriptions was varied from that of the reference model, the evolution recomputed, and the results compared. Each model is specified by the mass and angular momentum of the black hole, and the functional forms for the surface density of the disk (assumed to remain constant over time), the initial density profile of the cluster, and the star-disk interaction itself. The mass of the black hole was fixed at $M_8 \equiv (M/10^8\ \mathrm{M}_\odot) = 1$ in all models, and the star cluster was truncated at an inner radius $r_{\mathrm{min}} = 10\,M$ (approximately the tidal radius for the stars) and an outer radius of $r_{\mathrm{max}} = 10^4\,M$ (we will use units in which $G = c = 1$). The reference model assumed a maximally rotating black hole. A real cluster would of course extend far beyond r_{max} ($\approx 0.05\,M_8$ pc); limiting attention to this region is the result of several factors. First, since $r_{\mathrm{max}} \gg M$, relativistic effects will be unimportant outside this region. Second, the number of stars that can be included limits the range of radii that can be considered before small-number statistics make the results hard to interpret. In addition, at some larger radius ($\sim 10^5\,M$ in this case) the cluster and disk mass will become large enough that the black hole will no longer dominate the potential, contrary to the assumption. The cluster was initially spherically symmetric and isotropic, giving it a wide range of initial eccentricities and orbital inclinations.

The validity of the assumption that star-disk collisions dominate the evolution depends on both the density of the disk and the density of stars in the cluster. It is easily seen that in these simulations, tidal disruption and two-body relaxation are relatively unimportant; for a 10^8 $\mathrm{M}_\odot$ black hole, the tidal radius $r_{\mathrm{T}} \sim R_\star(M/M_\star)^{1/3}$ is approximately the horizon size for solar-type stars, and the two-body relaxation time $t_{\mathrm{r}} \sim v^3/(\pi G^2 M_\star^2 n_\star) \sim 10^{12}(n_\star/10^7\,\mathrm{pc}^{-3})^{-1}$ years at

$r_{\rm max}$. Stellar collisions, however, may or may not be important. To see this, note that the timescale for a star to be aligned with the disk is approximately the time needed for the star to intercept (in total) a quantity of disk material equal to its own mass. Since the star hits the disk twice per orbit, the alignment timescale relative to the orbital time is $t_{\rm align}/t_{\rm orb} \sim \Sigma_\star/(2\Sigma_{\rm d})$, where $\Sigma_\star \equiv M_\star/(\pi R_\star^2)$ is the effective stellar surface density and $\Sigma_{\rm d}$ is the disk surface density. To get an estimate for $\Sigma_{\rm d}(r)$, the Novikov-Thorne (1973) thin accretion disk model was used. Assuming moderate values of 0.1 for both the accretion rate relative to the Eddington limit, $\ell_{\rm E}$, and the mass to luminosity conversion efficiency, ϵ, one obtains

$$\Sigma_{\rm d}(r) \sim 2\times 10^{-3}\, M_8^{-1/2} \left(\frac{\alpha}{0.01}\right)^{-4/5} \left(\frac{\ell_{\rm E}}{\epsilon}\right)^{7/10} \left(\frac{r}{M}\right)^{-3/4} {\rm M}_\odot/{\rm R}_\odot^2 \qquad (4.1)$$

and

$$\frac{t_{\rm align}}{t_{\rm orb}} \sim 100\, M_8^{1/2} \left(\frac{\alpha}{0.01}\right)^{4/5} \left(\frac{\ell_{\rm E}}{\epsilon}\right)^{-7/10} \left(\frac{r}{M}\right)^{3/4} \qquad (4.2)$$

(for solar-type stars), where α is the viscosity parameter; $t_{\rm orb}/M \sim 2\pi(r/M)^{3/2}$ in these units. Note that since $r_\star \propto M_\star^{0.6}$ for above solar mass main sequence stars (e.g., Kippenhahn & Weigert 1990), $\Sigma_\star \propto M_\star^{-0.2}$ varies slowly with the stellar mass and the results of the simulations should be insensitive to the assumed mass spectrum; although most of the simulations used a range of masses for the stars (following a Salpeter distribution between 0.1 and 10 $\rm M_\odot$) since it incurred negligible cost, runs with equal (solar) mass stars confirmed this to be the case. (For sub-solar stars, the relation is more like $r_\star \propto M_\star^{0.8}$, but the conclusion is still valid.) The timescale for physical collisions between stars to occur is $t_{\rm coll} \sim 1/(4\pi R_\star^2 n_\star v_{\rm orb})$ (note that since $v_{\rm orb}/v_{{\rm esc},\star} \gg 1$ in the region under consideration, there is no gravitational focusing enhancement to $t_{\rm coll}$); assuming a cusp density of the form $n_\star(r) \propto r^{-7/4}$ (e.g., Bahcall & Wolf 1976), as assumed in the reference model, gives

$$\frac{t_{\rm align}}{t_{\rm coll}} \sim \frac{4\pi r \rho_\star(r)}{\Sigma_{\rm d}(r)} \sim 0.03\, M_8^{-1/4} \left(\frac{\alpha}{0.01}\right)^{4/5} \left(\frac{\ell_{\rm E}}{\epsilon}\right)^{-7/10} \left(\frac{\rho_\star(1\,{\rm pc})}{10^6\,{\rm M}_\odot\,{\rm pc}^{-3}}\right). \qquad (4.3)$$

Hence for $\alpha \lesssim 0.1$ or $\rho_\star(1\,\mathrm{pc}) \lesssim 10^7\,\mathrm{M}_\odot\,\mathrm{pc}^{-3}$, the alignment timescale is shorter than the average collision time and star-disk interactions dominate; evolution in the regime where collisions dominate will be considered in a future paper. For the reference model, $t_{\mathrm{align}}(r_{\mathrm{max}}) \sim 10^7$ years (and hence stellar evolution effects will also be unimportant for the initial development of the cluster in this scenario). For convenience, future references to the 'fiducial accretion disk model' will mean an α-disk with parameter $\alpha = 0.01$ and to the 'fiducial stellar density model' will imply $\rho_\star(r) = 10^6\,(r/1\ \mathrm{pc})^{-7/4}\ \mathrm{M}_\odot\ \mathrm{pc}^{-3}$. Note that the fiducial disk model has a total mass within radius r_{max} of $\sim 2 \times 10^7(\alpha/0.01)^{-4/5}\ \mathrm{M}_\odot < M$, hence self-gravity in the disk (which could induce fragmentation within the disk) should be a small effect in these simulations; similarly, the star cluster's mass within radius r_{max} is $\sim 2 \times 10^5\,(\rho_\star(1\,\mathrm{pc})/10^6\,\mathrm{M}_\odot\,\mathrm{pc}^{-3})\,\mathrm{M}_\odot$, well below the hole mass.

The reference model used $\rho_\star(r) \propto r^{-7/4}$ and $\Sigma_{\mathrm{d}}(r) = 10^{-3}r^{-1/2}\ \mathrm{M}_\odot/\mathrm{R}_\odot^2$ (the fiducial disk model of eq. (4.1) was not used because, due to the more rapid $r^{-3/4}$ fall off in surface density, using it would have required that each star be followed for too many orbital periods before being brought into the disk plane than was computationally feasible); note that since no star-star interactions are included, the absolute value of the stellar density is not relevant (i.e., the final density is directly proportional to the initial value), and was in any case limited by the number of stars that could be followed. Dependence on these choices was estimated by examination of the results of smaller simulations in which $\Sigma_{\mathrm{d}}(r) = 10^{-4}r^{-1/2}$, $10^{-3}r^0$, and $10^{-3}r^{-1}\ \mathrm{M}_\odot/\mathrm{R}_\odot^2$ and $\rho_\star(r) \propto r^{-1}$ and $\propto r^{-3}$ were individually substituted for the original prescription for Σ_{d} or $\rho_\star$.

The star-disk interaction was computed as follows. The point of intersection of the orbit with the disk plane was solved for numerically for each passage through the disk. The 4-velocities of the star and disk at the intersection point were then transformed to the locally non-rotating frame (e.g., Misner, Thorne, & Wheeler 1973), and the star's 4-velocity was then Lorentz boosted from that orthonormal frame into the local rest frame of the disk. Let $\mathbf{P}_{\mathrm{rel}}$ be the star's usual spatial 3-momentum in this latter frame. Using the impulse approximation, the change

in the star's momentum due to the drag induced upon passing through the disk is

$$\delta \mathbf{P} = -\frac{A_{\rm swept}\Sigma_{\rm d}}{M_\star}\,\mathbf{P}_{\rm rel} \approx -\frac{\pi R_\star^2 \Sigma_{\rm d}}{M_\star}\left(\frac{v_{\rm rel}}{v_\perp}\right)\mathbf{P}_{\rm rel} = -\left(\frac{\Sigma_{\rm d}}{\Sigma_\star}\right)\left(\frac{v_{\rm rel}}{v_\perp}\right)\mathbf{P}_{\rm rel}, \quad (4.4)$$

where $A_{\rm swept}$ is the effective disk area swept out by the star as it passes through the disk, $v_{\rm rel}$ and $v_\perp$ are the star's velocity relative to and velocity component perpendicular to the disk (in the disk's rest frame), respectively, and it is assumed that all disk material impacted is given the same velocity as the star (but is *not* actually accreted by the star). This form is similar to one used by Karas & Vokrouhlický (1993) and essentially identical to the one used by Artymowicz et al. (1993), and implicitly assumes that the disk thickness is greater than the size of the star (as it typically would be). The star's new momentum was then transformed back to the coordinate frame, the new constants of motion were computed, and the perturbed orbit followed until the next disk intersection, where the process was repeated. To judge the sensitivity of the results on this prescription, the factor $(v_{\rm rel}/v_\perp)$ was replaced with $(1+v_\parallel/v_\perp)$ in some simulations; the latter form would be appropriate if the disk was infinitely thin (the two forms differ by a factor of $\sqrt{2}$ at most). Orbits were evolved until they were brought within $\sim 1^\circ$ of the disk plane; in the reference model, this required $\sim 10^4 - 10^5$ disk crossings per star. Results of the calculations are described in the following section.

4.3. SIMULATION RESULTS

4.3.1 *Single Star Results*

The evolution of a cluster dominated by star-disk interactions is essentially an ensemble average of the (independent) development of each star's individual orbit. To allow easier interpretation of the full simulations, the results for a single star are given first. The time-varying quantities of interest are the semi-major axis, a, the eccentricity, e, and the inclination with respect to the disk, i. Figures 4.1, 4.2, and 4.3 each plot on a semi-log scale several $a(t)$, $e(t)$, and $i(t)$ curves, respectively, for an orbit with initial parameters $a_0 = 10^4$, $e_0 = 0.95$, and several

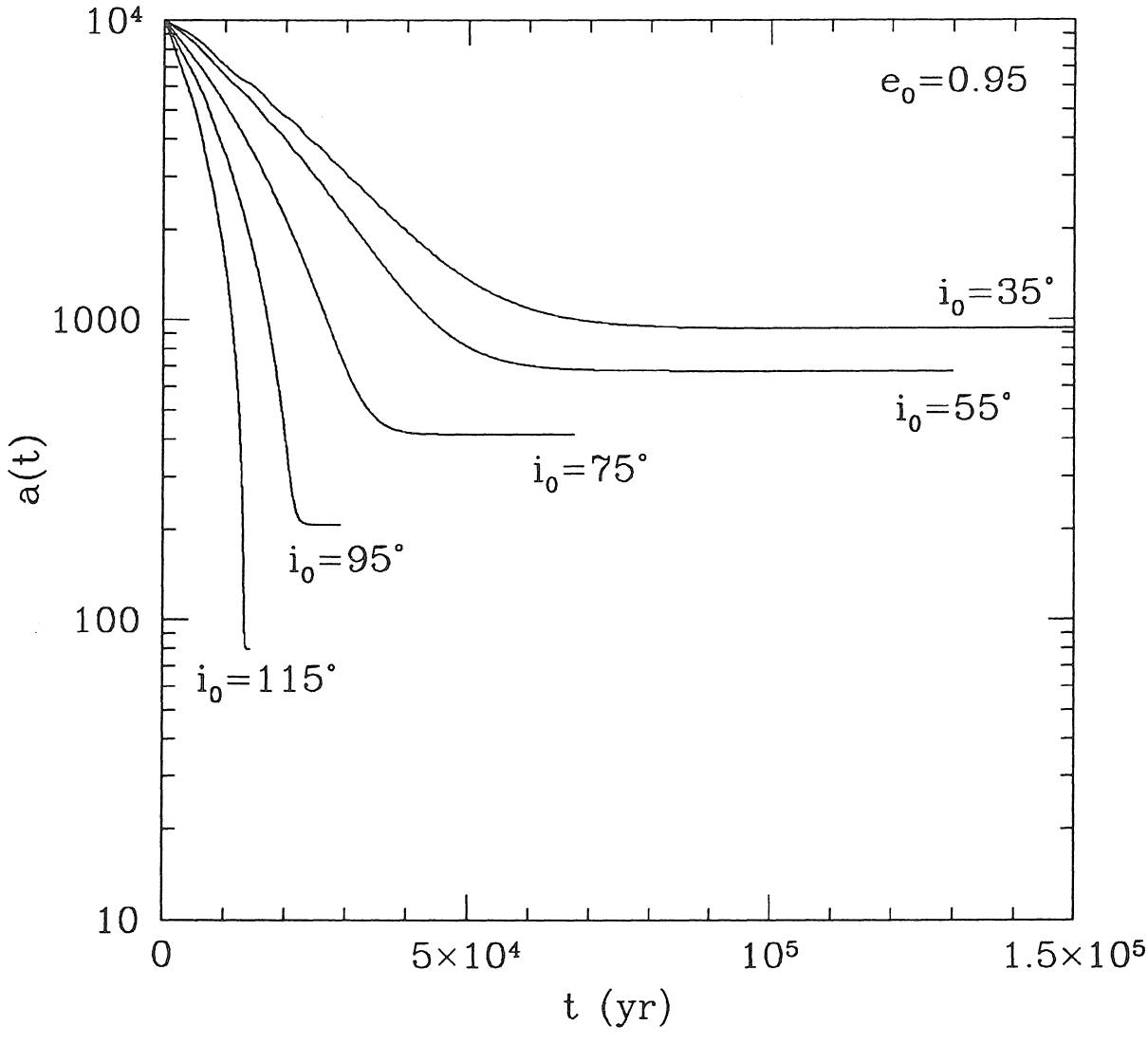

Fig. 4.1: The semi-major axis, a, as a function of time for five orbits under the influence of star-disk interactions in the potential of a 10^8 $M_\odot$, maximally rotating black hole. To avoid clutter the curves have been truncated some time after the asymptotic value has been reached. Initially, each orbit has $a_0 = 10^4$ (in units of M) and eccentricity $e_0 = 0.95$ but differing inclinations i_0 with respect to disk plane (marked). Highly inclined orbits become very tightly bound; in this case, orbits with $i_0 \gtrsim 135°$ were captured by the black hole. The physical time scales linearly with the hole mass and inversely with the disk surface density (the plotted value assumes $M = 10^8$ $M_\odot$ and $\Sigma_{\rm d}(r) = 2 \times 10^{-3}\,(r/M)^{-3/4}$ $M_\odot$ $R_\odot^{-2}$).

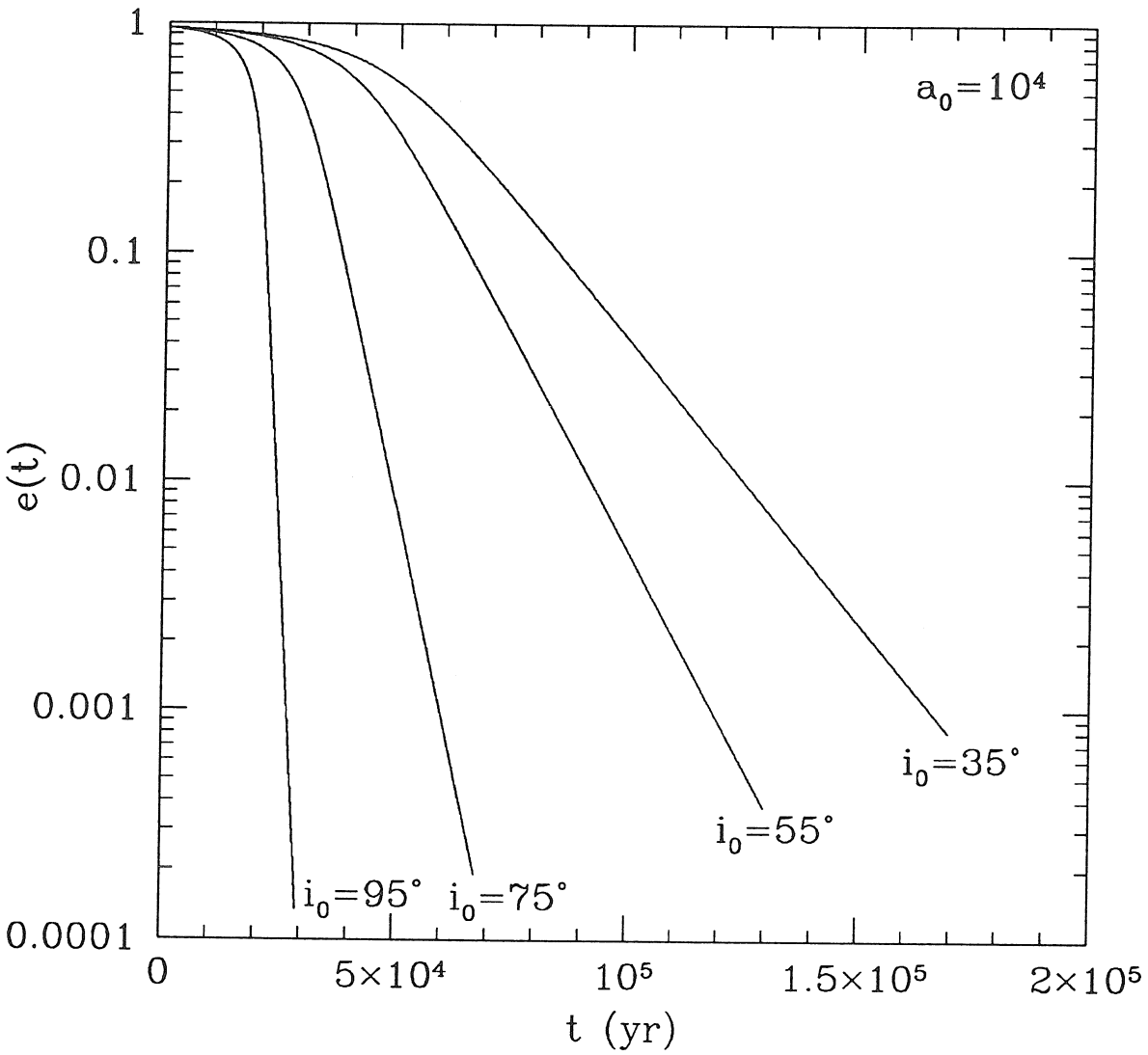

Fig. 4.2: As Figure 4.1 but showing the eccentricity e as a function of time. At late times (roughly where the inclination has dropped below $\sim 20 - 30°$) e falls exponentially. Note the semi-log scale.

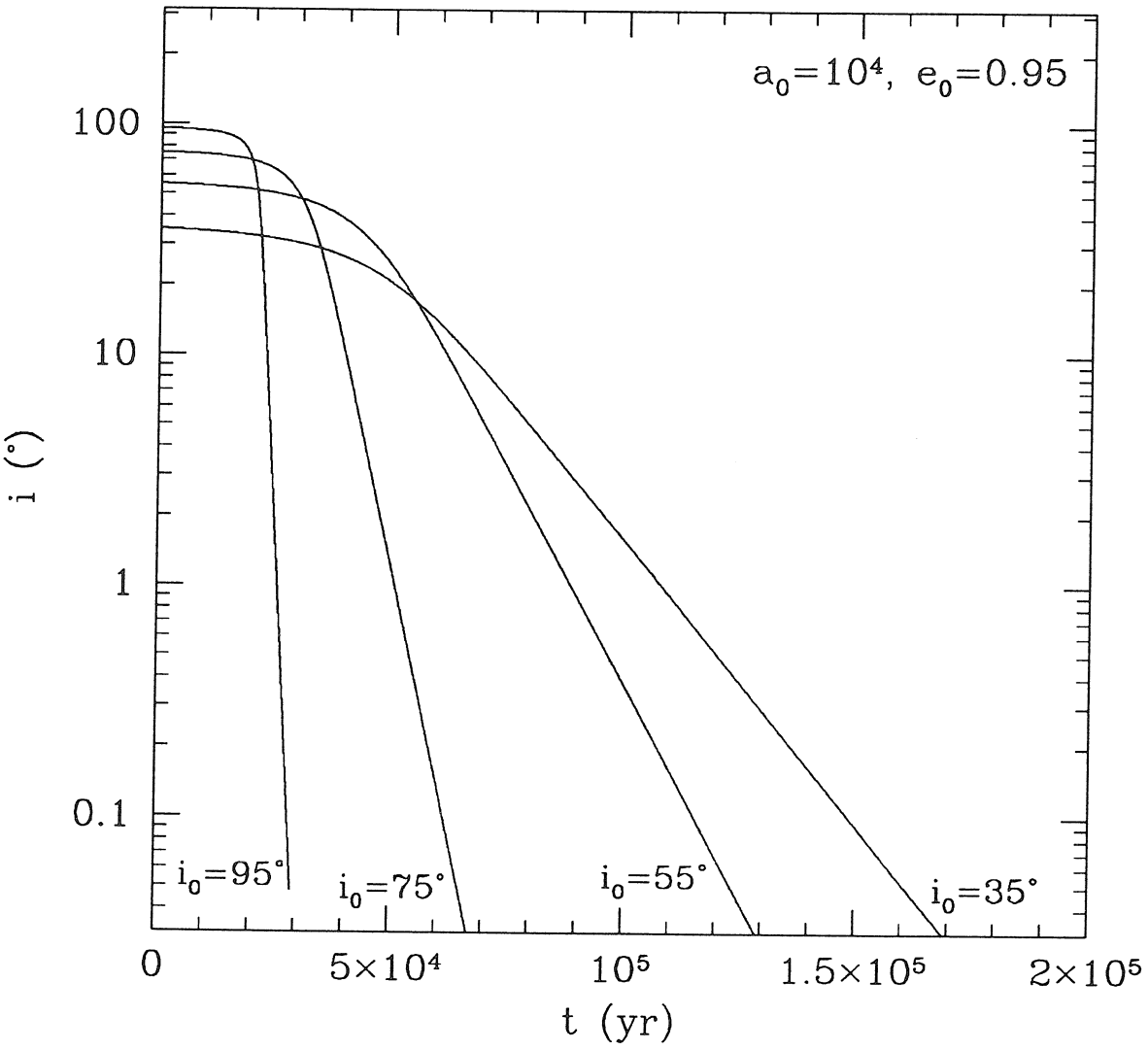

Fig. 4.3: As Figure 4.1 but showing the inclination with respect to the disk i as a function of time. The inclination, like the eccentricity, decreases exponentially once it is lower than $\sim 20 - 30°$; in this region, $e \propto i$ (cf. Figure 4.6).

values of i_0 (marked). The absolute value of the time coordinate given assumes the parameters $M_8 = 1$ and $\Sigma_{\rm d}(r) \sim 2 \times 10^{-3}\, r^{-3/4}$ $\mathrm{M}_\odot$ $\mathrm{R}_\odot^{-2}$ (the fiducial disk model) and scales as $t \propto a_0^{3/2}\, M_8\, \Sigma_{\rm d}^{-1}$, assuming a fixed functional form for $\Sigma_{\rm d}$ (recall also that a_0 is in units of M here). The rate and extent of the orbital development is strongly dependent upon the initial orbital inclination; orbits with large inclinations evolve much faster than those with smaller inclinations, as noted by Syer et al. (1991). This is due mainly to the fact that highly inclined orbits are quickly driven into tightly bound, short-period orbits, because of the relatively large kinetic energy and angular momentum losses suffered by the star on passing through the disk with a velocity vector nearly anti-aligned with the disk's. Note also that both $e(t)$ and $i(t)$ decrease exponentially (with the same time constant) at late times, the transition to exponential behavior occurring when the inclination falls below $\sim 20 - 30°$. Until this happens, a decreases more quickly than either e or i, and after the transition begins to approach a constant value.

The relative timescales are more easily seen in Figures 4.4, 4.5, and 4.6, which plot curves of e vs. a, i vs. a, and e vs. i, respectively. The tick marks on each curve are spaced at equal time intervals (although the size of the time step is somewhat different for each curve). For large values of e, the orbital size shrinks much more rapidly than the eccentricity, largely independent of the inclination; once the eccentricity finally does become relatively small, both a and e decrease at similar rates (Figure 4.4). Likewise, for large inclinations both the semi-major axis and the eccentricity of the orbit decrease at nearly constant inclination (Figures 4.5 and 4.6). Note also that $e(t) \propto i(t)$ for $i \lesssim 30°$ (i.e., when both e and i decrease exponentially) and that e undergoes stochastic oscillations when the inclination is large (Figures 4.4 and 4.6). In contrast to Karas & Vokrouhlický (1993), it was found that only *highly* retrograde orbits (roughly speaking, those with $i_0 > 135°$) became captured by the black hole before being brought into the plane, at least for large enough orbits (those with initial periapses $\gtrsim 100$). Rapidly rotating black holes in particular captured retrograde orbits less frequently than Schwarzschild holes, due to the fact that stable (prograde) orbits exist closer to the horizon of

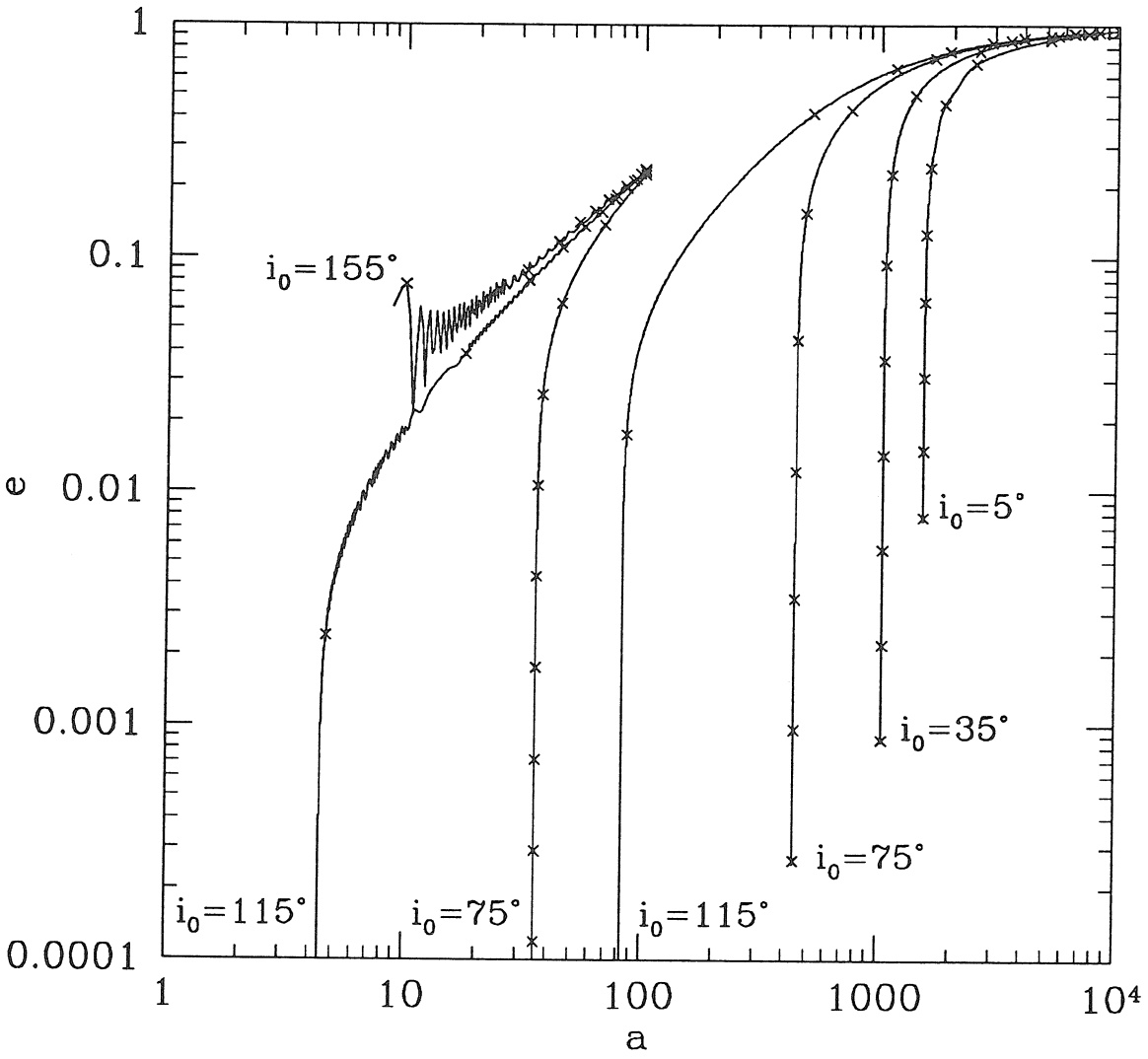

Fig. 4.4: Similar to Figure 4.1 but now plotting e against a for several orbits. For large e_0, e varies slowly as a decreases, while for smaller e_0 both quantities decrease at similar rates (until a eventually stabilizes). The tick marks on the curves denote increments of equal time, with approximately 10^4 years between each. The orbit marked $i_0 = 155°$ was captured by the black hole.

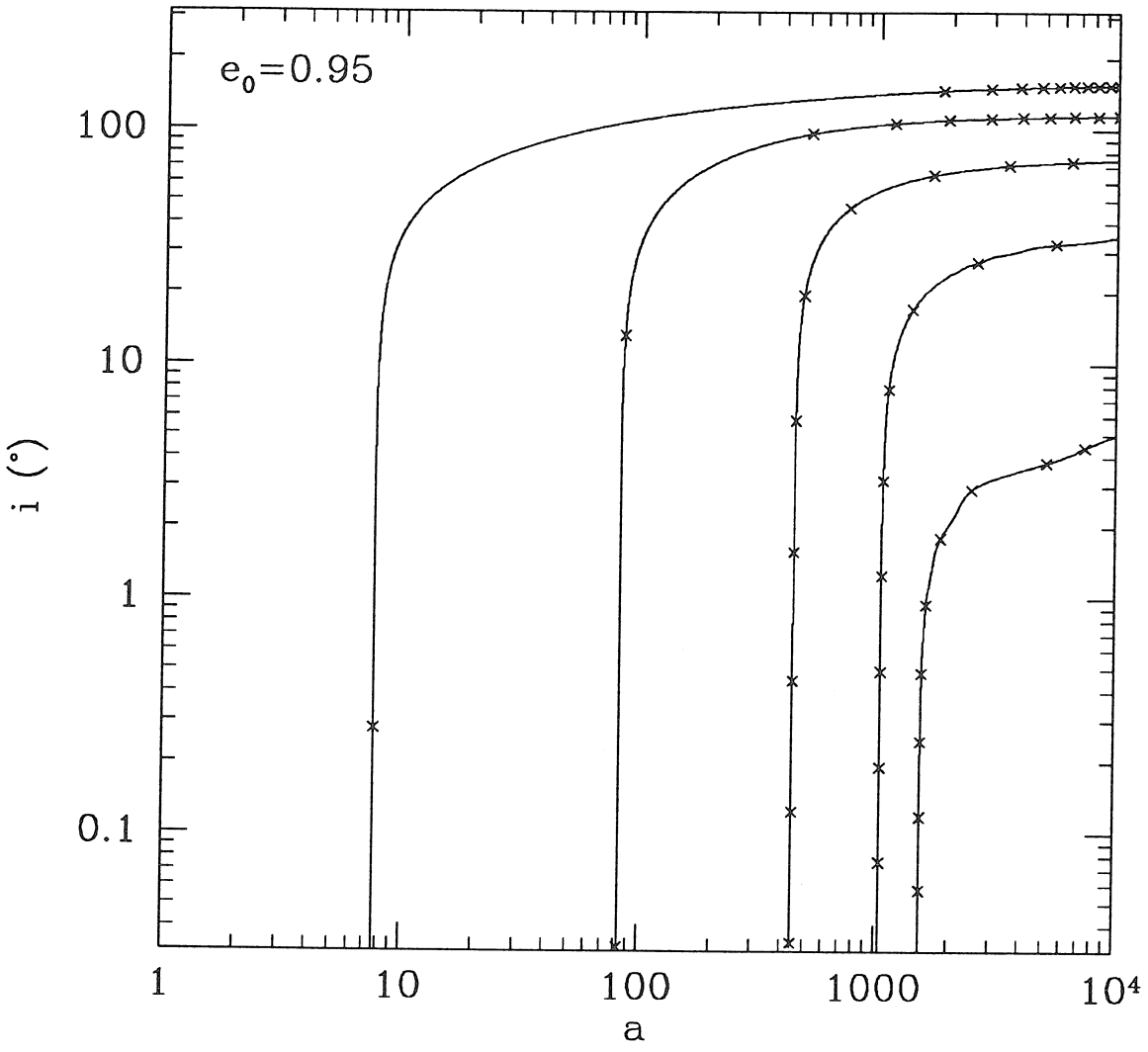

Fig. 4.5: As Figure 4.4 but now plotting i against a for several orbits. For $i_0 \gtrsim 10°$, a decreases at nearly constant i almost until it reaches its asymptotic value.

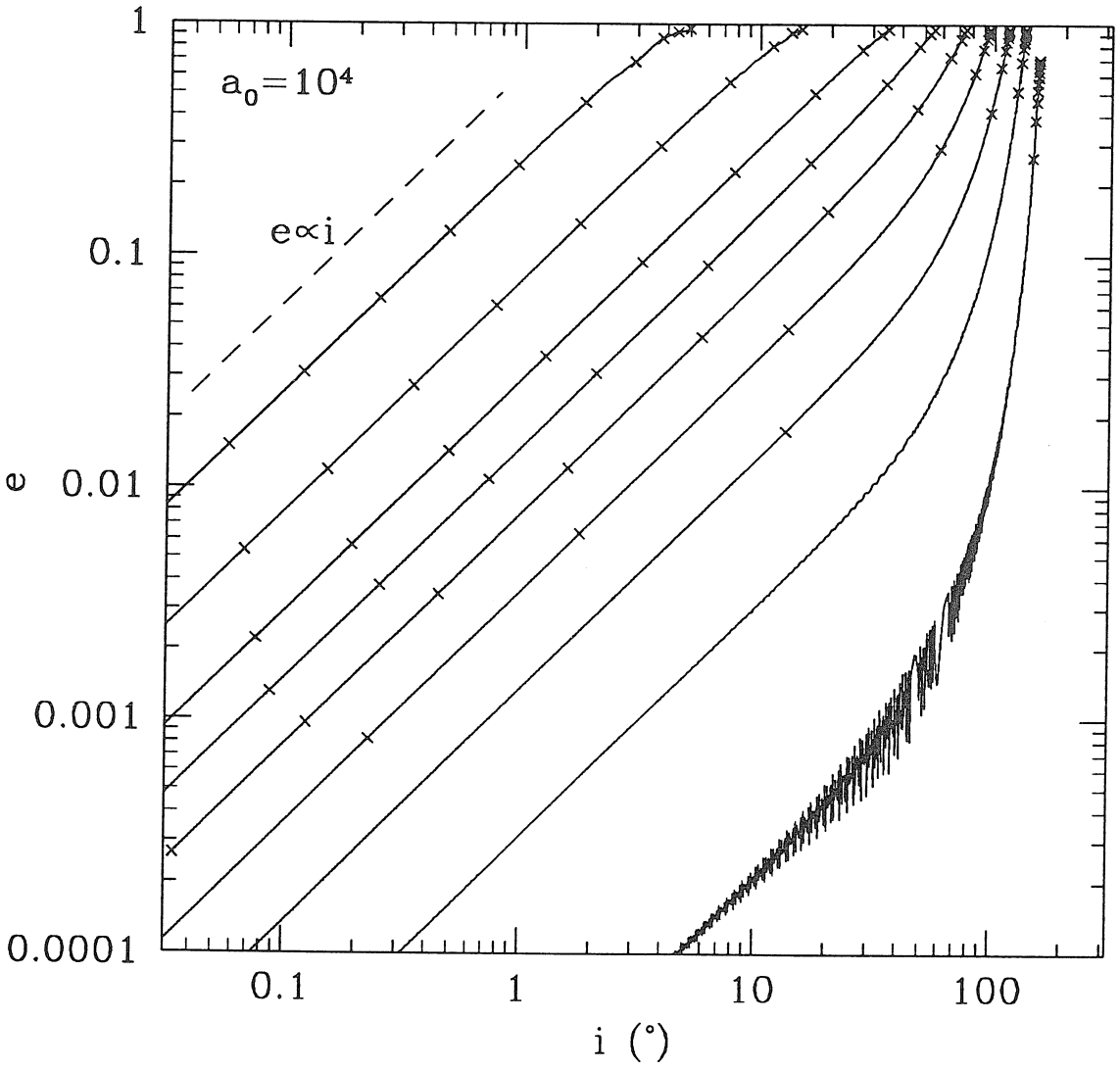

Fig. 4.6: As Figure 4.4 but now plotting e against i. For $i \lesssim 30°$, $e \propto i$. Note the stochastic oscillations suffered by the eccentricity for large values of i_0.

a rotating black hole than a non-rotating one—as long as the initially retrograde orbit becomes prograde by the time it has sunk close to the horizon, the hole's angular momentum helps "prop up" the orbit near the horizon, giving it a little more time to align with the disk and avoid being captured.

4.3.2 *Cluster Evolution Results*

Results for the cluster evolution are given as successive snapshots showing the distributions of a, e, i and the stellar density profile at several stages of the evolution (Figures 4.7, 4.8, 4.9, and 4.10, respectively). Each figure consists of four panels showing the state of the cluster initially, at two intermediate times, and at a later time when all stars have been either aligned with the disk or captured, for the reference cluster model described in § 4.2. The qualitative features of the development of the a, e, and i distributions are just what would be expected on the basis of the previous discussion. As the evolution progresses, the orbits with small a_0 evolve rapidly relative to the larger orbits, creating a small-a hump which subsequently expands outwards to larger radii (Figure 4.7). The initially symmetric (about $i = 90°$) distribution of inclinations becomes increasingly skewed towards small values, and early on becomes especially deficient in highly retrograde orbits, due to the more rapid development of these orbits (Figure 4.9). The eccentricities also become increasingly peaked at small values, as expected (Figure 4.8).

The stellar density profile shows considerable evolution (Figure 4.10). Of all the different prescriptions for the disk surface density, initial stellar cusp profile, etc., that were examined to judge the sensitivity on the results of the assumptions made in the reference model, only changes in the initial density profile of the cluster produced noticeable changes in the 'final' appearance of the cluster, and even here the effects were not large (in reality the cluster density would still change somewhat, as more distant stars inside or near the loss cone passed into the central regions, but this occurs more slowly, on the two-body relaxation timescale for the outer cluster). As seen in Figure 4.10, the initial density profile $\rho_{\star,\mathrm{i}}(r) \propto r^{-7/4}$ used in the reference model evolves to a final state well fit by another power

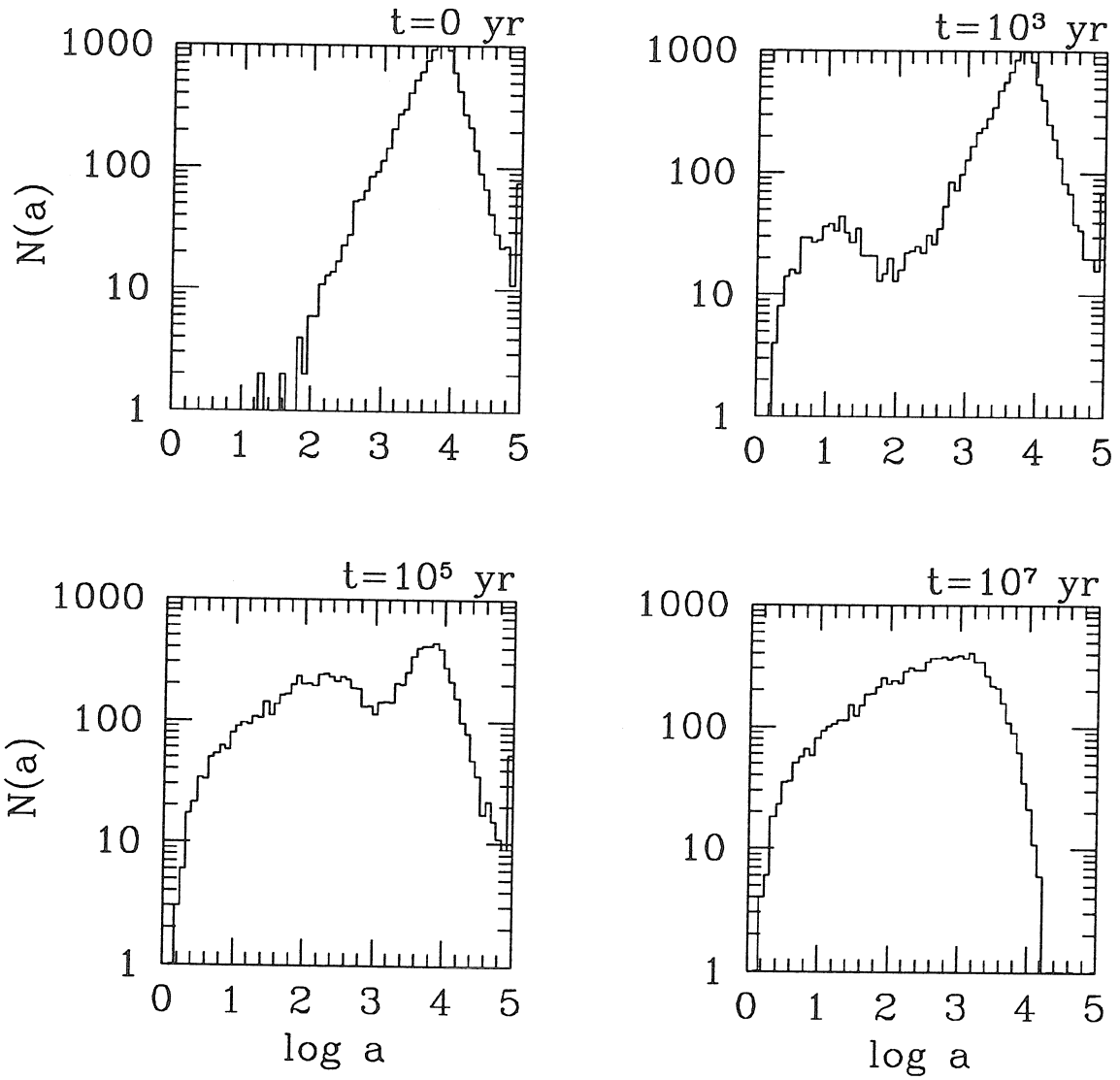

Fig. 4.7: The distribution of semi-major axes for the reference cluster model (§ 4.2) at four specific times during its evolution under the influence of an accretion disk, showing the initial distribution, two intermediate ones, and the final distribution (which is reached when all stars have been essentially aligned with the disk). The main effect of star-disk interactions is to broaden the distribution and shift it toward smaller radii.

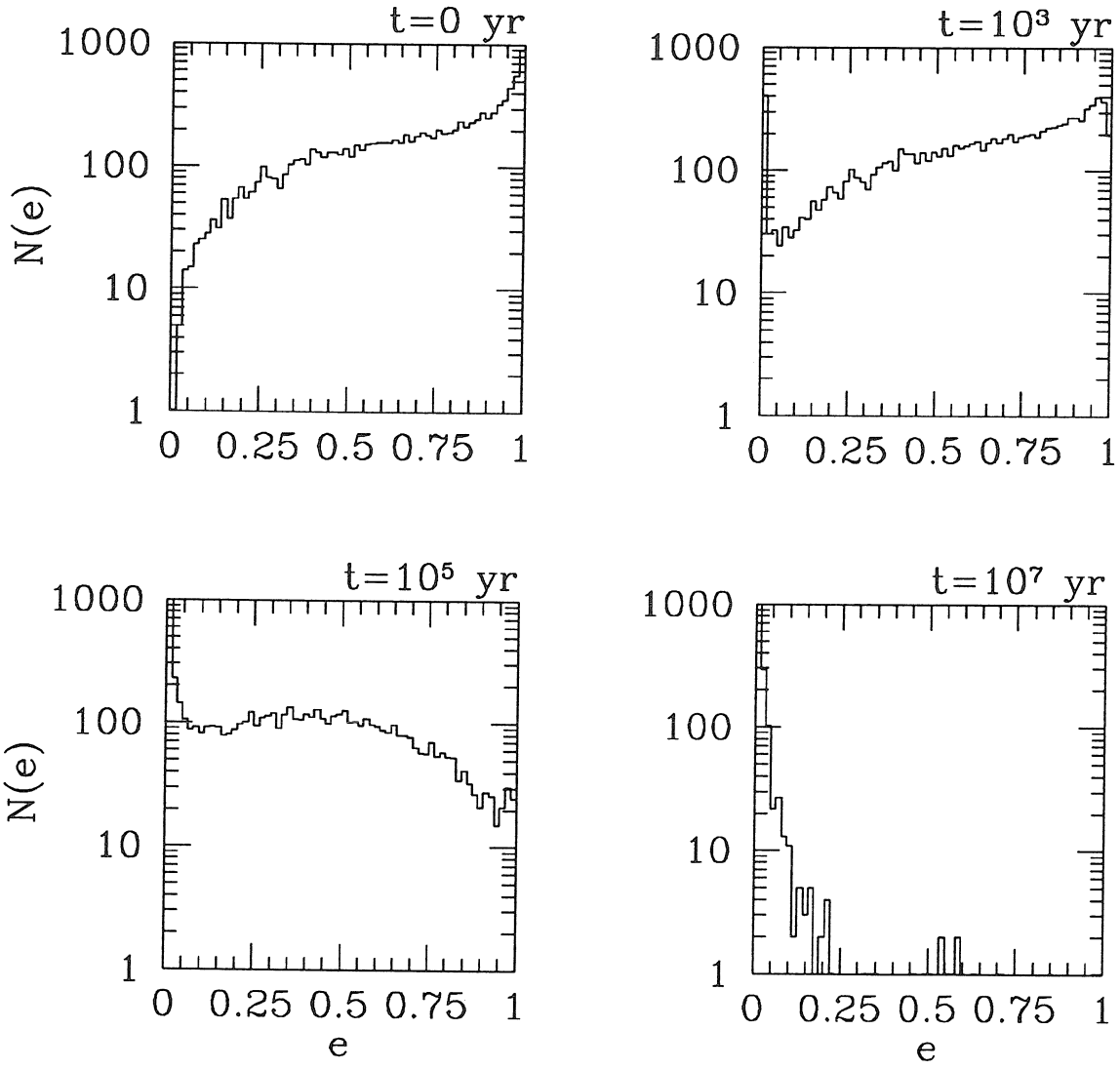

Fig. 4.8: Same as Figure 4.7 only now showing the distribution of eccentricities. The orbits are circularized in an orderly fashion.

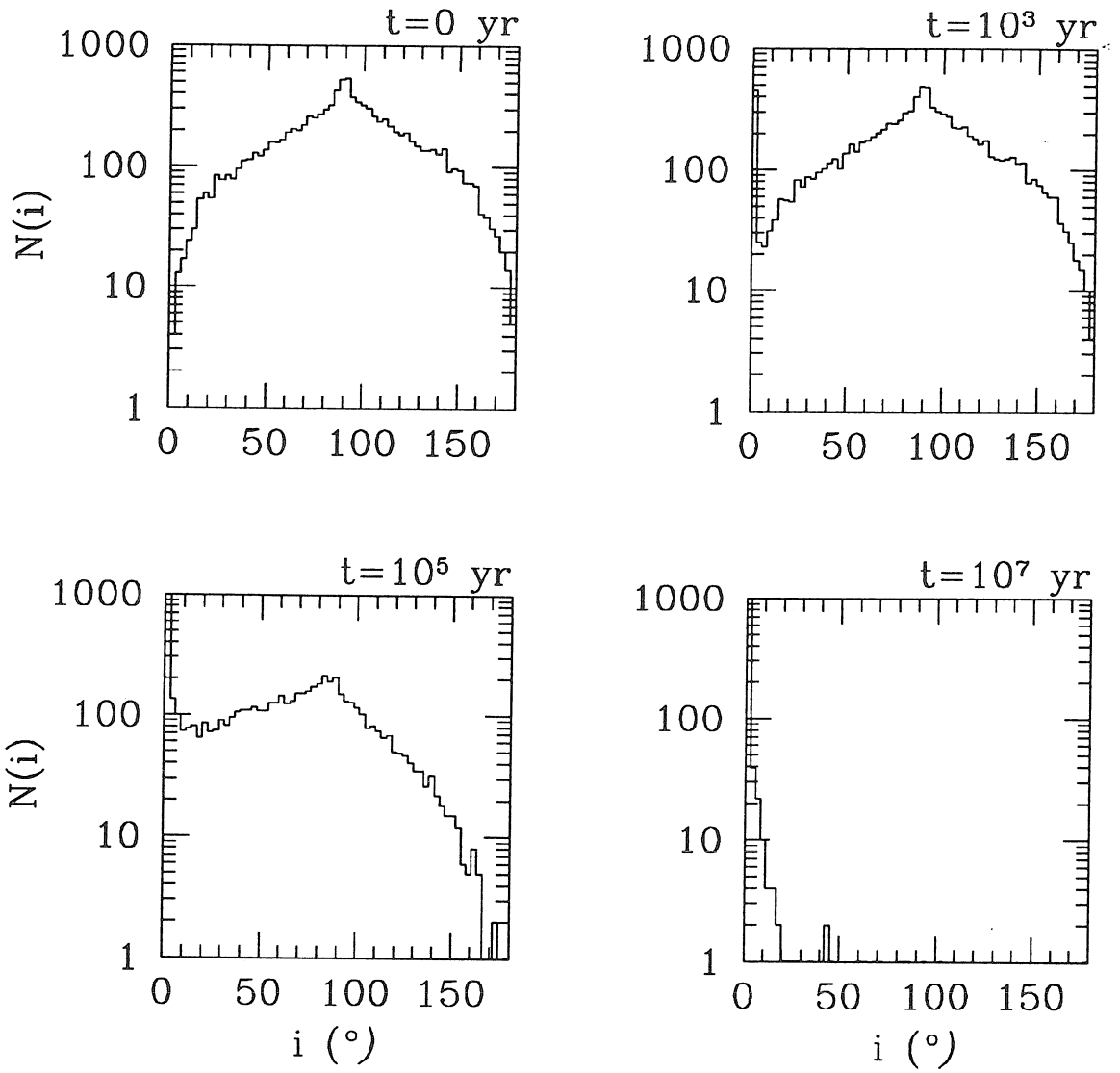

Fig. 4.9: Same as Figure 4.7 only now showing the distribution of orbital inclinations. The highly retrograde orbits evolve much more quickly than the others, either becoming captured or stabilizing at small radii, causing the distribution to become deficient in these orbits early on. Simultaneously, the less inclined orbits are driven towards smaller inclinations in a more orderly manner. At late times all stars which were not captured have been brought into the disk plane. About 7% of the stars in the simulation were captured.

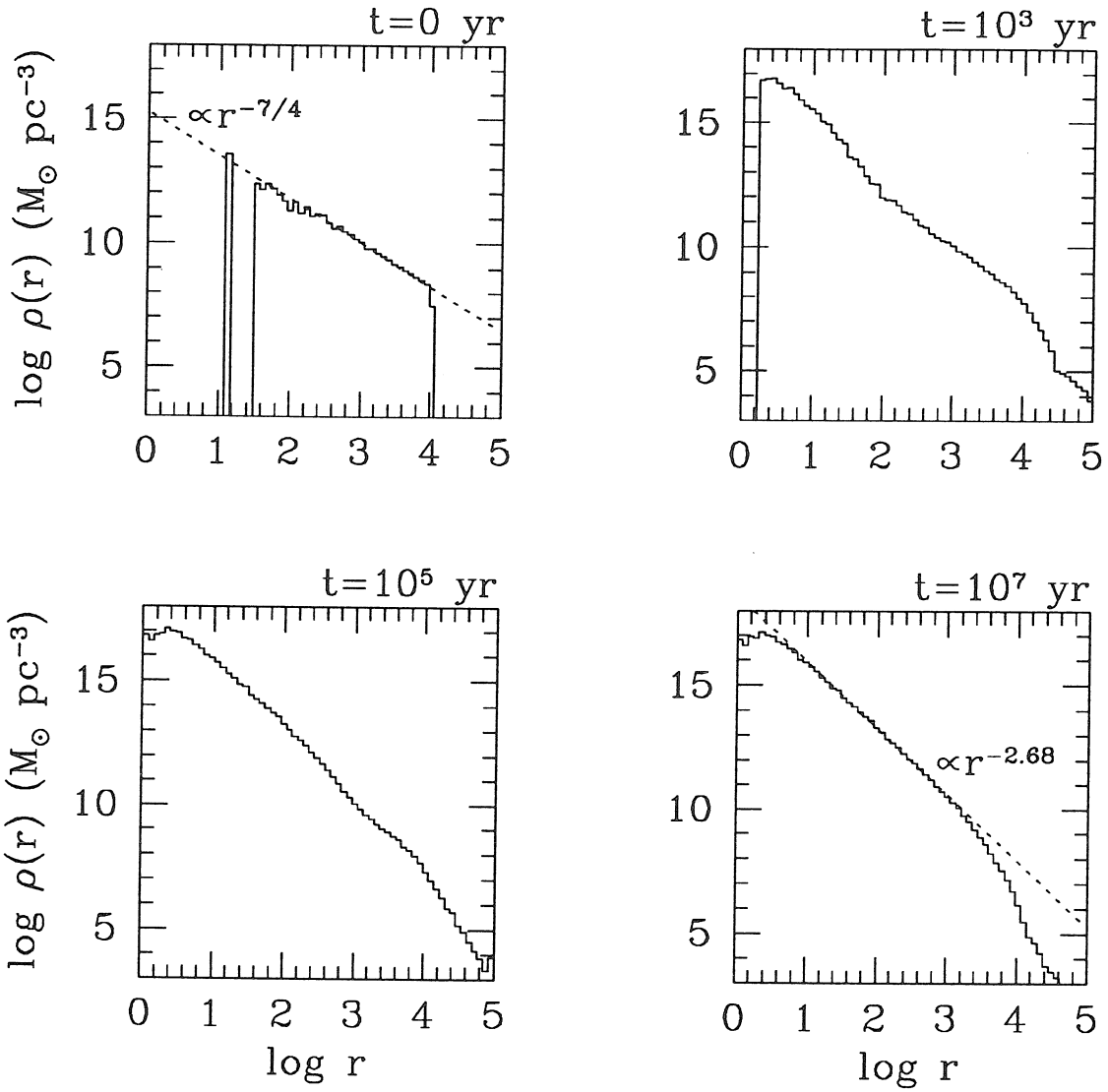

Fig. 4.10: Same as Figure 4.7 only now showing the evolution in the stellar density profile. Initially $\rho_\star \propto r^{-7/4}$, while at late times the profile is well fit by $\rho_\star \propto r^{-2.68}$ for $r \gtrsim 10$ (the falloff in density for $r \gtrsim 10^3$ is due to incompleteness, an artifact of having truncated the initial cluster at a radius of 10^4). The accretion disk drives the profile towards an 'equilibrium' r^{-3} value and increases the density close to the black hole at the expense of large radii (see § 4.3.2).

law, $\rho_{\star,\mathrm{f}}(r) \propto r^{-2.68\pm0.03}$, for $r > 10$; the final density profile is incomplete for $r > 1000$ due to having initially truncated the cluster at $r = 10^4$, but the power law should still apply here as long as sufficient time has elapsed for those larger orbits to evolve, at least at radii where the black hole still dominates the potential ($r \lesssim 10^5$). The cluster becomes denser at smaller radii at the expense of larger radii as the stars become more tightly bound under the action of the disk. It turned out that in this model, the density at $r \sim 3000$ was the same at late times as at early times; at smaller radii the cluster was denser at late times (by a factor of over 100 near the black hole), and less dense at larger radii. For $\rho_{\star,\mathrm{i}}(r) \propto r^{-1}$, the final density profile was found to be $\rho_{\star,\mathrm{f}}(r) \propto r^{-2.62\pm0.03}$, a small change. For $\rho_{\star,\mathrm{i}}(r) \propto r^{-3}$ (equal numbers of stars per decade of radius), the final density profile was $\rho_{\star,\mathrm{f}}(r) \propto r^{-3.03\pm0.03}$, essentially identical to the initial one. The main effect of the disk on the cluster, besides the generic ones of alignment and circularization, thus appears to be to drive the density profile towards an equilibrium r^{-3} distribution, in which stars are uniformly distributed in $\log r$. The consequences of this are discussed further in the following section.

4.4. DISCUSSION

We have shown that there is an astrophysically plausible regime in which an accretion disk around a massive black hole in an AGN or other galactic nucleus would dominate the dynamical evolution of a star cluster, through the process of star-disk interactions. In the innermost regions of the cluster that were considered in this paper, where the velocity dispersions are high (> 1000 km s^{-1}), only physical collisions between stars would be expected (in some cases) to compete with a disk in this respect, except possibly for tidal disruption of stars near the black hole, if the black hole is only 'modestly' supermassive ($M \sim 10^6 - 10^7$ M$_\odot$). Because of the steepening of the cusp profile and accompanying increase in the central densities (by several hundred times near the horizon), however, even if the evolution is originally dominated by the disk, densities could rise sufficiently to

make stellar collisions important again, preventing the cluster from being driven entirely into the plane of the disk. The subsequent evolution under these circumstances will likely depend upon the extent to which the orbits had been circularized and aligned with the disk at the time collisions became important, as the *relative* velocities between the stars, upon which the outcome of the collisional processes depend, could be considerably smaller than the absolute Keplerian velocity of the orbits if the stars are nearly comoving. If the relative velocities became less than the stellar escape velocities (~ 500 km s^{-1}), qualitatively different collision products, such as merged objects, could be produced. This same factor would also work to decrease the absolute rate of collisions, however, since the collision rate would also be tied to the typical relative velocity, which would no longer be the (much larger) orbital velocity. At this point, the timescales could become long enough for stellar evolution effects to become significant. It is quite possible that the longer term development ($> 10^7$ years in this case) of the nucleus will depend strongly on the detailed balance between these different evolutionary timescales. A series of supernovae occurring in the inner disk could destroy the disk in that region entirely, for example, leaving behind a tightly bound population of neutron stars and stellar mass black holes. The orbits of such a population of objects (which may also include white dwarfs, if the timescales are long enough) would no longer be significantly modified by either collisional processes (between the compact objects themselves) or by further interactions with the disk, as the timescale for both processes would increase by a factor $(\mathrm{R}_\odot/R_{\mathrm{co}})^2 \sim 10^4$ for white dwarfs and $\sim 10^9$ for neutron stars or black holes; for the reference cluster, the original timescale (at ~ 0.1 pc) of $\sim 10^7$ years now becomes longer than a Hubble time. The phase space distribution at this point becomes 'frozen in,' and in principle contains information on the density of the accretion disk around the time of freeze-out, regardless of conditions in, or even the existence of, the accretion disk at later times. If the black hole continued to (slowly) accrete matter, the orbits would each adiabatically conserve their angular momentum L and their component of L perpendicular to the equatorial plane L_z (Lee & Goodman 1989; see also Young

1980), so that the flattening of the cluster would be preserved. Although collisions between the remnants and remaining main sequence (MS) stars or giants could still be important, the significance and outcome of such events would again depend on the relative velocities involved. In high speed encounters, for example, the remnant would "core" the larger star's envelope but would emerge relatively unscathed (although the MS star could be disrupted), whereas in a low velocity encounter with a giant star, e.g., the compact object could be captured by the larger star and a common envelope system could be formed (Davies 1993).

Overall, relativistic effects on the results were subtle. Because most orbits were nearly circular and not very highly inclined by the time the typical orbit became relativistic, precession effects, such as the Lense-Thirring (L-T) precession of the orbital plane, could not significantly influence the orbits at that point. Orbits passing near the black hole that *did* have large inclinations, which were subject to large amounts L-T precession, were ultimately captured and hence did not affect the late-time density distribution. Inside $\sim 10\,M$, however, relativistic dynamics were very important; in particular, the radius of the smallest stable (prograde) circular orbit ($= 6\,M$ for a non-rotating hole and M for a maximally rotating one) strongly influences the locations of the inner edge of both the accretion disk and the cluster in this region. Outside $\sim 10\,M$ relativistic effects did not noticeably influence the gross properties of the cluster (cf. Figure 4.10).

There are other effects which could conceivably limit the importance of an initially dominant star-disk interaction. A dense enough cluster, for instance, could impart kinetic energy to the disk at a faster rate than the disk is gaining binding energy through mass accretion. For the fiducial cluster model, the binding energy inside $r_{\rm max}$ is easily estimated to be $E_{\rm b} \sim 2 \times 10^{56}$ erg; the typical star decreased its semi-major axis by about a factor of ten over the course of the simulation, hence roughly 2×10^{57} erg was transferred to the disk in the $\sim 10^7$ years that it took for the cluster to be brought into the disk plane. Since the average energy transfer rate ($\sim 10^{43}$ erg s^{-1}) this implies is much less than the Eddington luminosity limit of $\sim 10^{46}$ erg s^{-1} for a 10^8 M$_\odot$ black hole, an extremely

dense cluster ($\rho_\star(1\ \mathrm{pc}) \gtrsim 10^9\ \mathrm{M}_\odot\ \mathrm{pc}^{-3}$) would be needed for this effect to become important. The evolution of a cluster this dense would most likely be dominated by stellar collisions, not star-disk interactions (cf. eq. (4.3)). Observed galactic nuclei have central densities $\sim 10^6 - 10^7\ \mathrm{M}_\odot\ \mathrm{pc}^{-3}$ (Lauer 1989 and references therein). Another possible limitation is feedback of some of this binding energy into internal degrees of freedom of the stars. Since the binding energy of each star to the hole exceeds its internal binding energy by a factor $\sim (v_{\mathrm{orb}}/v_{\mathrm{esc}})^2 \gg 1$, only a small fraction of the dissipated energy would need to be deposited internally to disrupt the star before it becomes aligned with the disk plane. One way this could happen would be through excitation of oscillations inside the star as it successively smacks into each slab of disk material, which could conceivably build up faster than they could damp out. Since there is no reason for the oscillations to be in any way in phase with the impulses imparted by subsequent collisions with the disk, the degree of excitation will most likely be a random walk process. Whether stars would be disrupted by this process before their orbits are aligned with the disk is an important question whose answer depends on where within the star the modes are excited and damped. The calculation of this effect should be amenable to an analytic treatment and seems worthy of further investigation.

ACKNOWLEDGEMENTS

I am grateful to Roger Blandford and Melvyn Davies for motivational discussions and for suggestions on improving the manuscript. This work was supported by NSF grant AST92-23370.

REFERENCES

Artymowicz, P. 1994, ApJ, 423, 581

Artymowicz, P., Lin, D. N. C., & Wampler, E. J. 1993, ApJ, 409, 592

Bahcall, J. N., & Wolf, R. A. 1976, ApJ, 209, 214

David, L. P, Durisen, R. H., & Cohn, H. N. 1987a, ApJ, 313, 556

———. 1987b, ApJ, 316, 505

Davies, M. B. 1993, private communication

Duncan, M. J., & Shapiro, S. L. 1983, ApJ, 268, 565

Karas, V., & Vokrouhlický, D., 1993, MNRAS, 265, 365

Kippenhahn, R., & Weigert, A. 1990, Stellar Structure and Evolution. Springer-Verlag, Berlin

Kormendy, J. 1994, in Harris A., Genzel, R., eds, The Nuclei of Normal Galaxies: Lessons from the Galactic Center

Lauer, T. R. 1989, in Merritt, D., ed, Dynamics of Dense Stellar Systems. Cambridge University Press, Cambridge, p. 3

Lauer, T. R., Faber, S. M., Currie, D. G., & Ewald, S. P. 1992b, AJ, 104, 552

Lauer, T. R., Faber, S. M., Lynds, C. R., & Baum, W. A. 1992a, AJ, 103, 703

Lee, M. H., & Goodman, J. 1989, ApJ, 343, 594

Lugger, P. M., Cohn, H. N., Cederblom, S. E., & Lauer, T. R. 1992, AJ, 104, 83

Misner, C. W., Thorne, K. S., & Wheeler, J. A. 1973, Gravitation. W. H. Freeman Co., San Francisco

Murphy, B. W., Cohn, H. N., & Durisen, R. H. 1991, 370, 60

Ostriker, J. P. 1983, ApJ, 273, 99

Quinlan, G. D., & Shapiro, S. L. 1990, ApJ, 356, 483

Rauch, K. P., & Blandford, R. D. 1994, ApJ, 421, 46

Syer, D., Clarke, C. J., & Rees, M. J. 1991, MNRAS, 250, 505

Young, P. 1980, ApJ, 242, 1232

5

Collisional Stellar Dynamics Around Massive Black Holes in Active Galactic Nuclei†

Kevin P. Rauch

California Institute of Technology

ABSTRACT

Simulations of the dynamical evolution of the density cusp of a dense star cluster around a massive black hole are performed in a regime where stellar collisions dominate other dynamical processes and the black hole dominates the gravitational potential. The calculations are done fully relativistically using a cluster of $\sim 10^4$ discrete stars on geodesic orbits computed from the equations of motion of the Kerr (rotating black hole) metric, allowing relativistic effects to be accurately included. Numerical methods are developed to directly solve for collisions between geodesic orbits, without approximation, over part of the cluster and to statistically select collisions in an adaptive, self-consistent manner over the remainder of the cluster; useful by-products of these routines include an exact, compact procedure for determining the constants of motion from the orbital elements for bound Kerr orbits and a modified form of Kepler's Equation asymptotically valid in the Kerr geometry. Collision outcomes are determined using fitting formulae derived from an extended series of smooth particle hydrodynamics simulations of high velocity stellar impacts performed independently of this analysis. It is found that collisions produce a constant density inner core in the collisionally dominated region of the cusp which is mainly populated by stars on highly radial orbits. This is

† To be submitted to The Astrophysical Journal.

flatter than the $r^{-1/2}$ law seen in Fokker-Planck calculations; the origin of the discrepancy is not yet understood. Many grazing collisions are found to produce relatively little fractional mass loss even when a head-on collision would result in complete disruption, and a significant low mass tail consisting of post-collision stellar fragments develops in the stellar mass distribution. Collisional refilling of the loss cone is seen. The versatility of the simulation techniques makes them usable in a range of problems.

Subject headings: stellar dynamics – galaxies: nuclei – black hole physics

5.1. INTRODUCTION

Current observations continue to add support to the hypothesis, now over a quarter century old (Lynden-Bell 1969; cf. Hoyle et al. 1964), that many (if not all) galactic nuclei contain massive black holes (MBHs) at their centers. Recent HST observations of M87, for example, give strong evidence of a central mass of $\sim 3 \times 10^9 M_\odot$ (Harms et al. 1994). Observations of M31 (Richstone, Bower, & Dressler 1990; Bacon et al. 1994), M32 (Richstone et al. 1990; Lauer et al. 1992) and the Galactic Center (Genzel, Hollenbach, & Townes 1994), are consistent with central MBHs in the mass range $\sim 3 \times 10^7 M_\odot$ in M31 to $\sim 10^6 M_\odot$ for the Galactic Center. In addition, there are strong theoretical arguments in favor of this postulate. The leading explanation of energy production in active galactic nuclei (AGNs), for instance, involves accretion of gas (usually in the form of an accretion disk) onto MBHs having masses in the range $\sim 10^6 - 10^9 M_\odot$, and essentially all plausible alternatives, such as supermassive stars or dense starbursts, are almost guaranteed to produce an MBH during the course of their evolution if one is not already there (e.g., Rees 1984). Additional support can be found in the dense ($\rho \sim 10^6 - 10^8 M_\odot$ pc^{-3}) nuclear star clusters that have been directly observed in several nearby galaxies, including the Milky way (Lauer 1989 and references therein; Genzel et al. 1994). Besides the dynamical evidence for MBHs that observations of their central velocity dispersions and light profiles have already provided (see Kormendy 1994 for a review), numerical simulations of the dynamical evolution of these clusters themselves indicate that they are probably unstable to the formation of moderate mass ($\sim 10^3 M_\odot$) black holes, which would subsequently grow to $\sim 10^6 - 10^8 M_\odot$ through continued consumption of stars from the surrounding cluster (Quinlan & Shapiro 1990). In spite of all this, however, there is still no definitive proof that the central mass concentrations in galactic nuclei of the central engines of AGNs must be MBHs. Thus there is impetus from both sides of the debate to determine the probable influence of MBHs in galactic nuclei (active or not), either as a physical exploration because their existence is quite

likely, or, by uncovering unique, observable black hole signatures, as a possible way to help prove (or disprove) their existence.

From a physical standpoint, the existence of dense star clusters in galactic nuclei makes it a near certainty that collisions between stars will play an important role at some point during the cluster's dynamical evolution, at least at small radii, since both relaxation processes (which lead to eventual gravothermal collapse of the core) and central MBHs will tend to drive the inner regions towards higher densities and velocity dispersions until collisional processes become significant. Historically, before accretion onto a black hole came into favor as a way to power AGNs, several authors examined the evolution of nuclear star clusters (without black holes) in which, after a period of relaxation, stellar collisions became the dominant dynamical process, in part to see if the energy released in the collisions themselves could explain the luminosity of quasars (Spitzer & Saslaw 1966; Spitzer & Stone 1967; Colgate 1967; Sanders 1970). Spitzer & Saslaw (1966) and Spitzer & Stone (1967) studied clusters with high velocity dispersions ($v_{\rm rms} \gtrsim 10^3$ km s^{-1}) in which collisions were presumed to be disruptive, using a simple geometric model of mass loss from stellar collisions to compute the mass liberation rate, finding a peak luminosity of $\sim 10^{45}$ erg s^{-1} for a cluster of 10^8 stars with an initial radius of 0.05 pc. Colgate (1967) pointed out that when collisions first become important, they would occur at low enough relative velocity that coalescence would occur instead of disruption, and argued that the resulting buildup of massive stars, which would subsequently lead to a high rate of supernovae, could generate enough luminosity to explain energetic AGNs; a similar study by Sanders (1970) reached the opposite conclusion. A common feature of these studies is that they all used simple, qualitative prescriptions to calculate the outcome of a collision. The conflicting result of Colgate and Sanders, most likely resulting from the somewhat different prescriptions each used, highlights the uncertainty such an approach can introduce.

On the technological side, the continuing growth in computer power is allowing increasingly realistic simulations of both the dynamical evolution of galactic

nuclei and of the hydrodynamics of stellar collisions to be performed. The more recent studies of the evolution of galactic nuclei (Duncan & Shapiro 1983; David, Durisen, & Cohn 1987a,b; Quinlan & Shapiro 1990; Murphy, Cohn, & Durisen 1991) have used the Fokker-Planck method including progressively larger numbers of dynamical processes to examine cluster evolution, taking into account such processes as two-body relaxation, tidal disruption, stellar evolution and mass loss, loss cone effects, binary formation and heating, and stellar collisions. Although these studies represent major advances over previous research, they do have limitations. We note for one that it is assumed in these studies that the cluster can at all times be described by an isotropic distribution function $f(\mathcal{E})$; additionally, since by nature the Fokker-Planck approach is a statistical one, the above processes can only be included in an average sense. These studies have used the same or similar prescriptions for determining the outcomes of stellar collisions as used in the early studies, leaving them subject to the same uncertainties; advances in hydrodynamics simulations, however, and in particular the development and application of smooth particle hydrodynamics (SPH) techniques to the problem of stellar collisions (Benz & Hills 1987; Davies, Benz, & Hills 1991, 1992; Lai, Rasio, & Shapiro 1993), promise to greatly improve this shortcoming.

The purpose of this paper is to examine the dynamical evolution of the inner core of a dense star cluster around a massive black hole in the regime that stellar collisions dominate the evolution and the black hole dominates the potential, using a different (and more versatile) approach than has been used previously. Specifically, the cluster is modeled using a discrete collection of stars whose orbits are accurately traced using the geodesic equations of the Kerr (rotating black hole) metric. The intention is not only to compute the evolution of cluster properties such as the density profile, orbital distributions, and collisional mass liberation rate, but also to determine the importance of relativistic effects on the evolution, in particular as regards rotating (Kerr) against non-rotating (Schwarzschild) black holes. In the evolution of twisted accretion disks (Bardeen & Petterson 1975; Kumar & Pringle 1985; Pringle 1992), for instance, Lense-Thirring precession as-

sociated with Kerr black holes causes the inner portions of the disk to align with the equatorial plane of the hole (the Bardeen-Petterson effect). It is conceivable that the inherent anisotropy of the Kerr spacetime could manifest itself in the present case as well, since we will be considering the innermost regions of the cluster for which relativistic effects on orbits are noticeable. Additional discussion of Kerr effects is given in §5.2.2. The use of discrete stars also makes it easy to include a stellar mass spectrum and to follow its evolution, which could be substantial (Lee & Nelson 1988). Since the use of individual stars does restrict the volume around the hole that can be treated, due to the limited number of stars that can be included ($\sim 10^4$ in these simulations), the approach adopted here can be considered complementary to Fokker-Planck analyses, which are appropriate in the opposite limit of large volumes (where difficult-to-include relativistic effects are much weaker) and numbers of particles (which creates a well-populated phase space distribution).

As explained in §5.3, collisions between stars are found using a hybrid method in which collisions between the innermost ($\sim$ 500) cluster stars are solved for directly, without approximation, with collisions between the remaining ($\sim 10^4$) stars being randomly selected to occur at a dynamically computed rate that takes account of the current conditions in the cluster. This approach offers several advantages. First is its ability to allow the stars closest to the black hole, which are affected most strongly by relativistic effects, to be treated with a minimum of approximations, which could mask or bias subtle relativistic influences. Second, since finding collision events explicitly (especially between general quasi-elliptical Kerr geodesics, as is being done here) is expensive, the use of a zone of randomly generated collisions allows a much larger total volume and number of stars to be followed at minor additional cost. It is also versatile in that each individual strategy has enough flexibility to be used in several different ways, the simulations presented here representing only an initial application. The collision finding algorithm, for instance, can, with minor revision, be used to evolve a cluster undergoing two-body relaxation (gravitational 'collisions') in the vicinity of a Kerr

black hole. Another application is to sticky-particle type simulations in a Kerr environment, which could be useful for investigating the evolution of relativistic disks, for example. The adaptive behavior of the routine which randomly selects collisions makes it usable even when the distribution is changing rapidly or is significantly anisotropic. A modified form of Kepler's Equation asymptotically valid in the Kerr metric and a compact, exact procedure for computing the constants of motion from the orbital elements for an arbitrary quasi-elliptical Kerr orbit, derived in Appendix A (both are used by the collision finding routine), are also generically useful.

To alleviate as much as possible the uncertainties alluded to previously, the outcome of each individual collision (characterized by the mass loss and vector change in velocity suffered by each star), however it is found, is computed using fitting formulae derived from the results of a detailed set of SPH simulations of high velocity stellar collisions prepared independently of this work (Davies 1994).

The remainder of this paper is organized as follows. In §5.2, the basic dynamical processes relevant to the simulations are briefly reviewed; some discussion of the properties of Kerr black holes is also given. Section 5.3 explains in detail how the simulations were performed as well as the assumptions and simplifications made. Results and discussion are given in §5.4, and conclusions are offered in §5.5. Finally, Appendix A derives the modified Kepler's Equation, and Appendix B lists the SPH fitting formulae and gives related discussion.

Except where units are specifically given, this paper uses dimensionless variables defined by taking $G = c = M = 1$, where M is the mass of the black hole. Hence distances are measured in units of GM/c^2, time in units of GM/c^3, and velocities in units of c.

5.2. STELLAR DYNAMICAL PROCESSES IN AGNS

5.2.1 *Dynamical Influences on Cluster Evolution*

The physical processes which in principle can affect the evolution of a star cluster include two-body relaxation, tidal (i.e., loss cone) effects (if the cluster

contains a black hole), stellar collisions and mergers, stellar evolution and mass loss (supernovae), star formation, the creation of binaries and heating through binary encounters, and in the case of a cluster in an AGN, possibly also interactions with an accretion disk. For highly relativistic clusters energy losses due to gravitational radiation can also be important (e.g., Lee 1993; Quinlan & Shapiro 1989), although this is mainly relevant for clusters composed of compact objects such as neutron stars or stellar mass black holes, not main sequence stars. In the cluster cores treated in these simulations, which extend to distances of $r_{\rm sc} \sim 10^2 - 10^4$ from the MBH depending on the model, the typical velocity dispersion is quite high, $v_{\rm rms} \sim v_{\rm orb} \gtrsim r_{\rm sc}^{-1/2} \sim 0.03 \approx 10,000$ km s^{-1} (the choice of notation $r_{\rm sc}$ will become clear in §5.3), much larger than the surface escape speed of solar type stars, $v_{\rm esc} \sim 600$ km s^{-1}. Because of this, binaries and relaxation effects will be unimportant compared to collisions in determining the local dynamical equilibrium, and collisions will lead to eventual disruption of stars instead of mergers. As explained in §5.3.5, in some simulations relaxation is included in a limited way to account for injection of new stars into the inner core from the exterior regions of the cluster, but it is otherwise neglected. The short evolution times of most of these cores ($\sim 10^6 - 10^8$ yr) should make stellar evolution processes of little importance, especially since the more massive, quickly evolving stars also tend to be the first to collide due to their larger geometric cross-sections, and it was not included here; star formation is similarly neglected.

If the AGN contains an accretion disk, it will affect the evolution of the cluster through star-disk interactions (Syer, Clarke, & Rees 1991; Karas & Vokrouhlický 1993; Artymowicz, Lin, & Wampler 1993; Rauch 1994). The main effects are circularization of the orbits and alignment with the plane of the disk due to the drag force suffered by the stars as they hit the disk twice per orbit. It was originally hoped to include in the present paper, in the manner of Rauch (1994), simulations containing an accretion disk; preliminary calculations, however, indicated that results might be sensitive to how star-disk collisions are treated, and to allow for

a more careful treatment, this class of simulations will therefore be explored in an upcoming companion paper.

If evolution is dominated by relaxation, the effect of a central black hole (treated as a Newtonian point mass) on an isotropic cluster is to create a density cusp $\rho \propto r^{-7/4}$ within the hole's sphere of influence $r_{\rm h} \sim v_{\rm rms}^{-2}$ (Bahcall & Wolf 1976). However, the existence of black hole horizons, and, for holes $\lesssim 10^9 M_\odot$, a tidal radius, creates a loss cone which will modify (flatten) this cusp at small radii (Frank & Rees 1976; Lightman & Shapiro 1977), due to tidal disruption or direct consumption of stars entering the loss cone, which occurs on a dynamical timescale. Collisions will further modify the cusp solution, and one of the results of these calculations will be a quantitative estimate of this effect in the vicinity of the MBH. Hills (1975) suggested tidal mass loss and subsequent accretion of gas onto the black hole as a possible way to power AGNs. It is now generally agreed that this mechanism liberates insufficient amounts of gas to account for the more powerful AGNs unless collisional processes predominate. By incorporating the results of SPH simulations of colliding stars into the analysis, one aim of this paper is to realistically estimate the mass liberation rate produced by these collisions. Tidal mass loss is also accounted for in the calculations, but as shown in §5.4 this mass loss rate is significantly below the amount lost in collisions.

To summarize, in all calculations presented here disruptive collisions dominate the dynamical evolution of the cluster, and relaxation enters in a very limited way. Although expected to be of secondary importance, tidal effects are also included as the cost of doing so is trivial in our approach. To determine the range of cluster densities and radii over which these assertions are valid, we now compare the relative importance of these three processes more carefully. To this end, it is useful to define dimensionless, orbit-averaged collision, relaxation, and tidal depths defined by $\tau_i = \oint dt(r)/t_i(r)$, where the corresponding timescale for process i is t_i and the integral is over a complete orbital cycle of the particular star under consideration. Thus, for instance, $1/\tau_{\rm coll}$ is the average number of orbits traversed between collisions, $t_{\rm orb}/\tau_{\rm r}$ is the orbit-averaged relaxation time, and process i

dominates process j whenever $\tau_i \gg \tau_j$. The respective timescales are given by $t_{\rm coll} \sim (n_\star \sigma v_{\rm rel})^{-1}$, $t_{\rm r} \sim v_{\rm orb}^3/[15G^2 m_\star^2 n_\star \ln(0.4N_\star)]$, and $t_{\rm t} \sim (1-\epsilon^2)t_{\rm r}$, where $n_\star$ is the number density of stars, $\sigma \approx 4\pi R_\star^2$ is the collision cross-section (neglecting gravitational focusing, since $v_{\rm rel} \gg v_{\rm esc}$), $v_{\rm rel}$ ($\approx \sqrt{2}\, v_{\rm orb}$ for an isotropic cluster) is the typical relative velocity between stars, $v_{\rm orb} = [1/r - 1/(2a)]^{1/2}$ is the star's orbital speed, and $N_\star$ is the total number of stars in the cluster. The semi-major axis and eccentricity of the orbit are a and ϵ. The tidal timescale $t_{\rm t}$ is the typical time needed for the orbit to diffuse into the loss cone (assuming the loss cone orbit has $\epsilon \approx 1$). Note that $\tau_{\rm t}$ for an orbit with $a(1-\epsilon) \sim r_{\rm t}$ (where $r_{\rm t} \approx 2R_\star(M/m_\star)^{1/3}$ is the tidal radius) is essentially the function '$q(E)$' of Lightman & Shapiro (1977), defined as the ratio of the mean square change in angular momentum per orbital period due to relaxation to the width of the loss cone itself; thus $1/(1+\tau_{\rm t})$ is approximately the fraction of orbits lying near the loss cone which enter it (and disrupt) each orbital period. Assuming $n_\star = 10^6 n_6 (r/1\ {\rm pc})^{-7/4}\ {\rm pc}^{-3}$ and $M = 10^8 M_8\, M_\odot$, the numerical values of these depths are approximately

$$\begin{aligned}
\tau_{\rm coll} &\approx 5\times 10^{-4}\, n_6 M_8^{-3/4} \left(\frac{R_\star}{R_\odot}\right)^2 \left[a(1-\epsilon)\right]^{-3/4}, \\
\tau_{\rm r} &\approx 4\times 10^{-15}\, n_6 M_8^{-3/4} \left(\frac{m_\star}{M_\odot}\right)^2 \left[\frac{(1+\epsilon)^{5/4}(5-3\epsilon)}{(1-\epsilon)}\right] a^{5/4}, \qquad (5.1) \\
\tau_{\rm t} &\approx 4\times 10^{-15}\, n_6 M_8^{-3/4} \left(\frac{m_\star}{M_\odot}\right)^2 \left[\frac{(1+\epsilon)^{1/4}(5-3\epsilon)}{(1-\epsilon)^2}\right] a^{5/4}.
\end{aligned}$$

Note that a is still in units of M; in physical units, $M = 5\times 10^{-6} M_8$ pc. The condition $\tau_{\rm coll} \gtrsim \tau_{\rm r}$ implies $a \lesssim 10^5$ (assuming solar type stars), almost independent of ϵ; it follows that collisions are more important than relaxation in determining the distribution function at radii $r \lesssim 10^5 = 0.5 M_8$ pc, independent of the absolute density. Similarly, tidal effects are more important than relaxation for $r \lesssim r_{\rm crit}$, where $\tau_{\rm t} \sim 1$ for $r_{\rm crit}(1-\epsilon) \sim r_{\rm t}$, which implies $r_{\rm crit} \sim 5\times 10^4\, n_6^{4/13} M_8^{-7/39}$ (again assuming solar type stars). To complete the triad, the condition for collisions to dominate over tidal effects, $\tau_{\rm coll} > \tau_{\rm t}$, implies $a \lesssim 2000\, r_{\rm min}^{5/13}$, where $r_{\rm min} = a(1-\epsilon)$. Hence in the region where collisions dominate over relaxation, they also dominate

loss cone effects, except for a class of high eccentricity, large a orbits that violate the last inequality. Motivated by these considerations, the cluster initial conditions were chosen to be a truncated $r^{-7/4}$ power law with $a_{\rm min} \sim r_{\rm t}$ and $a_{\rm max} \sim 10^5$ (see §5.3.1).

The numerical values for the orbit-averaged timescales, $t_i = t_{\rm orb}/\tau_i$, are

$$
\begin{aligned}
t_{\rm coll} &\approx 0.2\, n_6^{-1} M_8^{7/4} \left(\frac{R_\star}{R_\odot}\right)^{-2} (1-\epsilon)^{-3/2} r_{\rm min}^{9/4} \ \ {\rm yr}, \\
t_{\rm r} &\approx 2\times 10^{10}\, n_6^{-1} M_8^{7/4} \left(\frac{m_\star}{M_\odot}\right)^{-2} \left[\frac{(1-\epsilon)^{3/4}}{(1+\epsilon)^{5/4}(5-3\epsilon)}\right] r_{\rm min}^{1/4} \ \ {\rm yr}, \qquad (5.2) \\
t_{\rm t} &\approx 2\times 10^{10}\, n_6^{-1} M_8^{7/4} \left(\frac{m_\star}{M_\odot}\right)^{-2} \left[\frac{(1-\epsilon)^{7/4}}{(1+\epsilon)^{1/4}(5-3\epsilon)}\right] r_{\rm min}^{1/4} \ \ {\rm yr}.
\end{aligned}
$$

These specific relations are only valid within the $(r^{-7/4})$ cusp, which extends to a radius $r_{\rm h} \sim 1\times 10^6 n_6^{-4/5} M_8^{-1/5} \approx 5\, n_6^{-4/5} M_8^{4/5}$ pc, outside of which the cluster, not the black hole, dominates the potential. Table 5.1 lists $r_{\rm h}$ and the maximum radii for which each timescale is shorter than a Hubble time, $\sim 10^{10}$ yr, for several hole masses, assuming solar stars and $\epsilon^2 = 0.5$ (the mean value for an isotropic cluster). Astrophysically the mass $M = 10^6 M_\odot$ could correspond to a modest Seyfert galaxy or to the Galactic Center, $M = 3\times 10^7 M_\odot$ to a bright Seyfert or M31, and $M = 10^9 M_\odot$ to a typical quasar or a nucleus such as M87's. Note that collisions are always important in the cusp and that relaxation processes are negligible for nuclei with very massive black holes. Since the $r^{-7/4}$ cusp is created on a relaxation timescale, the last point suggests that a well-formed cusp has not had time to develop in these nuclei; however, initial adiabatic growth of the black hole, which occurs on timescales much shorter than $t_{\rm r}$, creates a similar cusp, between $r^{-3/2}$ for isotropic orbits to $r^{-9/4}$ for purely circular orbits (Young 1980). The $r^{-7/4}$ form should therefore provide a reasonable description of the cusp even in this case.

5.2.2 *Effects of Kerr Black Holes*

One feature of the collision code is its use of a fully relativistic formalism for the computation of stellar orbital trajectories and collision center-of-mass frames,

TABLE 5.1

Maximum Radii for Dynamical Evolution Within a Hubble Time

	$n_6^{4/5} r_{\rm h}$		$n_6^{-4/9} r_{\rm coll}$		$n_6^{-4} r_{\rm r}$		$n_6^{-4} r_{\rm t}$	
M	(M)	(pc)	(M)	(pc)	(M)	(pc)	(M)	(pc)
$1 \times 10^6 M_\odot$	3×10^6	0.2	9×10^5	0.05	$> r_{\rm h}$	$> r_{\rm h}$	$> r_{\rm h}$	$> r_{\rm h}$
$3 \times 10^7 M_\odot$	2×10^6	2	7×10^4	0.1	$> r_{\rm h}$	$> r_{\rm h}$	$> r_{\rm h}$	$> r_{\rm h}$
$1 \times 10^9 M_\odot$	1×10^6	40	4×10^3	0.2	< 1	$\cdots$	< 1	$\cdots$

using the Kerr metric to model the spacetime environment. This is the appropriate model for astrophysical black holes where they dominate the gravitational potential, as in these simulations. This section gives a short review of the physics of Kerr black holes. For a detailed review of black holes in an AGN environment, see Blandford (1987).

In general relativity, all black holes are completely specified by three parameters (e.g., Misner, Thorne, & Wheeler 1973), their mass M, angular momentum J, and charge Q, with the additional requirement that $Q^2 + (J/M)^2 \leq M^2$ (violation of this constraint results in a naked singularity, not a black hole). In astrophysical contexts, Q is essentially guaranteed to be unimportant dynamically, as even a small relative excess of charge would be quickly quenched by accretion of charged particles due to the greatly mismatched strengths of gravity and electromagnetism. The angular momentum of the hole, however, can be substantial. (The remainder of this discussion will assume $Q = 0$ and define $j = J/M^2$.) This is particularly true for black holes in AGNs, which are expected to be spun up by accretion of matter (as from an accretion disk) to near maximal values, $j \sim 1$, regardless of their initial j (Bardeen 1970; Thorne 1974).

Both Kerr and Schwarzschild ($j = 0$) black holes exert qualitatively new effects on orbits compared to Newtonian gravity. In the Schwarzschild case spherical symmetry ensures that orbits still lie in a plane, but due to precession bound or-

bits no longer close on themselves; asymptotically far from the hole this precession amounts to $6\pi/[a(1-\epsilon^2)]$ rad per orbit (cf. eq. (5.17)). In the Kerr case the angular momentum introduces a preferred direction, and in general orbits no longer lie in a plane, although for distant orbits the motion can be well approximated by a slowly rotating planar orbit. Physically this non-planarity (or precession of the effective orbital plane about the spin axis of the hole) is the result of Lense-Thirring precession. In an AGN environment Lense-Thirring precession can have important consequences, such as the Bardeen-Petterson effect, as mentioned in §5.1. Another qualitatively new feature is the creation of a region of unstable orbits. In the Schwarzschild case circular orbits inside $6M$ are unstable (and do not exist at all inside $3M$), and small inward perturbations will cause them to spiral into the hole; additionally no quasi-elliptical orbits exist with $r_{\rm min} < 4M$. For Kerr holes these radii depend on both j and the orbital inclination, but qualitatively speaking, retrograde orbits become unstable at larger radii and prograde orbits are stable to smaller radii. A dynamical consequence of this is the creation of a loss cone even for objects which cannot be tidally disrupted by the hole. A final new property we will mention is the presence of negative energy orbits around Kerr holes, which exist only within the ergosphere (the region near the horizon in which frame-dragging forces all trajectories to rotate in the sense of the hole); the significance of these orbits is explained below.

In a dynamical sense, Schwarzschild black holes are 'dead'—their potentials can power AGNs through release of binding energy of inspiraling matter (up to $\sim 6\%$ of the rest mass), but their own rest-mass energy cannot be released. This is not the case for Kerr holes, whose spin energy can be extracted, up to a point, by several mechanisms (decreasing M in the process). One of these is the Penrose process (Penrose 1969), in which a particle on a negative energy orbit can cross the horizon and reduce the hole's mass. Energy can also be extracted electromagnetically (Blandford & Znajek 1977). In the calculations presented here, however, such effects have little chance to operate, and in terms of possible influences on cluster evolution, the salient relativistic effects are precession and possibly loss cone

effects; collision dynamics might also be influenced by the anisotropy of the Kerr spacetime (frame-dragging effects). Note that optical observations can only resolve distances $\sim v_{\rm obs}^{-2} \sim 1 \times 10^5 (v_{\rm obs}/10^3 \text{ km s}^{-1})^{-2} \sim 1 M_8 (v_{\rm obs}/10^3 \text{ km s}^{-1})^{-2}$ pc, on which scales these effects will be quite small, and hence that direct observation of these effects is implausible.

5.3. NUMERICAL METHOD

Each stellar cluster considered in the numerical simulations was separated into four distinct radial regions. The innermost zone, which will be termed the tidal region, normally contains few (if any) stars and delineates the volume around the black hole within which stars will be tidally disrupted should they enter. Immediately outside this zone lie the innermost stars of the cluster itself, in what will be called the 'collision finder' (CF) region; in this zone, physical collisions between the stars were found by advancing each star whose orbit crossed the CF region along its respective geodesic trajectory and searching for close encounters between all such pairs of stars whose orbits could possibly cross. Outside the CF region (which at any given time extends from $r_{\rm t}$ out to a radius to be called $r_{\rm CF}$) lies the 'statistical collisions' (SC) region, in which individual collisions were not solved for iteratively but were made to occur, by suitably altering the orbital phases, at a rate equal to that expected from each star's orbit-averaged collision timescale. This zone extends out to some fixed radius $r_{\rm SC}$; typically $r_{\rm SC}/r_{\rm CF} \sim 5 - 10$. Finally, the outermost region (not present in all simulations), labeled the 'reservoir,' consisted of all other stars in the cluster. The dynamical evolution of this region was not considered in the simulations and its only purpose, when applied, was to feed new stars into the inner regions to allow a steady state within the inner zones to be reached. The details of how each of these regions was dealt with, as well as the zone-independent assumptions employed, are explained in the following subsections.

5.3.1. *Fundamental Parameters and Initial Conditions*

Each simulation is characterized by specific values or treatment of the following elements:

a) the mass M and angular momentum j of the black hole,

b) the initial distribution function of the cluster, $f(\mathcal{E}, L, L_z)$,

c) the input stellar mass spectrum,

d) the number of stars in the CF (§5.3.3) and SC (§5.3.4) zones, and

e) the presence or absence of the outer cluster reservoir (§5.3.5).

One additional element, which was treated the same way in all simulations, is the determination of the outcome of a collision, for which the results of SPH simulations of colliding stars, performed independently of this work, were used (Davies 1994); for details, see Appendix B.

Hole masses in the simulations ranged from $\sim 10^6$ to $\sim 10^9 M_\odot$, roughly covering the range suggested by observations of galactic nuclei (see §5.1 for references). Most simulations used $j = 1$, for which any precession-induced effects would be maximized; to judge the significance of the black hole's spin on the results, a few runs were repeated using $j = 0$.

In the Newtonian point-mass potential $\Psi = 1/r$, all five orbital elements—the semi-major axis, a, eccentricity, ϵ, inclination, i, argument of periapse, ω, and argument of the ascending node, Υ—are constants of the motion. The correspondence to energy and angular momentum is $\mathcal{E} = 1/(2a)$, $L^2 = a(1-\epsilon^2)$, and $L_z = L\cos i$. We are using the conventions that $\mathcal{E} > 0$ is the binding energy and that $0 \leq i \leq \pi$, where $i = 0$ corresponds to a prograde orbit in the equatorial plane of the hole and $i = \pi$ to a retrograde orbit in that plane. In the Kerr spacetime geometry, geodesics possess only three constants of motion, which for quasi-elliptical orbits can be equated with a, ϵ, and i—or, (almost) equivalently, $\mathcal{E}, L$, and L_z—with precession effects leading to non-conservation of ω and Υ. (Strictly speaking, L is not conserved for Kerr orbits; cf. Appendix A.) The simulation code was equipped to create clusters with a distribution function (in an externally imposed

Newtonian potential $\Psi = 1/r$) of the form

$$f(\mathcal{E}, L, L_z) = \begin{cases} K\mathcal{E}^{n-3/2}\, e^{pL^2/L^2_{\max}(\mathcal{E})}\, e^{qL_z/L}, & \mathcal{E}_{\min} \le \mathcal{E} \le \mathcal{E}_{\max}\,; \\ 0, & \text{otherwise}, \end{cases} \tag{5.3}$$

where $L^2_{\max}(\mathcal{E}) = 1/(2\mathcal{E})$ is the maximum possible angular momentum for an orbit of energy $\mathcal{E}$ and the normalization,

$$K = \frac{(n-3)pqN_\star}{\sqrt{2}\,\pi^3(e^p - 1)\sinh q\,(\mathcal{E}^{n-3}_{\max} - \mathcal{E}^{n-3}_{\min})}, \tag{5.4}$$

is such that the total number of stars in the cluster is $\int f\, d^3r\, d^3v = N_\star$. As explained below, p and q are anisotropy parameters for velocity and real spaces, respectively, and n is the power law index of the radial density profile.

The value of K can be easily derived as follows. Create polar coordinates in both real space, (r, θ, ϕ), and velocity space, (v, η, ξ). It is convenient to let $\cos\eta = \vec{r}\cdot\vec{v}$ and ξ be the corresponding azimuthal angle. By performing the coordinate transformation $(r, \theta, \phi, v, \eta, \xi) \rightarrow (\mathcal{E}, L, L_z, v, \phi, \xi)$, noting that $\mathcal{E} = 1/r - v^2/2$, $L = rv\sin\eta$, and $L_z = rv\sin\theta\sin\eta\sin\xi$, and then integrating over $dv\, d\phi\, d\xi$, one finds $\int d^3r\, d^3v \rightarrow (\sqrt{2}\,\pi)^3 \int \mathcal{E}^{-3/2} d\mathcal{E}\, dL\, dL_z$ (each integral being done over all allowed ranges of each variable), from which the result for K follows. Since $f(\mathcal{E}, L, L_z)/N_\star$ can also be thought of as the joint probability distribution for the orbital parameters a, ϵ, and i, it follows therefore that the probability distributions of a, ϵ, and i are independent and given by $P(a) = (3-n)a^{2-n}/[a^{3-n}_{\max} - a^{3-n}_{\min}]$ (where $2\,a_{\max} = 1/\mathcal{E}_{\min}$ and $2\,a_{\min} = 1/\mathcal{E}_{\max}$), $P(\epsilon) = 2p\epsilon e^{-p\epsilon^2}/(1 - e^{-p})$, and $P(i) = q\sin i\, e^{q\cos i}/(2\sinh q)$. By creating the appropriate number of orbits with elements drawn from these distributions, a cluster with the distribution function (5.3) will be produced. Note that although this formalism is entirely Newtonian, once the elements of an orbit were drawn, the orbit was created using a procedure appropriate for the Kerr metric; for details, see Appendix A.

The specific analytic form of eq. (5.3) was chosen for its simplicity, ease in producing the desired initial density profile, and for its ability to model the overall effects of a cluster evolving in the presence of an accretion disk (cf. §5.2.1). For

$p = q = 0$, $f = f(\mathcal{E})$ only and the cluster is isotropic in both real and velocity spaces. For $p \neq 0$ the velocity dispersion is anisotropic, with orbits becoming increasingly radial on average as p tends to large negative values and increasingly circular for large positive values of p; in real space the cluster remains spherically symmetric. For $q \neq 0$ symmetry in real space is broken, with orbits being preferentially retrograde for $q < 0$ and prograde for $q > 0$. To good approximation, the density $\rho(r, \theta) = \int f \, d^3v$ scales as $\rho \propto r^{-n}$ for $a_{\rm min} \leq r \leq a_{\rm max}$ (and fixed θ), independent of p and q. In fact, all simulations reported here used $q = 0$ for the initial cluster; the value of p ranged from 0 to 10. Previous theoretical investigations into the equilibrium distribution function of a star cluster around a black hole, each making slightly different assumptions, give values of n in the range $n \sim 3/2 - 9/4$ (Peebles 1972; Bahcall & Wolf 1976; Young 1980). All runs presented here used $n = 7/4$, the Bahcall-Wolf result, with $N_\star$ chosen so that $\rho(1\,{\rm pc}) \approx 10^5 - 10^7 \, M_\odot \, {\rm pc}^{-3}$, in line with observations of real (but dense) nuclear star clusters. As explained in §5.2.1, the values $a_{\rm min} = r_{\rm t}$ (the tidal radius) and $a_{\rm max} = 10^5$ (the approximate radius beyond which relaxation is more important than collision) were used.

Most simulations used initially equal mass stars, specifically $M = M_\odot$ or $M = 10 M_\odot$. When a range of initial masses was included, a Salpeter mass function with $0.4 M_\odot \leq M \leq 4 M_\odot$ was used for convenience, and all masses were assumed to follow the same distribution function.

5.3.2. *The Tidal Zone*

The tidal zone, which is normally devoid of stars (except for those few in the process of disruption), extends from the event horizon of the black hole out to the tidal radius of the cluster stars, and is the region within which a passing star will be tidally disrupted by the hole. Since the tidal radius $r_{\rm t} \approx 2 R_\star (M/m_\star)^{1/3} = 4.4 M_8^{-2/3} (m_\star/M_\odot)^{1/3}$ (assuming $R_\star \propto m_\star^{2/3}$) is mass-dependent, this region is of definite size only for clusters with equal mass stars. During the simulations, $r_{\rm t}$ was computed individually for each star, and whenever a collision placed the

surviving star on an orbit passing within its own $r_{\rm t}$, it was removed from the simulation and its mass added to the 'tidal disruption mass loss' total, $M_{\rm t}$. When setting up the cluster initially, if the randomly drawn orbit happened to pass inside its tidal limit, it was simply rejected and another orbit created until one not suffering this condition was obtained; to help keep this from occurring in the first place, $a_{\rm min} = 1/(2\mathcal{E}_{\rm max})$ (see the previous section) was normally taken to be the tidal limit of a typical mass star in that simulation. When adding stars from the reservoir (see §5.3.5) during the course of a simulation, however, such an orbit was not rejected but instead the associated mass was immediately added to $M_{\rm t}$, and no star was actually added to the inner cluster; this procedure allows the mass loss rate due to stars entering the tidal loss cone, $\dot{M}_{\rm t}$, to be realistically estimated and compared to the collisional mass liberation rate, $\dot{M}_{\rm coll}$. In most simulations, however, the tidal radius was small and close to the horizon, and in all cases it turned out that $\dot{M}_{\rm t} \ll \dot{M}_{\rm coll}$ (see §5.4). In simulations possessing a steady state, the ratio of the two mass loss rates should be roughly $\dot{M}_{\rm t}/\dot{M}_{\rm coll} \sim L_{\rm t}^2/L_{\rm sc}^2$ ($\ll 1$ for the specific parameters used in most simulations), i.e., the fraction of orbits inside $r_{\rm sc}$ which also cross $r_{\rm t}$, the L_i being defined such that orbits with $L < L_i$ pass inside r_i. Overall, therefore, the tidal zone was expected to play a small role in most of the simulations, and such was indeed found to be the case.

5.3.3. *The 'Collision Finder' Zone*

The CF zone is not so much a specific region of space as it is a region of orbital parameters. It consists merely of those $N_{\rm CF}$ stars with the smallest periapses, and extends from the inner boundary of the cluster (set by the tidal radius) out to a radius $r_{\rm CF}$ determined by $N_{\rm CF}$ and by the current conditions within the cluster. In this region actual collisions between the stars are solved for, without approximation. This is done by advancing each star in the CF region along its quasi-periodic geodesic trajectory (also computed exactly from the equations of motion in the Kerr geometry), orbit by orbit, and searching for close encounters between all overlapping pairs of orbits; the precise method used to accomplish

this is given in complete detail below. In an ideal universe containing unlimited computing power, one would set $N_{\rm CF} = N_\star$ and solve for the true collisions over the entire cluster; in the real universe and with current CPU technology, the algorithm used to sift through the orbits in search of collisions reached its practical limit at $N_{\rm CF} \sim 500$, for which runs covering physically worthwhile periods of time took ~ 1 week on a state-of-the-art desktop workstation. Although not large by N-body standards, this number did allow a significant portion of the innermost region of the cluster, the part which can be affected by relativistic effects most strongly, to be treated without approximations (excluding the uncertainties in the SPH fitting formulae), which could conceivably introduce bogus effects or mask real ones, as in the question of whether or not the anisotropy of the Kerr metric can act to leave post-collision stellar remnants in preferentially low-inclination orbits, for example. It is also important to remember that unlike Newtonian N-body simulations, in the present case the gravitational force is not $1/r^2$, coordinate time is not the computationally convenient independent variable, and in general the orbits being followed neither lie in planes nor close on themselves, all of which make the problem of searching for collisions considerably more challenging (and time-consuming). Since it was nonetheless desirable, and in some cases necessary, to extend the radial range and total number of stars present in the simulations, an intermediate region that could contain $\sim 10^4$ stars was created in which collisions were not solved for but were forced to occur at a self-consistent rate; the treatment of this region is described in §5.3.4.

In broad terms the functioning of the collision finder can be understood as follows. Imagine separating the continuous geodesic trajectory of each star into a series of segments, each segment beginning at one periapse and extending to the next (future) one; for orbits lying far from the black hole these segments are nearly ellipses and have a well-defined orbital plane, but in general neither of these properties will hold. Now consider finding all collisions (close encounters) between a specific (but arbitrary) pair of stars in the CF zone; obviously, only stars whose orbits overlap radially need be considered, for they cannot possibly

collide otherwise. Associated with each orbital segment is the corresponding time interval that the star lies within that particular segment, and it is likewise clear that only those orbital segments that overlap in time need be examined further for close encounters. Thus by considering in turn each successive orbital segment for one of the stars and finding for each those segments of the other star's orbit which overlap in time, an exhaustive search for collisions between that pair of stars will result; by repeating this procedure for all possible pairs of stars in the CF region, all collisions occurring within the cluster will be found, and this is the basic strategy used by the collision finding routine. What remains to be worked out is an efficient method for finding close encounters between any given pair of orbital segments, or more appropriately, since in fractional terms exceedingly few segments will actually have a collision point, for *rejecting* non-colliding orbital segments as quickly as possible. Although to some this strategy may seem rather awkward and roundabout, it manages to substantially avoid the computational problems suffered by conceptually simpler time-based algorithms. In time-based approaches, such as the step-by-step N-body integration of the evolution of a globular cluster, time is the natural variable of integration, and searching for collisions is in principle a near triviality, since at each time step the position and velocity of each particle are already known and close encounters (or nearest neighbors) can be hunted for directly. In the Kerr spacetime, however, a much more efficient solution to the geodesic equations can be obtained if radius (or, in principle, the polar angle) is used as the independent variable instead of time, since when this is done the equations of motion can be recast as definite integrals, some of which can be solved (and even inverted) analytically (for details, see Rauch & Blandford 1994). Once this is done, one can efficiently find the value of coordinate time given the radius, and it is not even necessary to integrate around each orbit to find the successive times of periapse passage, or in other words, since the geodesic solutions are space-based, whole series of orbits can be 'skipped' essentially for free, a property not shared by methods attempting solution in the form of a set of ordinary differential equations. The collision finding strategy was designed to

exploit the aforementioned properties of the space-based Kerr geodesic routines, making it well-suited to the problem at hand.

Synchronization presents a special problem. Since the orbital periods of individual stars can vary widely, keeping the orbital segments of one star in step with those of every other star requires careful treatment. This was dealt with by dynamically creating a stack frame containing, for each star, all segments of that star's orbit traversed between fixed starting and stopping times, the number of segments needed being directly proportional to the star's orbital period. All collisions occurring within the given time interval are then found by comparing corresponding pairs of segments, the starting and stopping times are incremented, and the entire process is then repeated.

Clearly, quite a large number of pairs of orbital segments, $\sim (t_{\rm max}/\bar{t}_{\rm orb})N_{\rm CF}^2$ ($\sim 10^{10}$ in the simulations), where $t_{\rm max}$ is the maximum time reached in the simulations and $\bar{t}_{\rm orb}$ is an average star's orbital period, need to be examined in the course of a run; hence it is crucial to be able to discard pairs without a collision point as quickly as possible. The heart of the collision finding algorithm lies in its application of a series of increasingly refined (and time-consuming) 'cuts' to the particular pair of segments, designed to reject as quickly as possible pairs which don't collide, without falsely rejecting pairs that do. As these are competing goals, a suitable balance between speed and accuracy needs to be achieved, and in practice acceptability was defined as a false rejection rate (the fraction of collisions not found by the algorithm) of $\lesssim 1\%$.

As the existence of time is not to be denied regardless of its computational liabilities, central to this endeavor was the development of a modified version of Kepler's Equation valid in the Kerr geometry, derived as an expansion in $\delta = 1/[a(1-\epsilon^2)]$ and valid to $O(\delta^{5/2})$. Although orbits in Kerr do not, in general, lie in a plane, the non-planarity of any single orbital segment will be small as long as the orbit does not pass too close to the hole. Thus by approximating single segments as lying in a mean orbital plane defined by the star's position and velocity vectors at apoapse (where relativistic effects for a given orbit are smallest) and applying

the modified Kepler's Equation, the approximate position of the star in that plane as a function of time can be economically computed. In this way the basic strategy one would use to seek a collision between two Newtonian elliptical orbits can be applied to the Kerr segments with (conceptually) only minor amendments. The modified Kepler's Equation and related formulae are derived in Appendix A.

In final detail the segment rejection algorithm works as follows. Having excluded all pairs of stars whose orbits do not radially overlap, the following cuts were applied to each pair of segments until the possibility of collision was excluded or, in the end, confirmed. First, the modified Kepler's Equation was used to find the time intervals during which each star resided in the radial overlap zone, in which all collisions must of course occur; if the time intervals did not overlap, the pair was rejected immediately. If this failed, use was made of the fact that each segment lies nearly in a plane. Were each orbit strictly planar, clearly they could collide only along the line of intersection of the two planes; for small deviations from planar motion, a collision point must lie somewhere *near* the line of intersection, unless the orbits are almost coplanar. Quantitative estimates of the maximum perpendicular deviation from the mean plane over a single orbit were made and used to reject segment pairs whose spatial separations at the line of intersection exceeded this limit; if a collision appeared spatially possible, the times at which the stars crossed the line of intersection were computed and exclusion was attempted on that basis. A complication over the Newtonian case is that due to general precession inside the mean orbital plane, one periapse-to-periapse orbit can cross the line of intersection more than twice, and possibly many times, each at a different radius, so that one must be careful to test all possible intersection points for overlap before rejecting a pair of orbits. Finally, if a close encounter was found by this procedure and the pair still could not be rejected, the full Kerr routines were used to find the exact point of closest approach of the orbits, using the approximate location as an initial estimate. If the minimum distance found was less than the sum of the stars' radii, a collision was flagged and sent to the collision processing routine (cf. Appendix B).

To verify that the collision finder was working properly, tests were done with full scale clusters ($N_{\rm CF} > 100$) similar to those used in the actual simulations in which numerous randomly selected pairs of stars were given initial conditions that put them on a collision course at a known time and place in the future, normally many orbital periods ahead. The false rejection rate was then found by counting the number of collisions missed by the collision finder; the accuracy of the recovered collision coordinates and minimum distance was also checked. In all tests the loss rate was $< 2\%$, with $\lesssim 1\%$ being the typical case, and the recovered collision parameters were accurate. Additionally, during the production runs histograms of the recovered minimum collision distances (i.e., impact parameters) were tallied and examined and the resulting distributions agreed with that expected from an equal-areas probability law, giving runtime assurance that no bias against the recovery of collisions with any particular impact parameter was being introduced. The rate at which collisions were being found was also examined and found to be consistent with estimates based on the average collision rate determined from the initial distribution function. A foreseen exception is with the outermost stars in this zone, most of whose collision probability is due to stars in the SC region, where these particular CF zone stars spend nearly all of their time. Qualitatively this should not alter the results, since the absolute number of collisions involved with these stars is small, and otherwise identical simulations redone with only a SC zone (which does not suffer from this problem) showed this to be the case.

5.3.4. *The 'Statistical Collisions' Zone*

Conceptually the SC zone is distinguished from the CF zone solely by its method of "finding" the collisions occurring between its $N_{\rm SC}$ stars; in practice this zone contained most of the stars included in a given simulation and spanned most of the included volume, these useful increases in size and number being the motivation for its introduction, as mentioned previously. Unlike the CF zone the SC region possesses a fixed radial boundary $r_{\rm sc}$, which was chosen to be the

largest radius such that all orbits (in the average sense) passing within $r_{\rm SC}$ could be physically included in the simulations.

The treatment of collisions in the SC region proceeded as follows. At the beginning of each iteration—one iteration being the time interval covered by a single CF stack frame (§5.3.3)—the orbit-averaged collision depth, $\tau_{\rm coll}$, was computed for each star in the SC zone from the expression

$$\begin{aligned}\tau_{\rm coll} &= 2\int_{a(1-\epsilon)}^{a(1+\epsilon)} \frac{dt(r)}{t_{\rm coll}(r)} = 2\int_{a(1-\epsilon)}^{a(1+\epsilon)} \frac{dr}{v_r(r)t_{\rm coll}(r)} \\ &= 2a^{3/2}\int_{1-\epsilon}^{1+\epsilon} \frac{\bar{n}_\star\big(ax,\theta(ax)\big)\,\langle\sigma_{\rm coll}(v_{\rm rel})v_{\rm rel}\rangle x dx}{[(1+\epsilon-x)(x-1+\epsilon)]^{1/2}},\end{aligned} \tag{5.5}$$

where $v_r^2(r) = [a(1+\epsilon)-r][r-a(1-\epsilon)]/(r^2 a)$ is the radial component of the star's orbital velocity, $\sigma_{\rm coll}(v_{\rm rel}) = \pi(R_\star + R)^2(1 + v_{\rm esc}^2/v_{\rm rel}^2)$ is the cross section for the given star (of radius $R_\star$ and surface escape velocity $v_{\rm esc}$) to collide with one of radius R at relative velocity $v_{\rm rel}$, and $\bar{n}_\star(r,\theta)\langle\sigma_{\rm coll}(v_{\rm rel})v_{\rm rel}\rangle$ is the local collision rate for *that particular* star's orbit, averaged over the current cluster distribution function. Note that $\tau_{\rm coll}$ implicitly depends on both the orbital elements of the star and the distribution function of the cluster.

The determination of $\bar{n}_\star$ and the averaging procedure require explanation. Since the integral for $\tau_{\rm coll}$ was being computed every iteration for each of $N_{\rm SC} \sim 10^4$ stars, the values of $\bar{n}_\star$ and $\langle\sigma_{\rm coll}(v_{\rm rel})v_{\rm rel}\rangle$—the latter quantity requiring in principle a multi-dimensional integration over the distribution function to find the proper average—needed to be economically calculable. To compute the mean cluster density $\bar{n}_\star = \bar{n}_\star(r,\theta)$ (axisymmetry of the cluster was assumed), a 2-dimensional grid spanning the relevant decades of radius and range of θ was created anew every iteration, and by computing the fraction of time that each of the $N_{\rm SC}$ stars occupied the discrete bins, the average occupation of each bin (which could thus be $\ll 1$ in principle), i.e., the average number density, was found. Although the average occupation was calculated assuming the orbits were Newtonian, because the mean occupation in $r-\theta$ space depends on the argument of periapse ω and because the amount of relativistic precession (change in ω) suffered was

typically $\sim 0.1 - 1$ rad per orbit, the average occupation was itself computed as the average of the mean occupations for $\omega = 0$ and $\omega = \pi/2$ (any two orthogonal directions would have been adequate). A proper averaging requires integration over ω and would have been prohibitive, but as the mean occupation was normally found to vary by at most a factor of two between $\omega = 0$ and $\omega = \pi/2$, averaging the two was satisfactory for present purposes. The mean density obtained by this procedure is only accurate for $r < r_{\rm SC}$. To compute the mean density at larger radii, a least-squares fit to the distribution function given in eq. (5.3) was done on the stars in the SC zone to estimate the current effective values of n, p, and q, and by matching the easily calculated density of the resulting model to the actual density at $r_{\rm SC}$, the profile for larger radii was obtained.

The value of $\langle \sigma_{\rm coll}(v_{\rm rel}) v_{\rm rel} \rangle$ was approximated by

$$\langle \sigma_{\rm coll}(v_{\rm rel}) v_{\rm rel} \rangle \approx \pi \langle (R_\star + R)^2 \rangle (\langle v_{\rm rel} \rangle + v_{\rm esc}^2 / \langle v_{\rm rel} \rangle),$$

which in tests was found to be accurate to $\lesssim 10\%$. The value of $\langle (R_\star + R)^2 \rangle$ was found by direct averaging over the stars in the SC region. Computation of $\langle v_{\rm rel} \rangle (r)$ was done by first defining $g_{\rm coll}(r) = \langle v_{\rm rel} \rangle (r) / v_{\rm orb}(r)$ and then pre-computing a (5-dimensional!) table of values A and B such that $g_{\rm coll}(r) \approx A + Br$, depending on a, ϵ, i, p (the velocity space anisotropy parameter), and q (the real space anisotropy parameter) (dependence on n, the power law index for the density, was extremely weak and was neglected), from which the values of A and B for the given star were interpolated, allowing very efficient calculation of $\langle \sigma_{\rm coll}(v_{\rm rel}) v_{\rm rel} \rangle$.

Most of the preceding, rather elaborate procedure was done to allow reasonably accurate computation of $\tau_{\rm coll}$ when an accretion disk was added to the simulations, which causes all three parameters n, p, and q to systematically increase over the course of a simulation. Comparison of the values for $\tau_{\rm coll}$ computed by this run-time procedure with accurate routines which explicitly performed the multi-dimensional integrals for each individual case showed the above procedure to have typical errors of $\sim 10 - 20\%$, with few values off by more than 50%, which was deemed quite acceptable, particularly since $\tau_{\rm coll}$ affects only the *rate* at

which collisions were made to occur, not the circumstances or outcomes of those collisions—by way of analogy, if the dynamical evolution of the stars in the SC zone is like a movie, then the value of $\tau_{\rm coll}$ doesn't alter the movie's ending, only whether it progresses at normal speed or in slow motion (if $\tau_{\rm coll}$ is too small) or fast forward (if it's too large).

Once the value of $\tau_{\rm coll}$ was computed for a star, it was given the opportunity to suffer a collision that iteration with probability $1 - e^{-\tau_{\rm coll}}$. The result of doing this for all stars was a list of stars that were to suffer a collision that iteration, which were then randomly paired off with each other and set up to collide by suitably altering the orbital elements ω and Υ for each, with the radius of the collision point being randomly drawn from a probability distribution properly weighted by the input density profile. Subsequent treatment of the collisions was handled identically as for ones found by the collision finding algorithm.

The results of the simulations indicate that the above procedure produced collisions nearly indistinguishable from ones which would have been found by the collision finder, as judged by the smooth transition between the two regions or similarity in form in the histograms describing the circumstances and outcomes of the collisions (cf. §5.4.1 and Figs. 5.7-5.8). This afforded confidence in doing simulations lacking a CF entirely, allowing a considerably greater number of models to be examined. Repeating a simulation with and without a CF zone also allowed the significance of the loss in collision probability for the outer CF stars mentioned in the previous section to be estimated. Since in absolute terms the number of collisions occurring in the SC region was considerably larger than in the CF region, this loss of probability was not found to alter the results noticeably. Complete results are given in §5.4.

5.3.5. *The Reservoir*

The 'reservoir' consists of all stars not explicitly included in the simulations, i.e., all stars with periapses greater than $r_{\rm SC}$. The dynamical evolution of these stars is not considered here, and because of the longer timescales involved at

greater distances, that which can be expected to occur during the time interval covered normally would not be substantial in any event. However, relaxation processes in the outer cluster would lead to the occasional injection of a star into the inner parts that are being treated numerically, and to reach a quasi-steady state solution this action must be taken into account. On the other hand, there are plausible circumstances under which a one time injection of stars could occur, such as the tidal disruption of a globular cluster or dwarf spheroidal passing close to a larger galactic nucleus, in which the stars would be slowly consumed without replenishment. To examine this latter case, some of the runs did not include the reservoir at all but instead allowed the cluster to slowly deplete itself until densities dropped to such a level that collision timescales became very long. For the simulations in which it was desired to reach a steady state, the injection of stars by the reservoir was treated as follows.

By Jean's Theorem, since $\mathcal{E}, L$, and L_z are all conserved in a $1/r$ potential (as well as in the Kerr geometry, with suitable identification of the constants of motion), any distribution function of the form $f = f(\mathcal{E}, L, L_z)$ is a solution of Vlasov's Equation and describes a valid equilibrium state, assuming relaxation is the dominant dynamical process. In equilibrium, the rates of relaxation into and out of an orbit of given a, ϵ, and i must be the same, so that the distribution of orbital parameters is time-independent. Hence the newly relaxed parameters of those reservoir orbits which have changed so as to bring them inside $r_{\rm sc}$ must follow the same distribution as the orbits already inside $r_{\rm sc}$, and the orbital elements of each star injected can be had by the same procedure used to set up the initial cluster, which was described in §5.3.1. What remains to be computed is the rate at which these stars should be added to the inner cluster. To estimate this, recall that the relaxation time for a star on a given orbit can be thought of as the timescale on which the orbit's mean square angular momentum changes by $L^2_{\rm max}(\mathcal{E}) = 1/(2\mathcal{E}) = a$ (e.g., Lightman & Shapiro 1977). For a star with given a and ϵ such that $a(1-\epsilon) > r_{\rm sc}$, we want to know the average length of time it will take for this star to diffuse onto an orbit with $a'(1-\epsilon') < r_{\rm sc}$; by then averaging

over all relevant a and ϵ using the distribution function, the net rate at which stars should be added will be found. The rate of change of L^2 can be written

$$\frac{d(L^2)}{dt} \sim \frac{L_{\rm max}^2}{t_{\rm r}} = \frac{\tau_{\rm r}}{2\pi a^{1/2}}, \tag{5.6}$$

where the orbit-averaged, dimensionless relaxation depth $\tau_{\rm r}$ is given by

$$\begin{aligned} \tau_{\rm r}(a,\epsilon) &= 2\int_{a(1-\epsilon)}^{a(1+\epsilon)} \frac{dt(r)}{t_{\rm r}(r)} = 2\int_{a(1-\epsilon)}^{a(1+\epsilon)} \frac{dr}{v_r(r)t_{\rm r}(r)} \\ &= \mathcal{C}a^{3-n}\int_{1-\epsilon}^{1+\epsilon} \frac{(2/x-1)^{-3/2}x^{1-n}\,dx}{[(1+\epsilon-x)(x-1+\epsilon)]^{1/2}} \equiv \mathcal{C}a^{3-n}g(\epsilon), \end{aligned} \tag{5.7}$$

where the relations $t_{\rm r}(r) \sim v_{\rm orb}^3(r)/[15G^2m_\star^2 n_\star(r)\ln(0.4N_\star)]$, $v_{\rm orb}^2 = 2/r - 1/a$ (the star's orbital speed), $v_r^2 = [a(1+\epsilon)-r][r-a(1-\epsilon)]/(r^2a)$ (the radial component of $v_{\rm orb}$), and $n_\star(r) = n_0 r^{-n}$ have been used, and $\mathcal{C} = 15n_0(m_\star/M)^2\ln(0.4N_\star)$. The dimensionless function $g(\epsilon)$ can be approximated to $\lesssim 25\%$ by $g(\epsilon) \approx (5-3\epsilon)(1+\epsilon)^{3-n}/[2(1-\epsilon)]$ (this formula becomes exact as $\epsilon \to 1$). Now the given orbit will diffuse into the region $r < r_{\rm sc}$ when the condition $L^2 < L_{\rm min}^2 \approx a(1-\epsilon_{\rm min}^2)$ is met, where $a(1-\epsilon_{\rm min}) = r_{\rm sc}$. The amount of time this takes is $\Delta t(a,\epsilon) \sim 2\pi a^{3/2}(\epsilon_{\rm min}^2 - e^2)/\tau_{\rm r}$. The distribution function was then used to average this result over the outer cluster to derive the net injection rate. For $n = 7/4$, $p = q = 0$, $n_\star(1\,{\rm pc}) = 10^6 n_6\ {\rm pc}^{-3}$, and $M = 10^8 M_8\ M_\odot$, the numerical value of this injection timescale is

$$t_{\rm inj} \approx 5000 M_8^{3/4} n_6^{-1} \left(\frac{m\star}{M_\odot}\right)^{-2} \left(\frac{N_\star}{10^6}\right)^{-1}\ {\rm yr}, \tag{5.8}$$

and is a very slowly increasing function of p. For solar mass stars, note that this implies an average mass injection rate of $\sim 10^{-3} - 10^{-2} M_\odot\ {\rm yr}^{-1}$ for typical simulation parameter values; this is roughly comparable to the rate required to power a moderately luminous Seyfert galaxy.

5.4. SIMULATION RESULTS AND DISCUSSION

The numerical results will be presented in a set of diagrams for a series of six fiducial models encompassing a range of hole masses, densities, and initial amounts

TABLE 5.2

Parameters for the Fiducial Models

Model	M_8	p	min/max $m_\star$[a]	n_6	reservoir?
1	0.01	0	0.4/4.0 $M_\odot$	10	no
2	0.3	0	1.0/1.0 $M_\odot$	1	no
3	10	5	10./10. $M_\odot$	0.1	no
4	0.3	0	1.0/1.0 $M_\odot$	1	yes
5	0.01	10	0.4/4.0 $M_\odot$	10	yes
6	0.3	5	0.4/4.0 $M_\odot$	1	yes

[a] Salpeter mass function ($f(m) \propto m^{-2.35}$) assumed.

of anisotropy. For convenience the models have been assigned reference numbers; these and the corresponding parameters are listed in Table 5.2. The five variable parameters are the mass of the hole, $M = 10^8 M_8 M_\odot$, the velocity anisotropy parameter of the initial cluster, p, the initial stellar mass spectrum, $f(m)$, the density scale $n_6 = n_\star(1\ \mathrm{pc})/(10^6\ \mathrm{pc}^{-3})$, and whether or not the reservoir was included to reach a steady state in the core or instead the stars were allowed to deplete themselves. As can be seen in Table 5.2, in fact the value of n_6 was always kept the same for a given value of M_8; this was essentially done for convenience to keep the number of stars within a fixed number of gravitational radii comparable for each model, giving them similar dynamic ranges in radius. Most results scale easily with n_6 so this is not a significant restriction. In addition, the input clusters were all spherically symmetric (flattening parameter $q = 0$) with initial densities $\rho(r) \propto r^{-7/4}$. All models presented here used $j = 1$ for the spin of the MBH; the comparison models computed using $j = 0$ showed no perceptible difference from the corresponding $j = 1$ model and are not shown. We will comment on this issue in §5.5. A rather large number of plots have been included here for perusal, and commenting on each individually would become quite tedious. Instead of this an integrated discussion of each model will be given, highlighting the salient

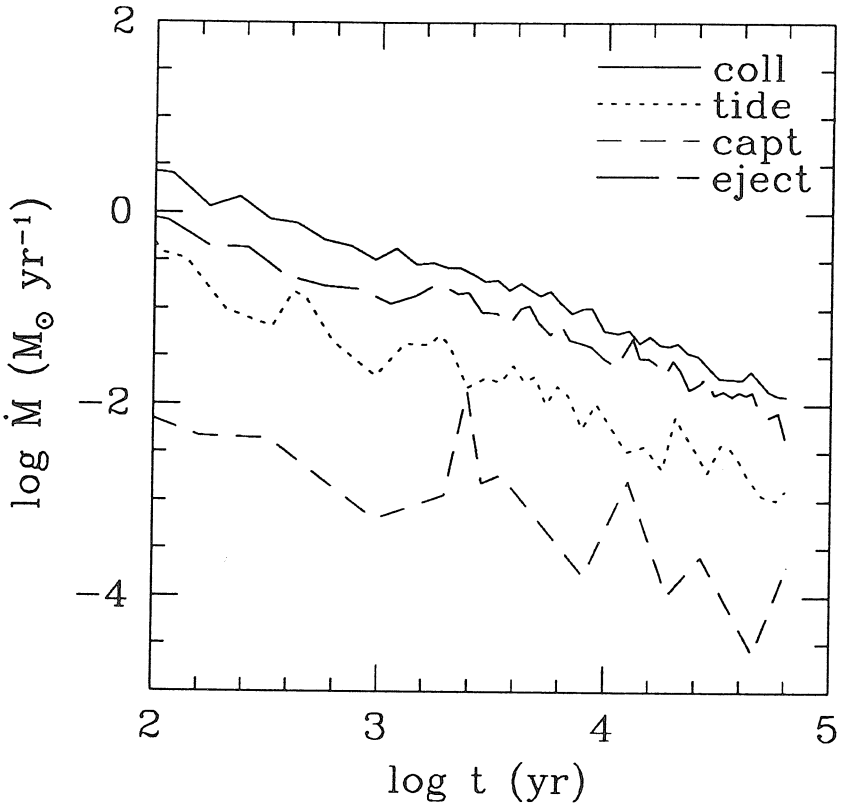

Fig. 5.1: Cluster mass depletion rates due to collisions, tidal disruption, and post-collision stars whose orbits are either captured by the hole or unbound (ejected) with respect to it, for cluster model 1 (see Table 5.2 for numerical parameters and text for definitions thereof.) The approximate scalings at late times are $t^{-0.5}$. Collisions from this initially rather dense ($n_6 = 10$) cluster can support Eddington-limited accretion for $\sim 10^5$ yr (at 10% conversion efficiency).

features or differences compared with other models. We begin with the models not incorporating a reservoir.

5.4.1. *Single Injection Models*

Models 1-3 did not include a reservoir and hence display continually decreasing densities and collision rates. Such models could correspond to a sudden influx of stars to the nucleus as from a tidally disrupted globular or close interaction with another galaxy, which are afterwards gradually depleted through collisions

and consumption by the black hole. They are also useful as a reference to compare to similar steady state models, which include injection of new stars from an extended outer cluster (§5.4.2). The fundamental quantities of interest are the mass loss rates over time and the evolution of the density profile and stellar mass spectrum. Other useful diagnostics shown for some of the models are the changing distributions of orbital parameters a, ϵ, and $\mu = \cos i$ (or equivalently the distribution function $f(\mathcal{E}, L, L_z)$).

The results for model 1 are given in Figs. 5.1-5.8. This model consists of a $10^6 M_\odot$ black hole and an isotropic, relatively dense cluster containing a range of stellar masses. Figure 5.1 shows the mass release rates as a function of time into four different channels: collisional, tidal, captured, and ejected. The collisional mass loss is simply the total amount of gas unbound from colliding stars. As this mass is typically ejected at high velocity and at random angles from the radial direction, roughly half of this mass would be expected to remain bound to the hole and half to be lost from the core; in all simulations this mass loss rate is also the largest of the four, usually by factors of a few. The tidal mass loss rate consists of the mass of all tidally disrupted stars, and in the single injection models represents those stars which, after suffering a collision, found themselves on a modified orbit crossing their tidal radius; for the models including a reservoir, it also includes orbits driven into the loss cone by relaxation processes. The captured mass rate is due to stars whose post-collision orbits have so little angular momentum that they cross the horizon and are consumed; although related, this mass was not included in the tidal mass loss because nearly all of the former will be bound to the hole and is likely to plunge into it with little energy release, whereas in the latter case roughly half the mass is likely to be bound and the rest unbound to the hole, and this mass could take many dynamical times to be accreted. The ejected mass rate is due to stars whose post-collision orbits are no longer bound to the hole and hence these stars are ejected from the core in a dynamical time (if not from the nucleus itself). The pattern in Fig. 5.1 was typical, with collisional mass loss being largest, followed by ejected mass, with tidal losses a sometimes close,

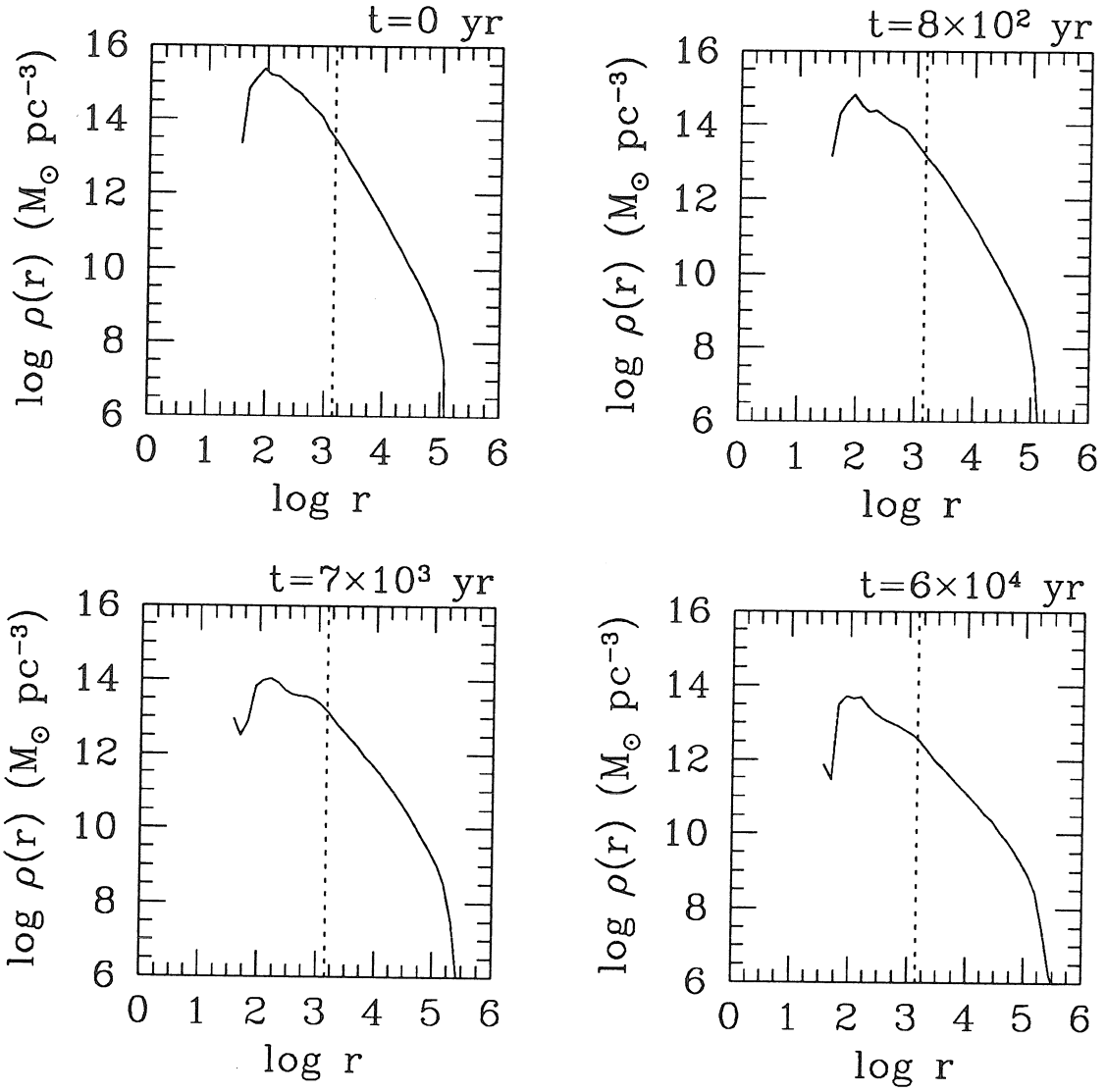

Fig. 5.2: The density of cluster model 1 at several epochs. The dashed line denotes the value of r_{sc} ($= 1400M \sim 10^{-4}$ pc), the maximum simulation radius; outside this limit the density is incomplete. Collisions lead to the development of an inner core of nearly constant density (within statistical fluctuations).

sometimes distant third, with mass loss by direct capture being smallest. Since the cross-section for capture is inside that for tidal disruption (unless the hole is so massive that the horizon size exceeds the tidal radius), one would expect the capture mass rate to lie below the tidal rate under normal circumstances, and this was indeed the case in most models. In Fig. 5.1 the approximate late time scaling of these rates is $\propto t^{-0.5}$.

The evolution of the density in model 1 is shown in Fig. 5.2. As will be seen more clearly in the other models, the collisions produce a core of nearly constant density which increases in radius as evolution continues. In this case the simulation was terminated before the core was well-developed. Figure 5.3 displays the evolution of the mass spectrum. When a range of masses is included, the qualitative trend is the production of a low mass tail consisting of surviving post-collision stellar fragments and depression of the high mass part of the spectrum, as these stars have the largest cross-sections and so will collide sooner on average than the other stars. In single injection models the mass spectrum will be substantially altered when most stars have had an opportunity to collide, but for steady state models the addition of new stars stabilizes this process, as discussed below. Figures 5.4-5.6 show the distributions of orbital parameters at several epochs for model 1. As with high mass stars, stars with small semi-major axes have shorter collision timescales and thus will be preferentially removed from the cluster, as seen in Fig. 5.4. Since all stars in the simulation have $a(1-\epsilon) < r_{\rm sc}$, the large-$a$ orbits are biased towards high ϵ and hence removal of the small-a stars leads to increasingly eccentric remaining orbits, as shown in Fig. 5.5. Returning to the question raised in §5.1 concerning the possibility of Kerr effects producing a flattening of the cluster during the course of its evolution, Fig. 5.6 show the distribution of orbital inclinations including the difference between the first and last. The short answer is that no evidence of this or any other type of non-sphericity was found in any of the simulations; this will be discussed further in §5.5.

Of the six models shown here, only model 1 used the collision finding algorithm; the rest were done with fully randomly selected collisions. This was mainly

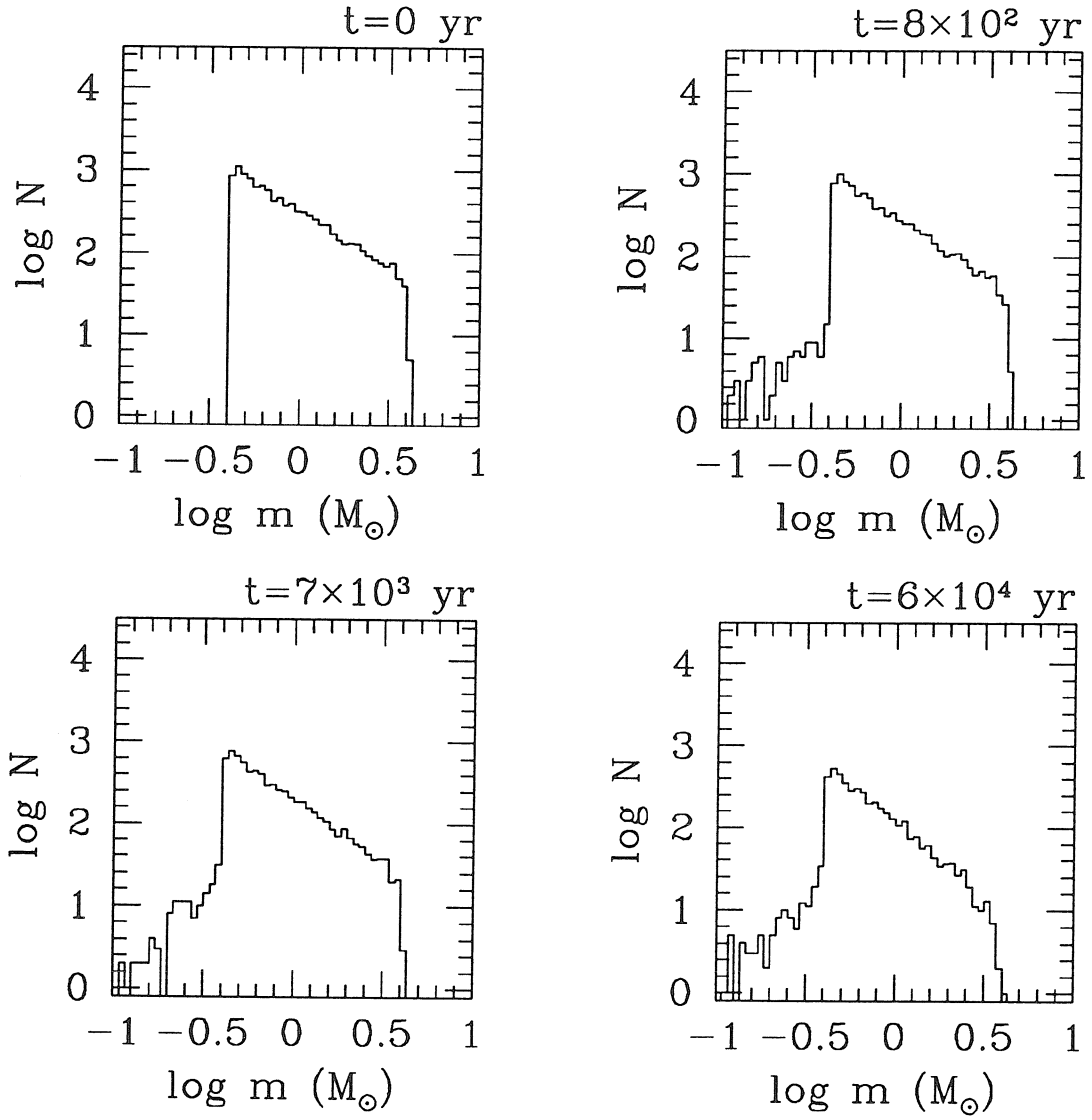

Fig. 5.3: The mass function of cluster stars at several times for model 1. The larger geometric cross-sections of high mass stars leads to their depletion relative to other stars, with post-collision stellar fragments creating an excess of low-mass stars. The overall effect is small here as most stars had not collided when the simulation was terminated.

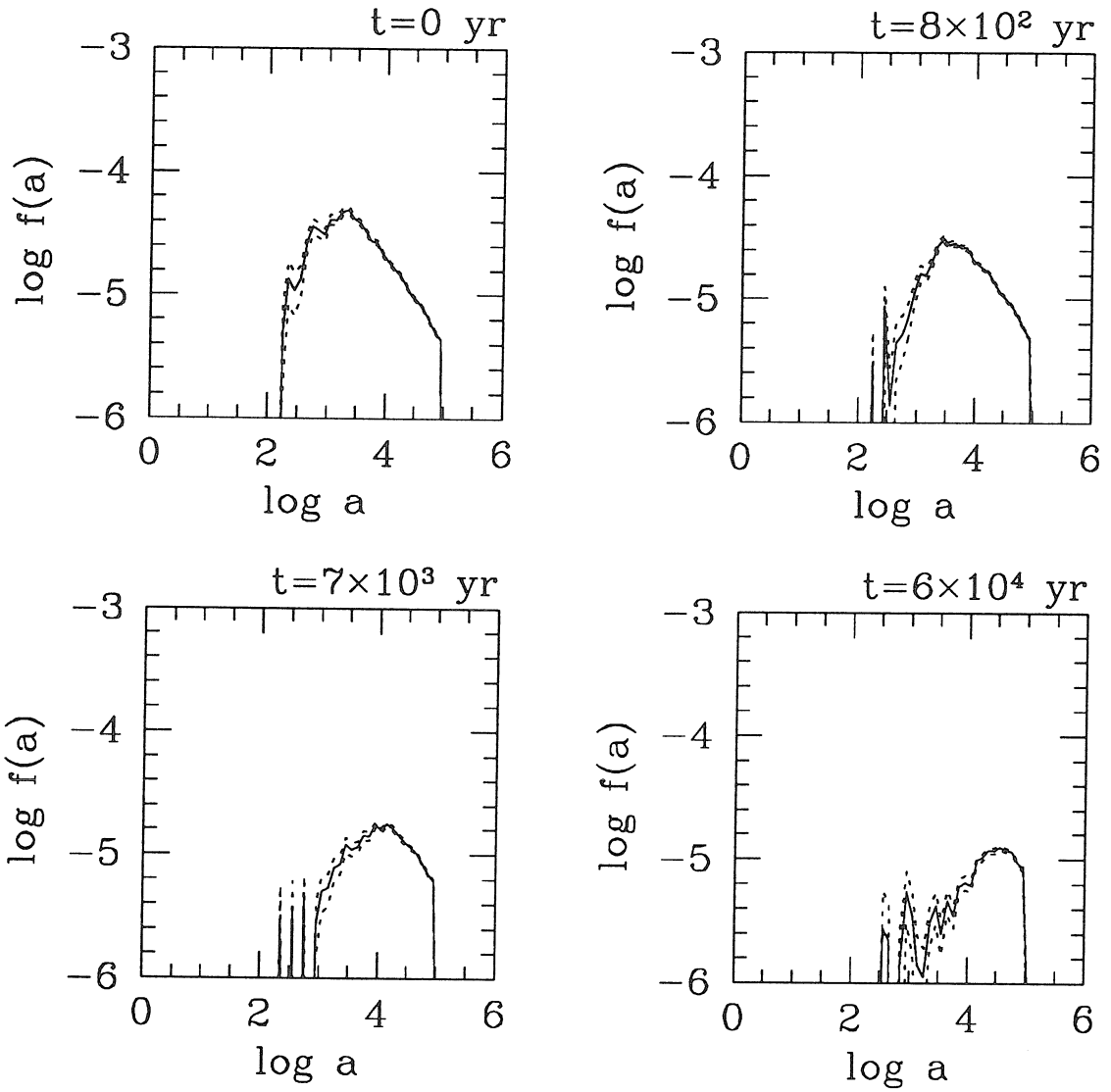

Fig. 5.4: The normalized ($\int f(a)da = 1$) distribution at several times of semi-major axes of included stars for model 1. Small a orbits are preferentially removed from the cluster because of their short collision timescales. The dashed lines are $1-\sigma$ error estimates based on Poisson statistics.

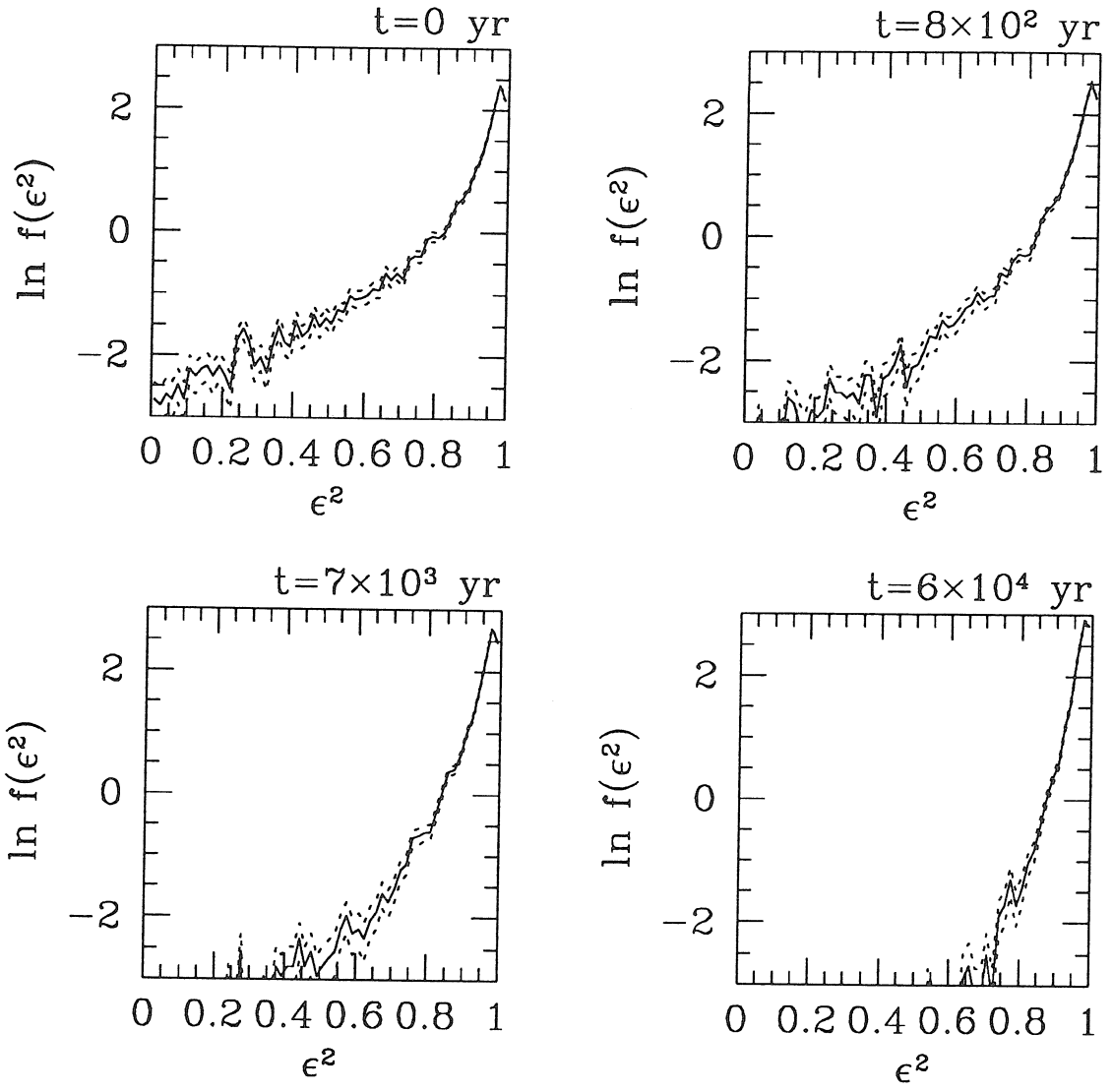

Fig. 5.5: The normalized distribution of eccentricities of included stars for model 1. Velocities in the cluster core become progressively anisotropic as orbits with small ϵ, which of necessity have relatively small values of a since they must cross r_{sc}, tend to have highest collision frequencies (cf. Fig. 5.4). Dashed lines are $1-\sigma$ error bars. Scales have been chosen so that an unbiased sample of stars obeying the distribution function (5.3) will have a line of slope $-p$ on these plots; in this case, the initially isotropic ($p = 0$) cluster does not produce a flat line because the stars are biased by the requirement that they have $a(1-\epsilon) < r_{sc}$.

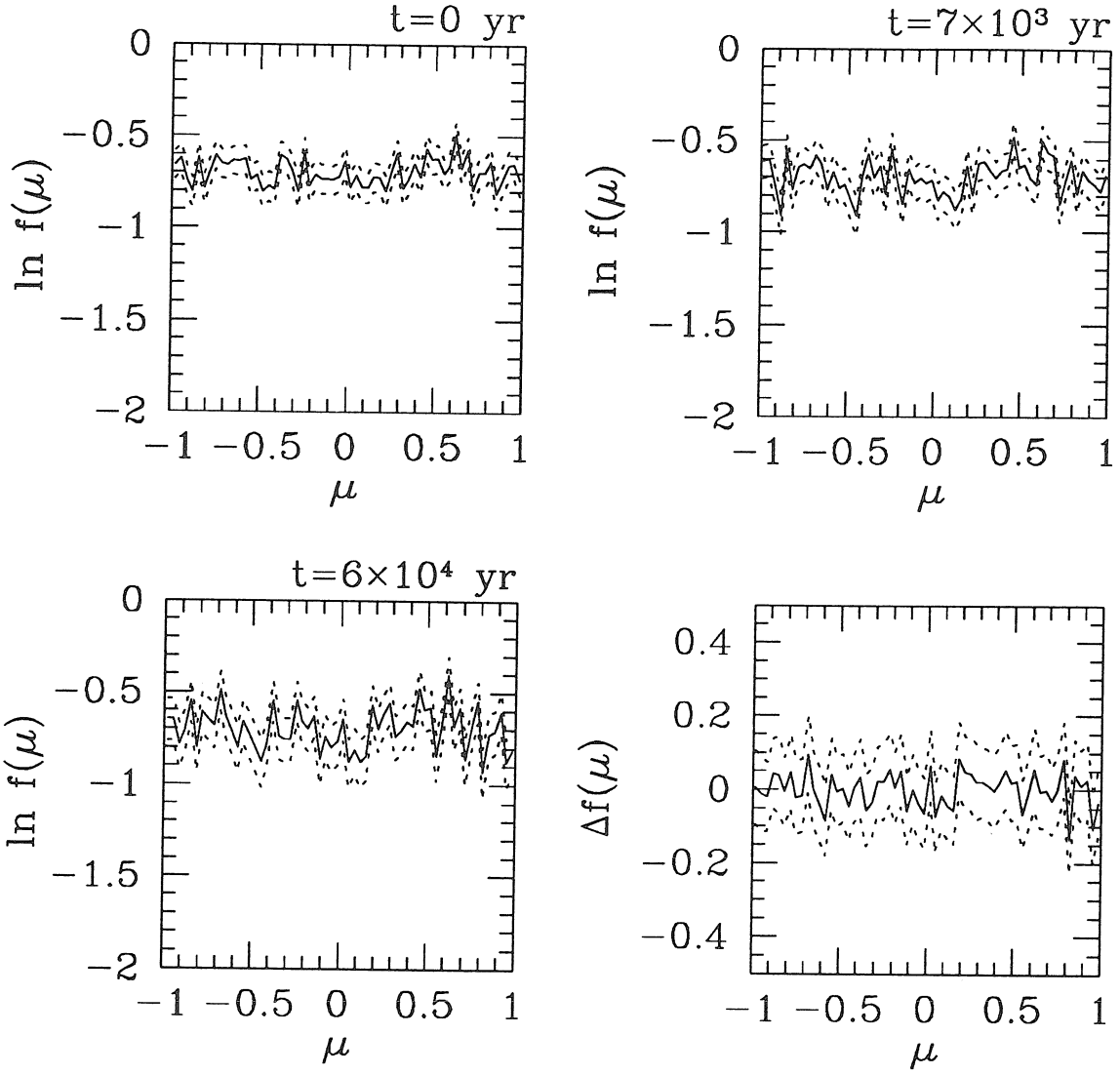

Fig. 5.6: The normalized distribution of inclinations $\mu = \cos i$ for model 1, with $1-\sigma$ error bars (dashed lines). The lower right panel plots the difference between the $t = 6 \times 10^4$ and $t = 0$ distributions, which is consistent with 0, and hence there is no statistically significant evidence indicating the development of spatial anisotropy (such as flattening towards the equator) in the cluster. These curves are typical; in no model was any significant departure from spherical symmetry detected.

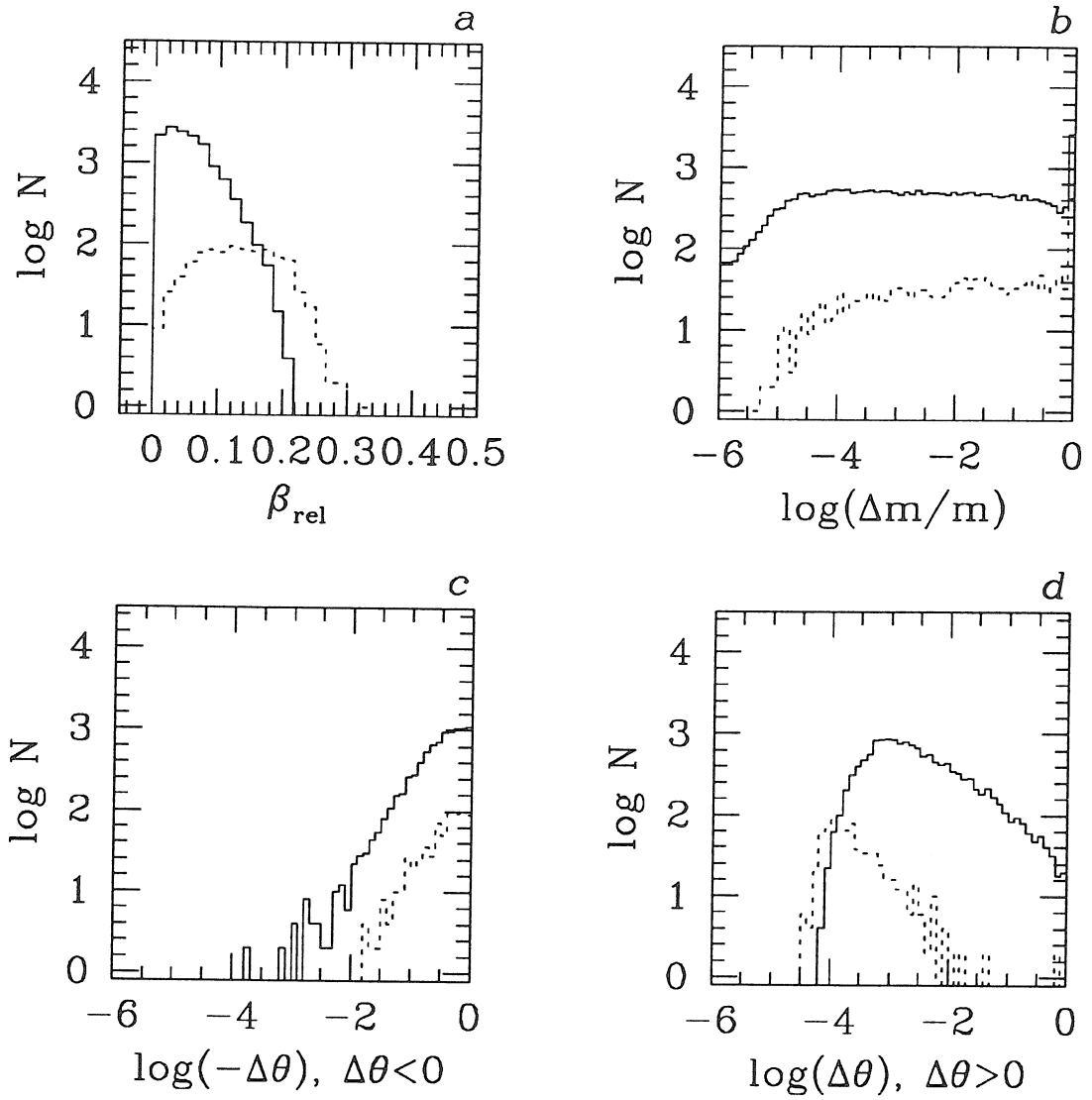

Fig. 5.7: Histograms of various collision statistics for model 1; the solid curve is for statistically selected collisions (see §5.3.4) and the dashed is for those found explicitly using the algorithm of §5.3.3. Fig. 5.7*a* shows the distribution of relative velocities in the (relativistic) center-of-mass frame. Fig. 5.7*b* plots the distribution of relative mass loss suffered per collision; the peak at $\Delta m/m = 1$ is from stars suffering complete disruption in a collision. Figs. 5.7*c* and 5.7*d* show the deflection angles suffered by the stars, on log scales for $\Delta\theta < 0$ and $\Delta\theta > 0$, respectively. Positive values of $\Delta\theta$ denote deflections towards the center-of-mass. Note the smooth transition between collisions found explicitly and statistically. See text for discussion.

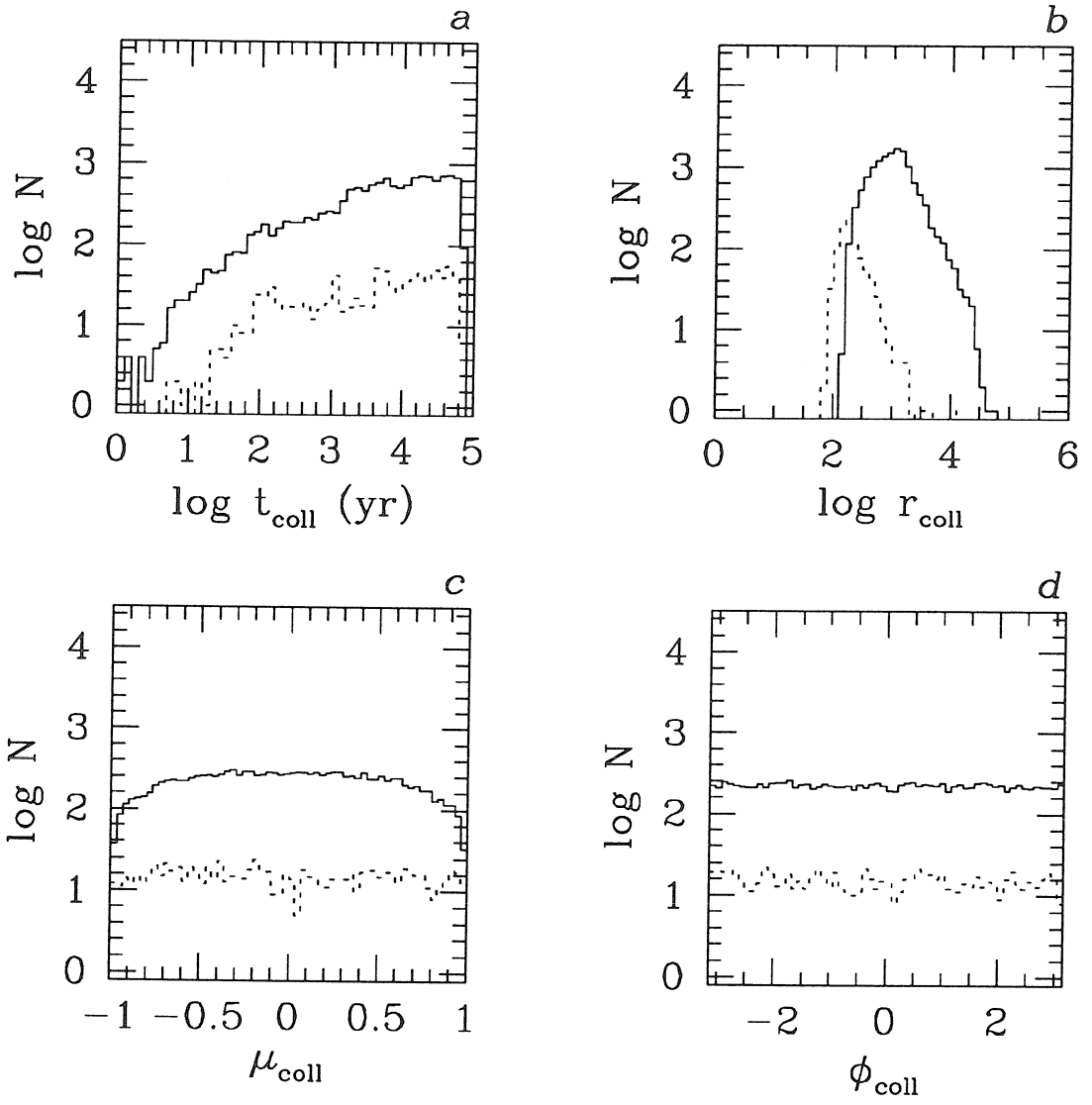

Fig. 5.8: Similar to Figs. 5.7*a-d*, now showing distributions of collision coordinates (t, r, μ, ϕ). As in Fig. 5.7, note the similarity in the distributions between collisions solved for and those produced randomly; see text.

done out of desire to examine the late-time behavior of the models; in particular, use of the collision algorithm was too time-consuming to be able to follow the evolution until a steady state was finally achieved in those models including the reservoir. Figures 5.7 and 5.8 show diagnostic output from which the relative performance of the collision finding algorithm and the statistical collisions routine can be ascertained. Figure 5.7 plots for both routines the distributions of relative velocities, mass loss of each star, and deflection angles of the collisions produced by each routine. The quite similar appearance of these curves indicates that the randomly drawn collisions are basically indistinguishable from those that the collision algorithm would have found. There are systematic shifts in the relative velocities and deflection angles, but these are real differences arising from the fact that the collision algorithm is always applied to stars at the smallest radii, where velocities are systematically larger and hence where collisions will tend to produce smaller deflections. Figure 5.8 shows the distributions of the recovered collision spacetime coordinates; once again the two routines produced very similar results, except for a slight tendency of the statistical collision routine to underproduce collisions near the poles, which comes about because the routine selects pairs of orbits to collide without regard to their inclinations, thereby forcing the collision to occur at or below the lesser of the two inclinations, leading to excess collisions near the equator at the expense of the poles. However, as no evidence of nascent spatial anisotropy was found at large or small radii (where collisions were found directly), it seems unlikely that this minor shortcoming had any detrimental effect on the results. For these reasons most runs were done with entirely statistically selected collisions as they are much faster, with the collision algorithm being used in a few runs to search for relativistic influences and to act as a control.

Figures 5.9-5.13 show the simulation results for model 2 and are analogous to Figs. 5.1-5.5 for model 1. This model consisted of an 'AGN-class' black hole of $3 \times 10^7 M_\odot$ surrounded by an isotropic, moderately dense cluster of equal solar mass stars. One thing of note in Fig. 5.9, which shows the mass release rate, is the near equality of $\dot{M}_{\rm t}$ and $\dot{M}_{\rm cap}$, due to the much smaller tidal radius ($r_{\rm t} \sim 10$

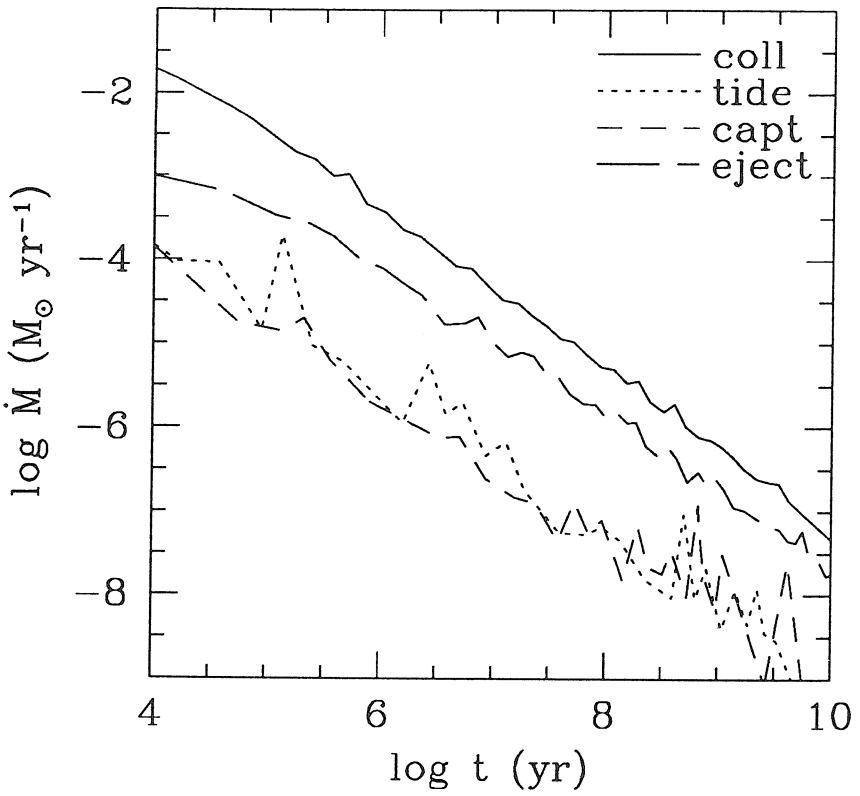

Fig. 5.9: Mass liberation rates for cluster model 2 (cf. Fig. 5.1). For this medium density cluster ($n_6 = 1$), Eddington-limited accretion rates can be sustained for $< 10^3$ yr, assuming 10% efficiency, before asymptotically declining; at late times $\dot{M}_i \propto t^{-1}$.

as opposed to ~ 100 for model 1) associated with model 2's larger hole; the rates decrease at late times as t^{-1}. The development of the constant density collision core is readily apparent in Fig 5.10. Qualitative results for the remaining figures match those for model 1, except that in general the effects are more pronounced due to the more advanced state of evolution.

Figures 5.14-5.18 show the corresponding results for model 3, consisting of a 'quasar-class' black hole ($M = 10^9 M_\odot$) and a modestly dense and anisotropic cluster of high mass stars. Note the very small tidal contribution to the mass loss in Fig. 5.14; for this model $r_t \sim 2$, which is so close to the horizon (at $r = 1$) that nearly all orbits inside r_t plunge directly into the hole; asymptotically the rates

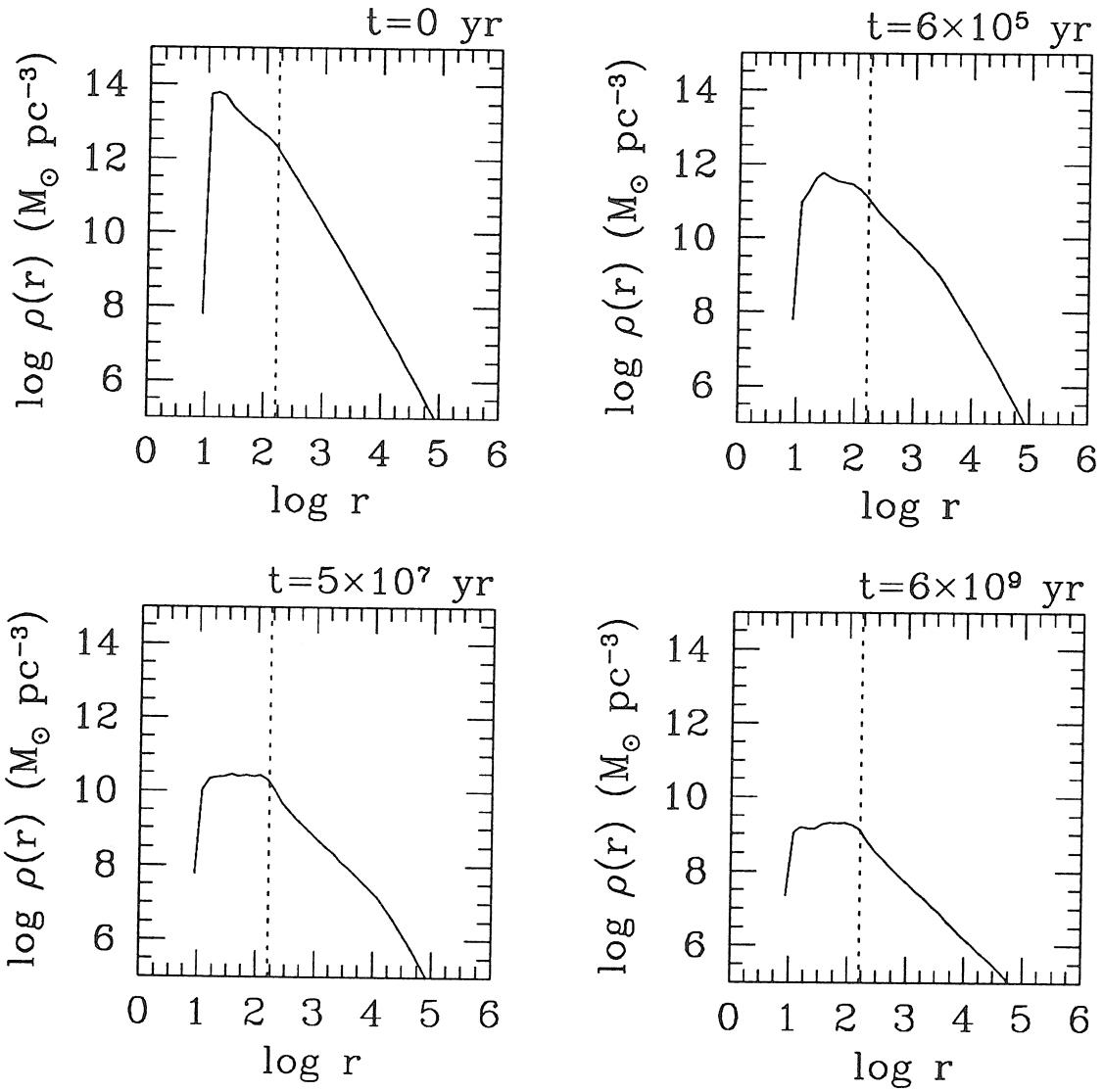

Fig. 5.10: Evolution of cluster density for model 2 (cf. Fig. 5.2). The completeness limit (dashed line) for this model is $r_{sc} \approx 160 \approx 2 \times 10^{-4}$ pc. The constant density core is well-developed at later times.

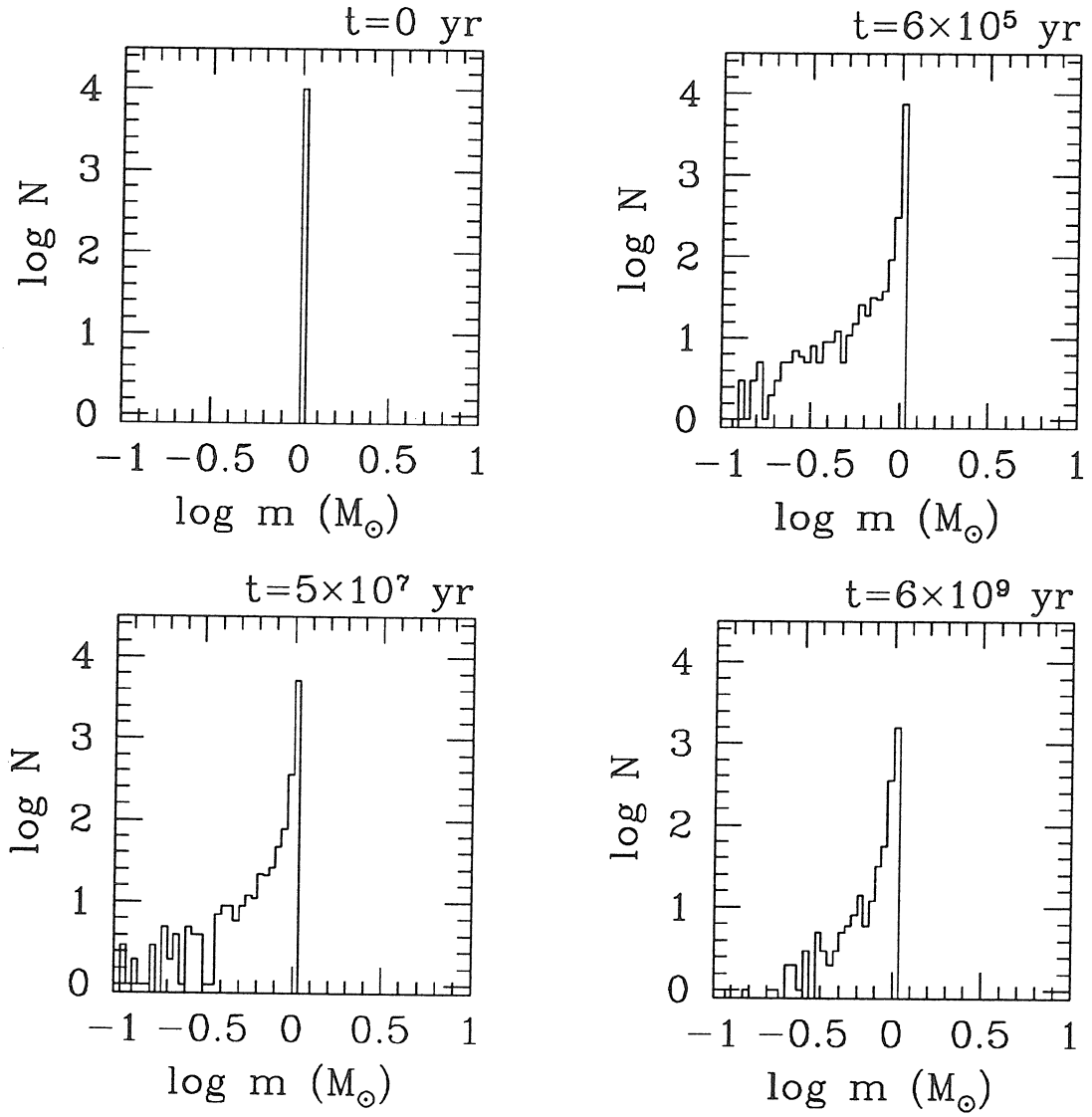

Fig. 5.11: Evolution of the mass spectrum for model 2. The relative occupation of the collision-induced tail is substantial at late times.

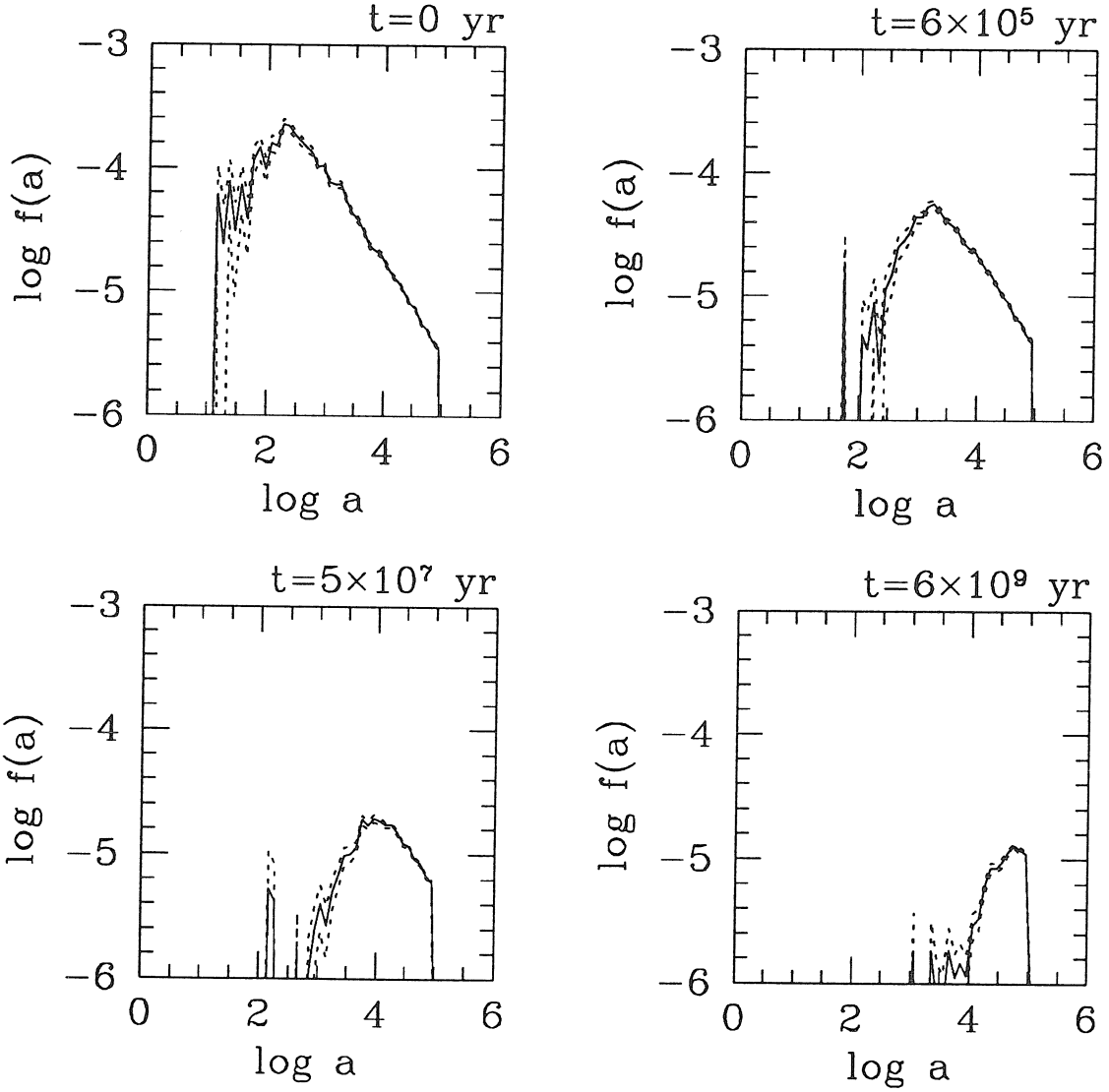

Fig. 5.12: Semi-major axis distributions for model 2. Behavior is similar to that in model 1 (see Fig. 5.4).

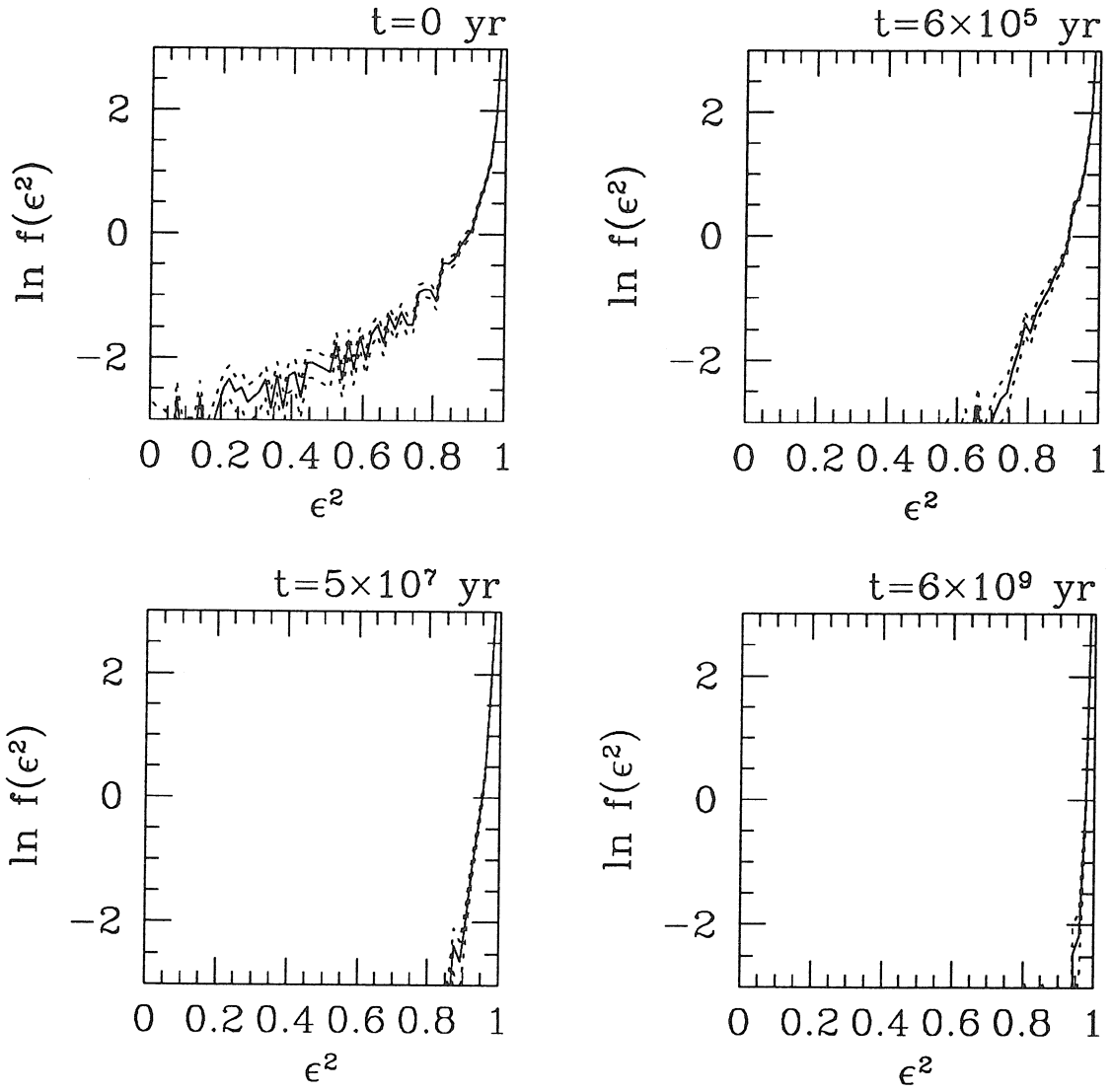

Fig. 5.13: Eccentricity distributions for model 2. Qualitative behavior is similar to model 1, only more extreme due to the advanced state of evolution.

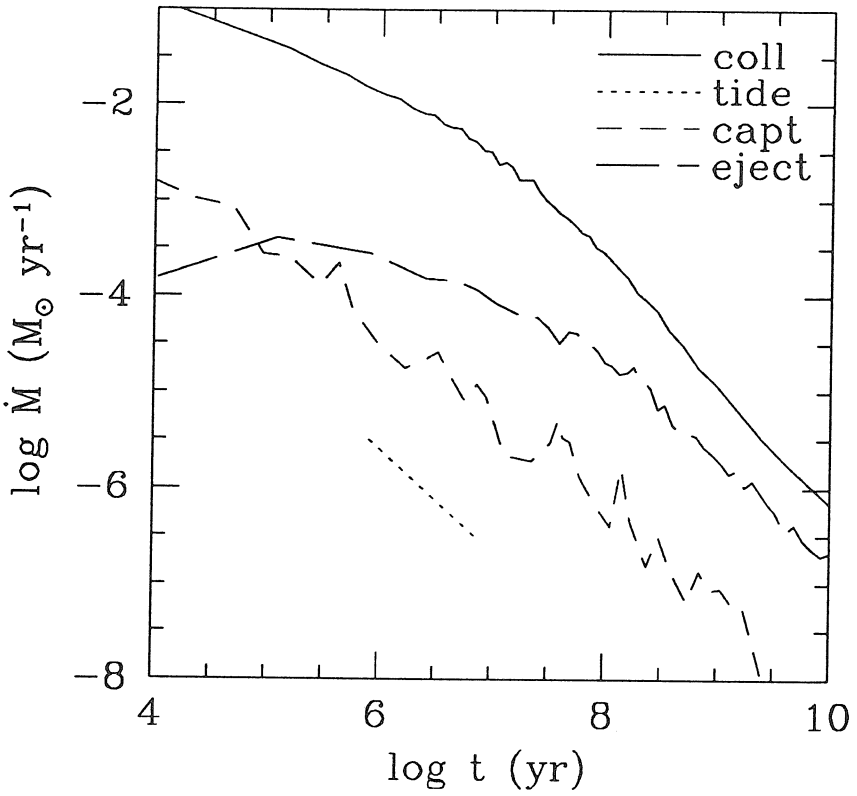

Fig. 5.14: Mass loss rates for cluster model 3, consisting of a very massive ($M_8 = 10$) black hole inside a relatively low density cluster ($n_6 = 0.1$) containing only high mass ($10M_\odot$) stars. Tidal effects are especially weak in this case as the tidal limit is nearly inside the horizon even for these large stars (which also prohibited meaningful calculation of the tidal curve at general times, hence the truncation of it). Scaling at late times is approximately $\dot{M}_i \propto t^{-1.3}$.

scale approximately as $\propto t^{-1.3}$. Qualitative behavior of the remaining figures is similar to the other models; see the figure captions for a few additional comments.

5.4.2. *Steady State Models*

As we are simulating only the inner cusp of a presumably more extended nuclear cluster, the normal expectation is that new stars will occasionally enter the simulation region through relaxation processes. The way this was accomplished in these calculations was described in §5.3.5. Models 4-6 are three examples of

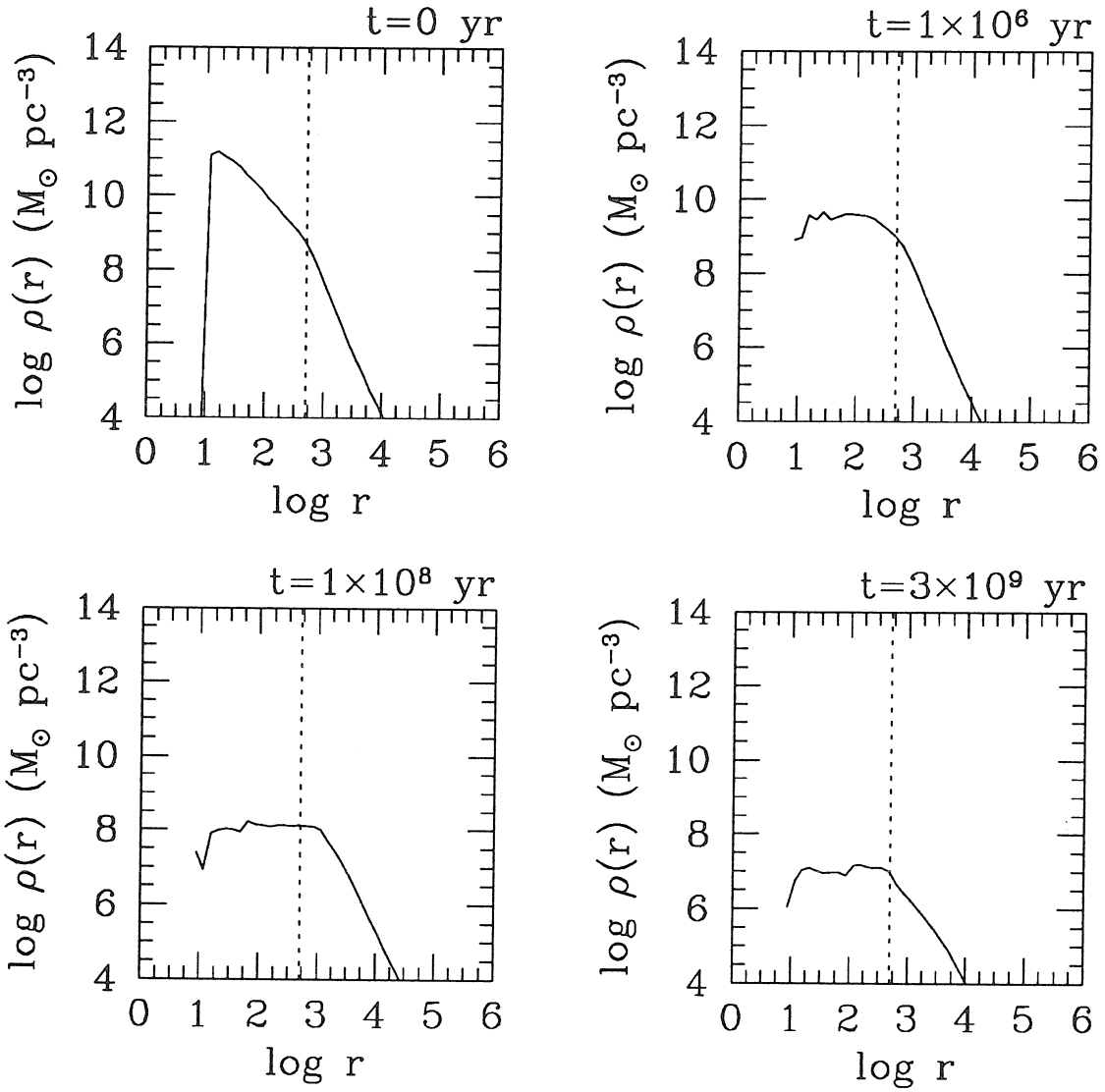

Fig. 5.15: Density profiles for model 3, again exhibiting a flat core. The completeness limit (dashed line) is $r_{sc} \approx 520 \approx 0.03$ pc.

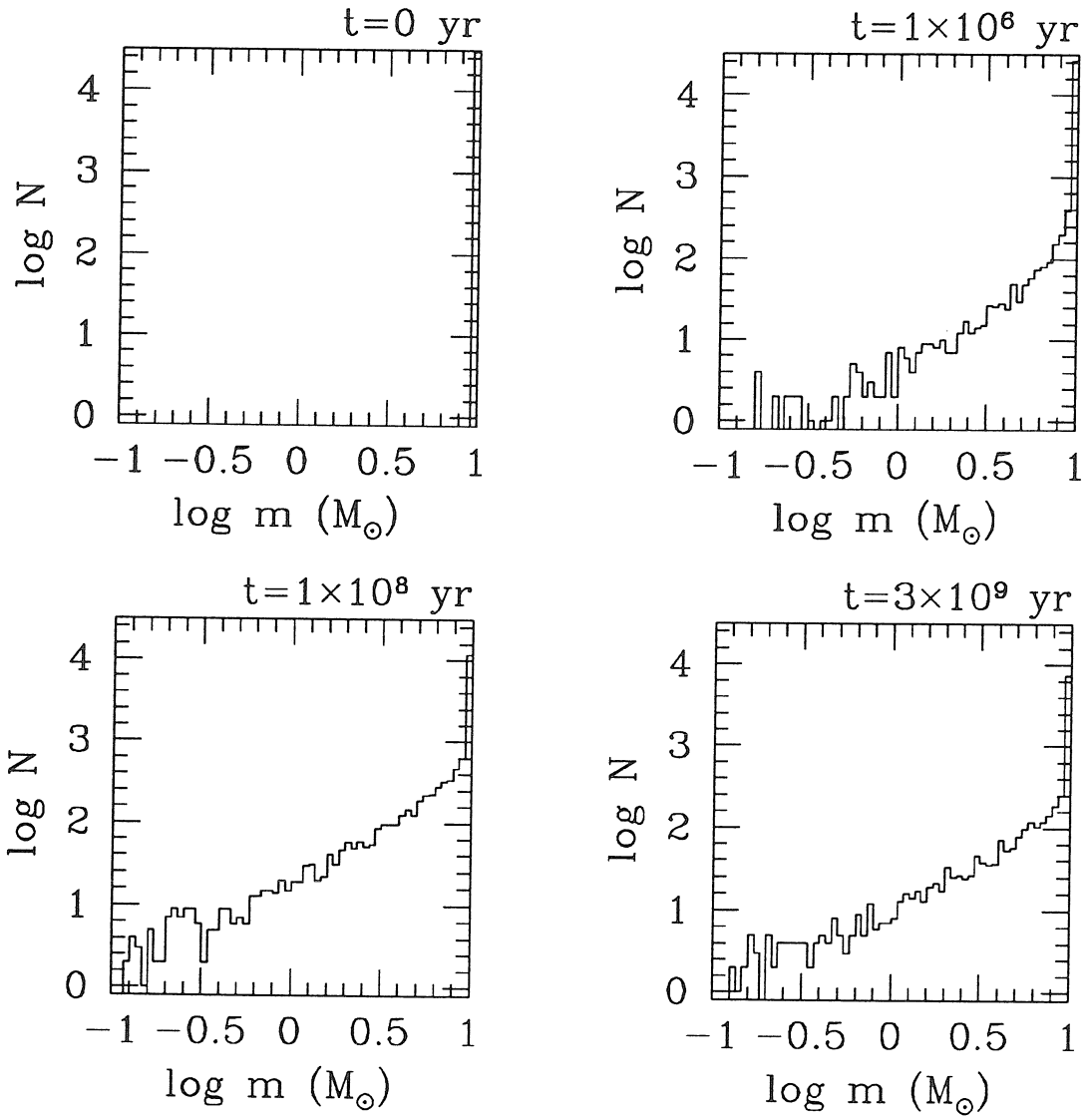

Fig. 5.16: Mass spectrum evolution for model 3. A quite extended tail develops at later times.

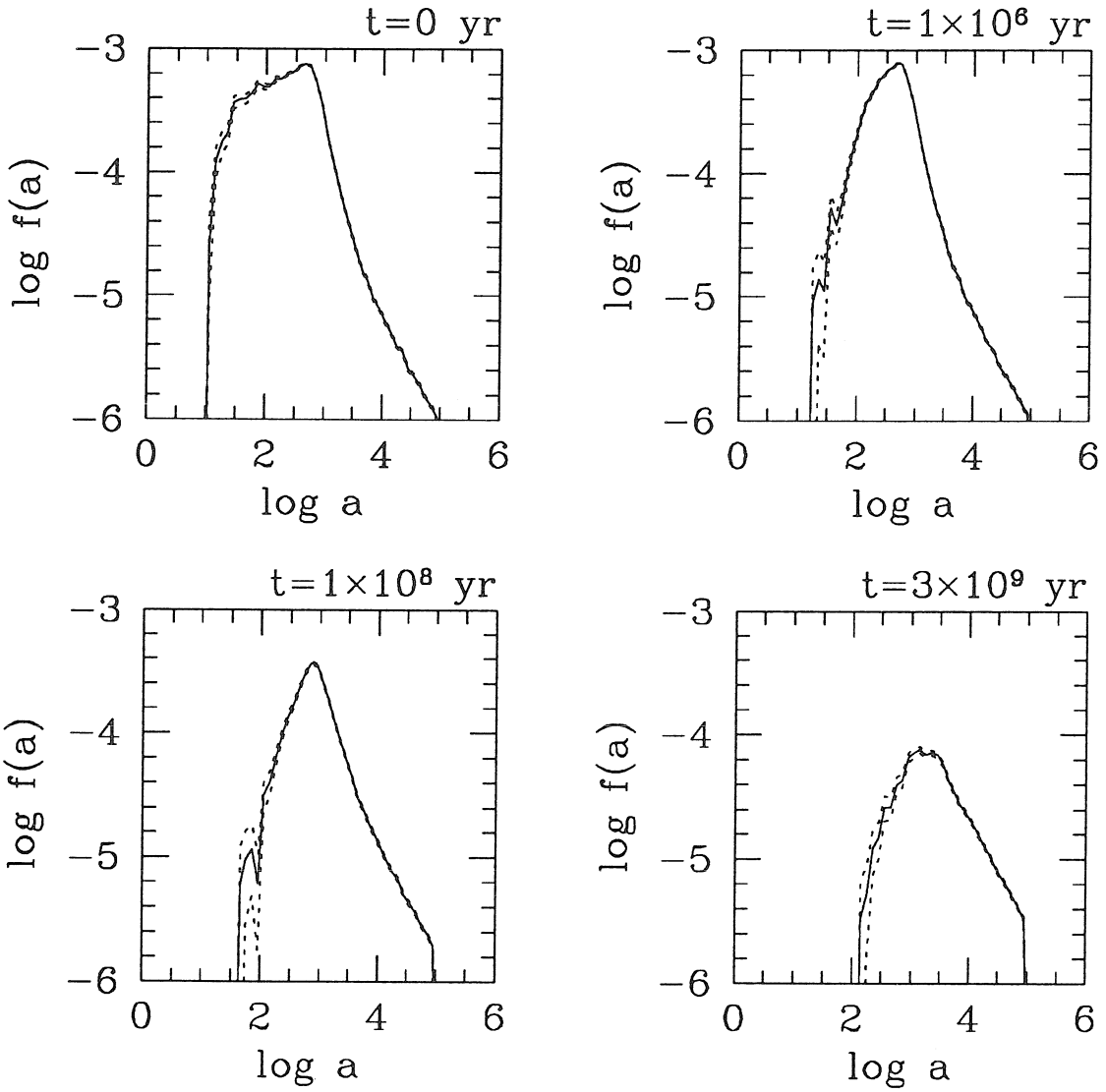

Fig. 5.17: Semi-major axis distributions for model 3. The shift away from small-a orbits is particularly noticeable in this case; cf. Figs. 5.4 and 5.12.

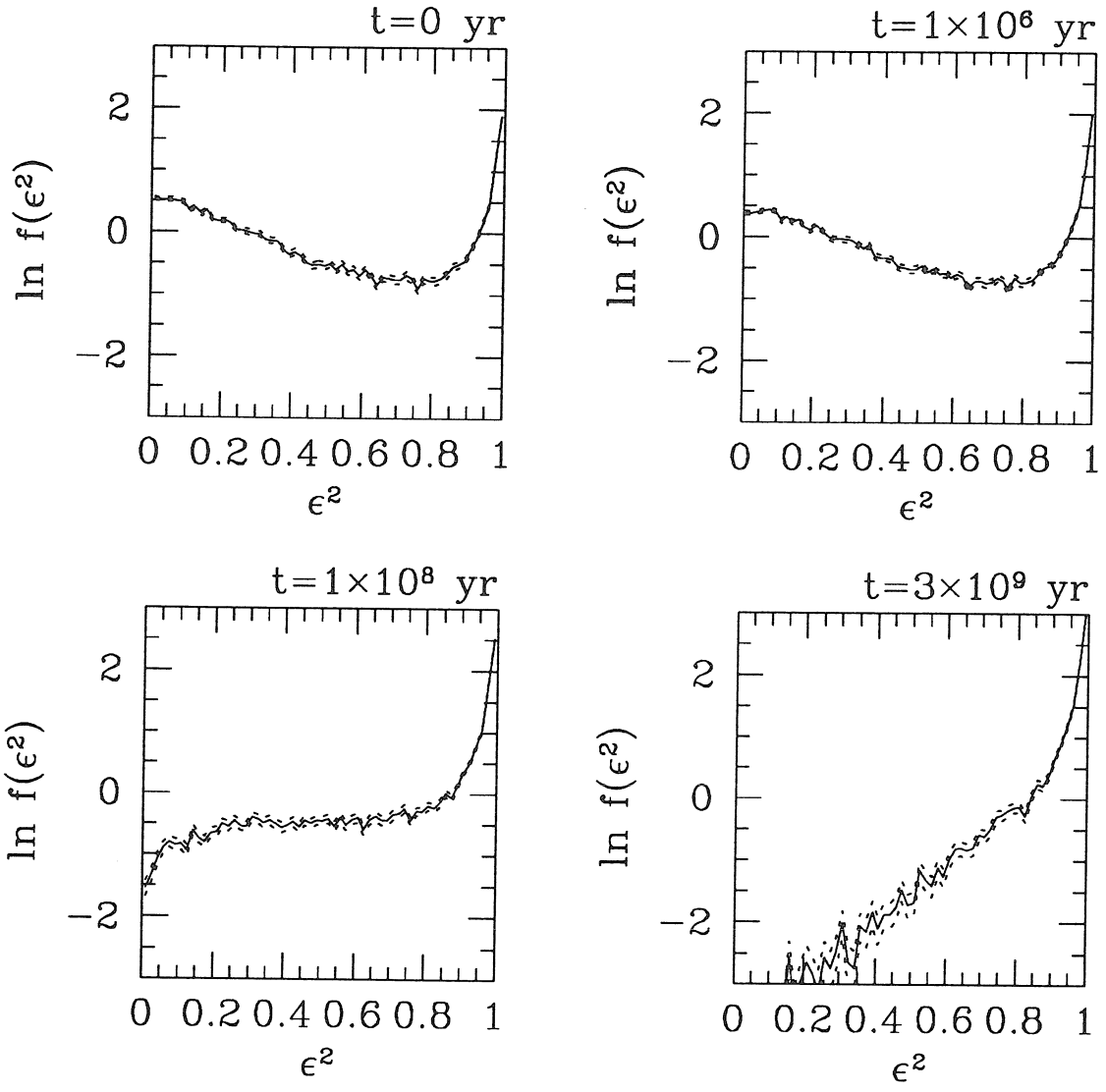

Fig. 5.18: Eccentricity distributions for model 3. The initially mild anisotropy ($p = 5$) is quenched as evolution proceeds; with continued evolution the cluster would be expected to become anisotropic in the opposite direction, towards negative effective values of p, as in Fig. 5.13.

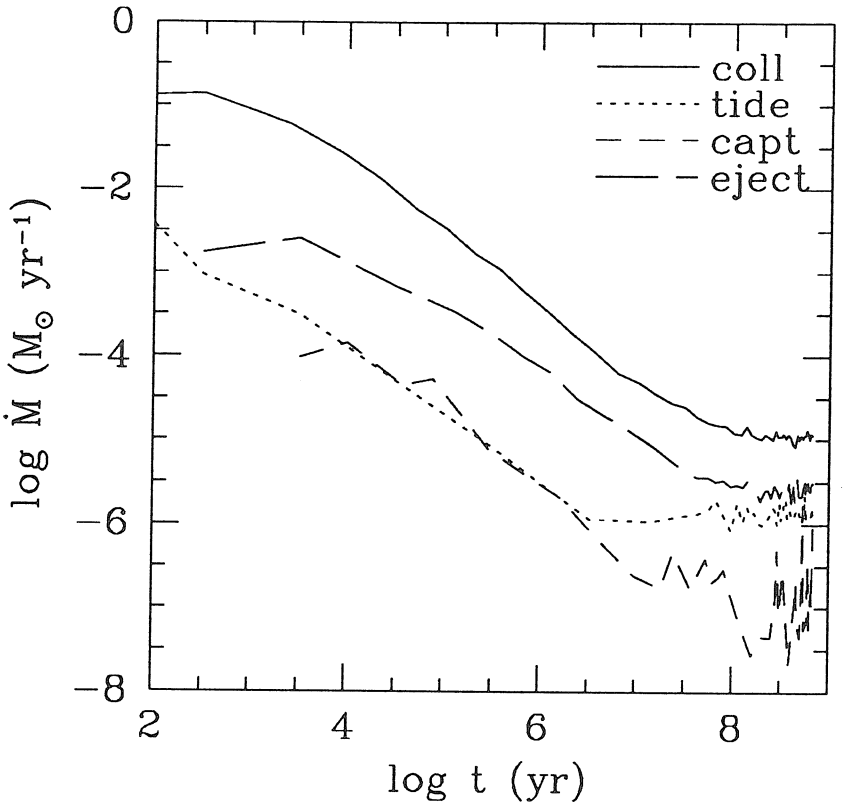

Fig. 5.19: Mass production rates for model 4. This model includes injection of stars from the outer cluster by relaxation and hence the rates reach a steady state. The steady state rate of $\dot{M}_{\rm coll} \approx 10^{-5} M_\odot$ yr^{-1} could be sustained for over a Hubble time under the assumed conditions, although the luminosity would correspond at best to that of a modest Seyfert galaxy ($L_{\rm bol} < 10^{42}$ erg s^{-1}).

the outcome of this procedure. We mention first that no simulations of quasar-class black holes are shown here as they were found to be uninteresting due to their very large relaxation times and hence slow injection rates, which did not produce a steady state mass loss rate within a Hubble time; for the same reason, the equilibrium rate would in any case be quite small.

Figures 5.19-5.22 exhibit the results for model 4, consisting of an AGN-class black hole and an isotropic cluster of equal solar mass stars. This model is identical to model 2 (Figs. 5.9-5.13) except for the injection of new stars. As seen in Fig.

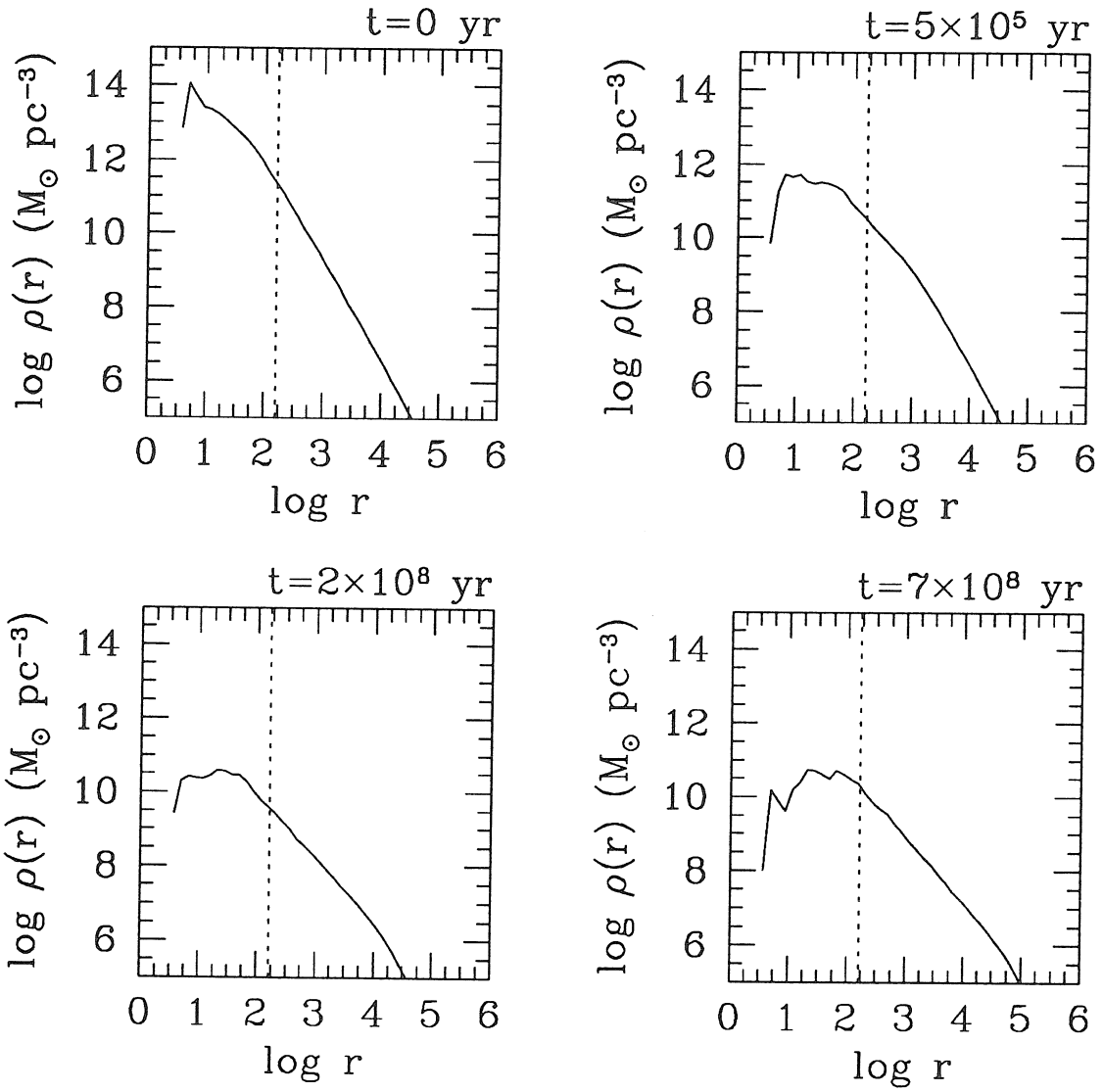

Fig. 5.20: Density profiles for model 4; cf. Fig. 5.10 (profiles for model 2), which is for the same model as here only without injection of new stars. The completeness limit is again $r_{\rm sc} \approx 160 \approx 2 \times 10^{-4}$ pc.

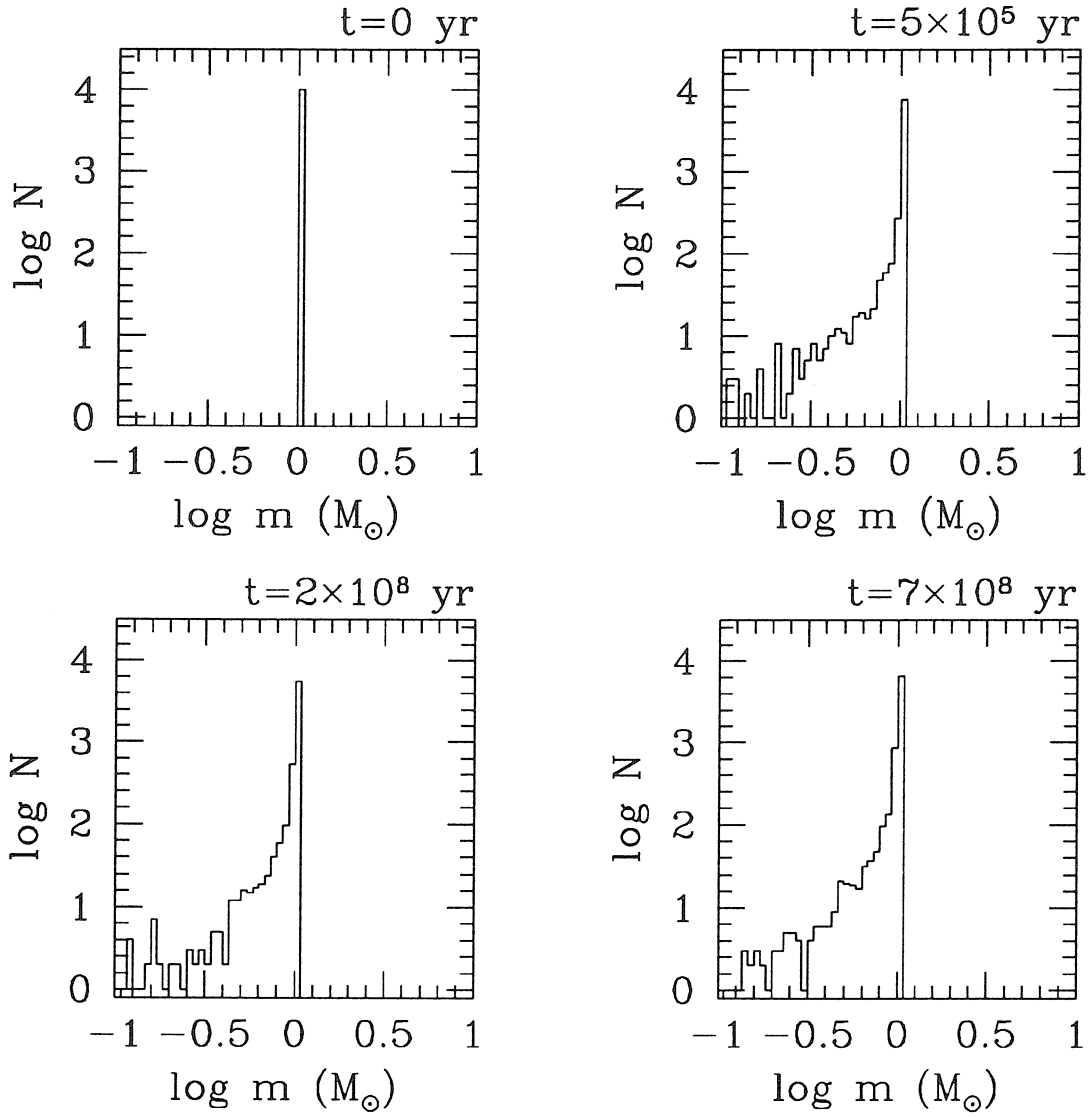

Fig. 5.21: Mass spectrum evolution for model 4; cf. Fig. 5.11.

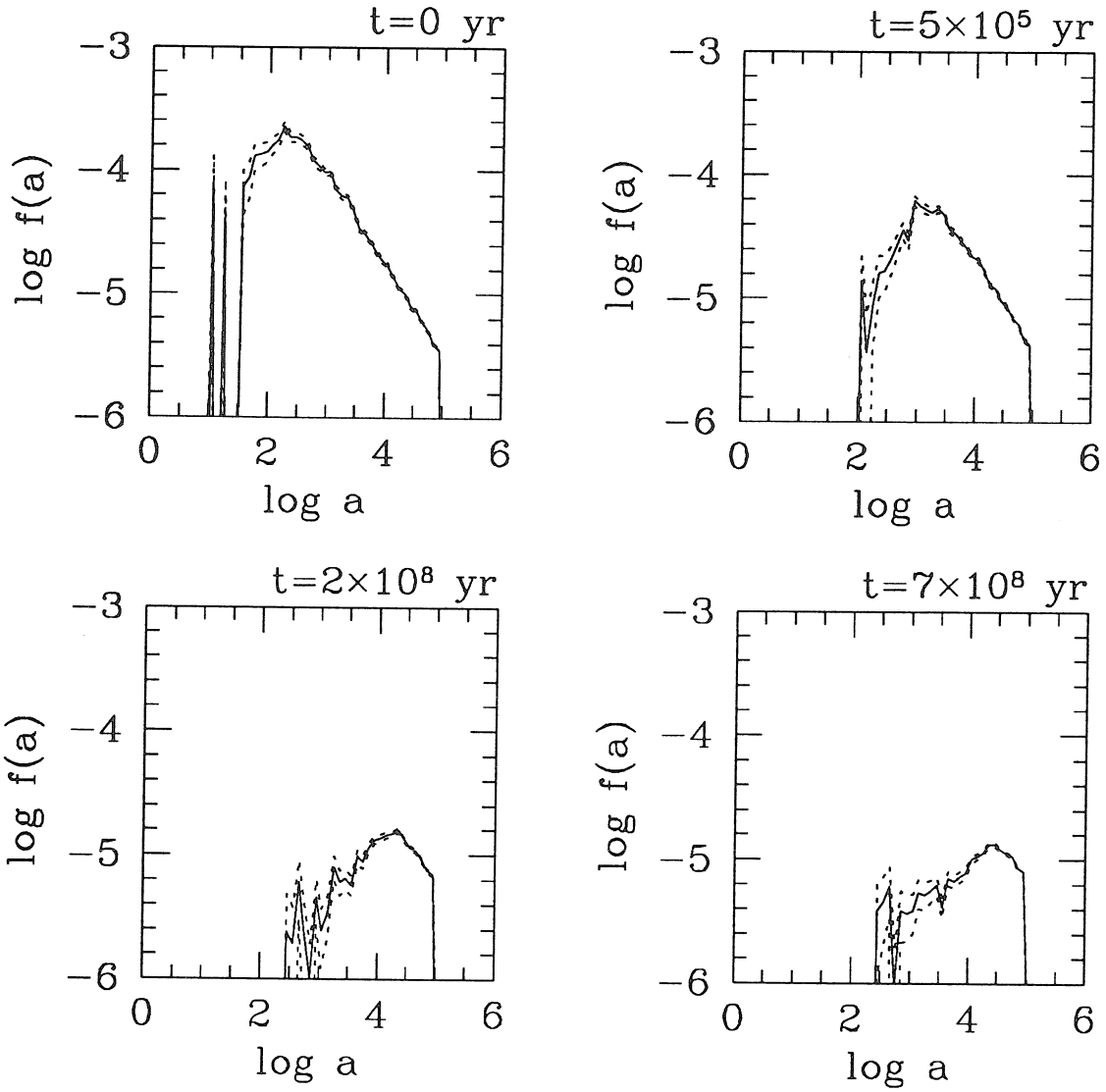

Fig. 5.22: Semi-major axis distributions for model 4; cf. Fig 5.12.

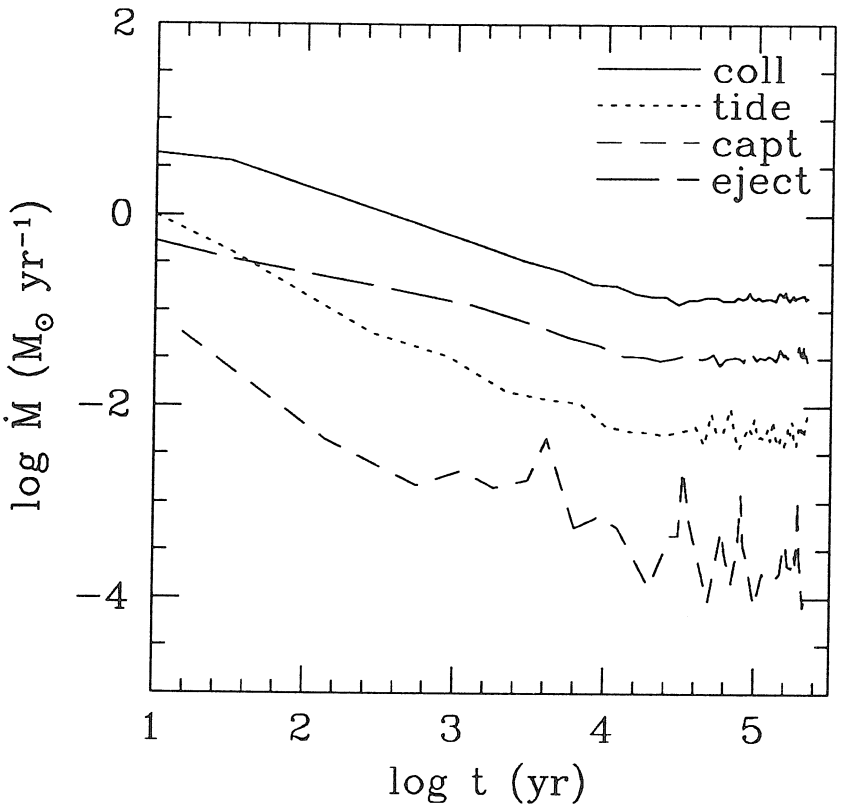

Fig. 5.23: Mass liberation rates for model 5. The smaller hole mass ($M_8 = 0.01$) and higher densities ($n_6 = 10$) compared with model 4 (with $M_8 = 0.3$ and $n_6 = 1$) results in a substantially higher equilibrium gas production rate (due to much reduced relaxation times and hence more rapid injection rates), $\dot{M}_{\rm coll} \approx 0.1 M_\odot$ yr^{-1}, which could in principle support Eddington-limited accretion for $\sim 10^7$ yr until the cluster substantially depletes itself.

5.19, the mass loss rates reach a steady state after $\sim 10^8$ yr; the equilibrium collisional mass loss rate, which dominates all others, is $\dot{M}_{\rm coll} \sim 10^{-5} n_6 M_\odot$ yr^{-1}. The remaining figures show behavior comparable to the corresponding diagrams for model 2; one difference occurs in the mass spectrum (Fig. 5.21). Here continued injection leads also to an equilibrium mass spectrum representing the balance between the fraction of stars in the final state which have or have not suffered significant collisional mass loss, unlike the single injection models in which the

spectrum is increasingly deformed once nearly all stars have collided once or more. In this model the equilibrium low mass tail contained only $\sim 5\%$ of the stars in the simulation.

Model 5 is similar to model 1 except that here the cluster was moderately anisotropic ($p = 10$) and the reservoir was included. The results are given in Figs. 5.23-5.27. Compared with the previous steady state model, the equilibrium mass rates are much higher, $\dot{M}_{\rm coll} \approx 10^{-2} n_6 M_\odot \ {\rm yr}^{-1}$, due mainly to the shorter relaxation times as well as the anisotropy of this cluster, which results in systematically smaller eccentricities of the included stars (Fig. 5.27) and hence shorter orbital periods and larger collision rates; the evolutionary trend, however, is still towards increasingly eccentric orbits. The low mass tail in the equilibrium mass spectrum (Fig. 5.25) is a substantial fraction of the total; in fact, 'equilibrium' is rather a misnomer here since the collision-induced tail continually grows at a slow rate relative to the remaining spectrum even in the steady state models because these low mass fragments have smaller cross-sections and so collide less frequently on average than newly injected stars, and this process is visible upon careful inspection of the figures.

The results for the final model are displayed in Figs. 5.28-5.30, and is similar to model 4 except that a range of stellar masses and a small amount of anisotropy in the cluster are included. The results are comparable and the equilibrium mass loss rates are similar though systematically smaller in this case by a factor of two, which is also the difference in injection rates due to the anisotropy.

5.5. CONCLUSIONS

One of the motivations for these simulations was to search for relativistic influences on cluster evolution around a massive black hole, which is normally treated as a Newtonian point potential in Fokker-Planck calculations. In this regard the results are perhaps disappointing, as their was little evidence for relativistic effects of any type in the results outside of the mass loss rate due to stars on plunge

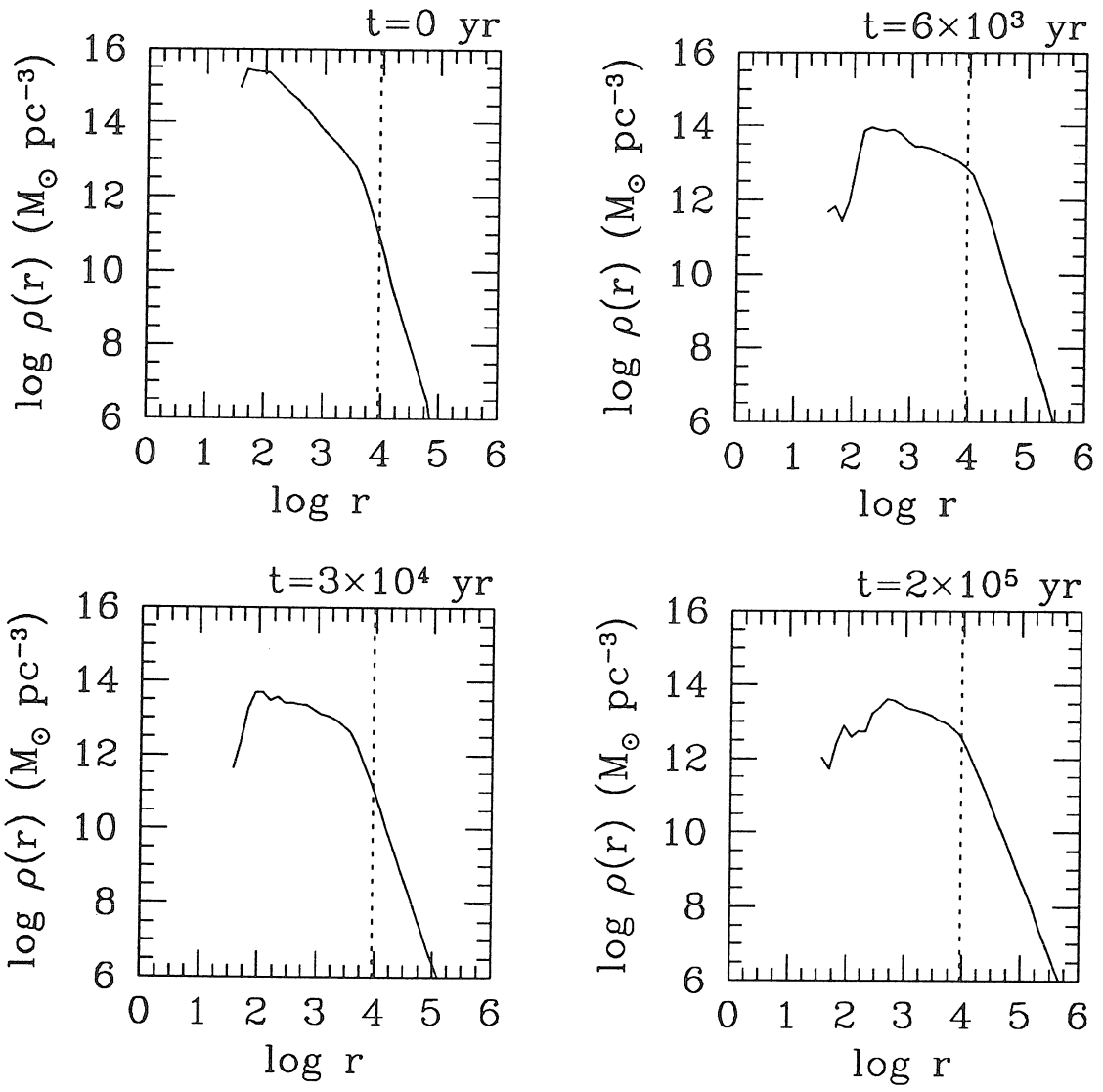

Fig. 5.24: Density evolution for model 5. In $\sim 10^5$ yr a constant density core has already substantially developed. The completeness limit (dashed line) is $r_{sc} \approx 9200 \approx 5 \times 10^{-4}$ pc.

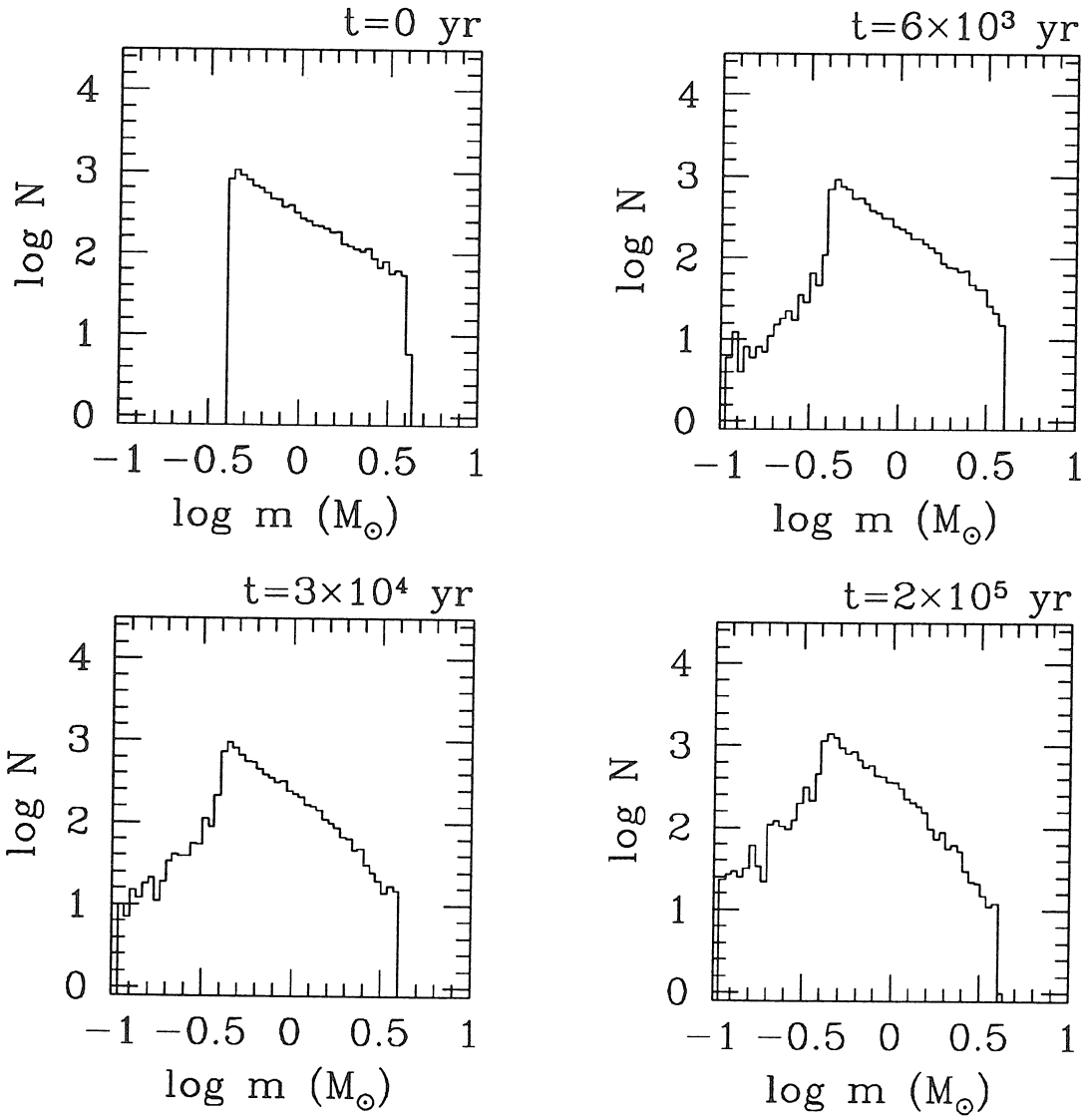

Fig. 5.25: Mass spectrum evolution for model 5; cf. Fig. 5.3.

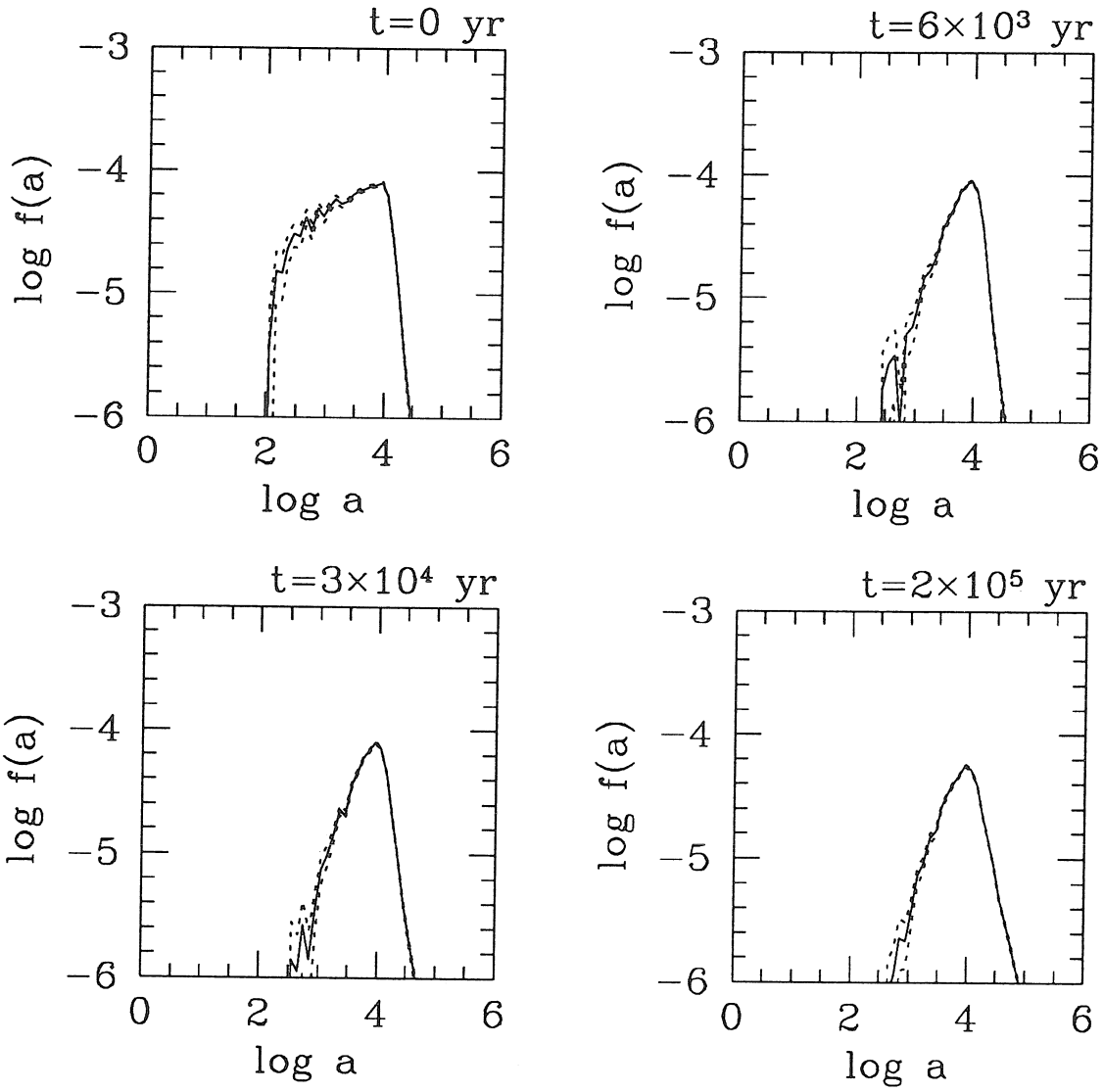

Fig. 5.26: Semi-major axis distributions for model 5.

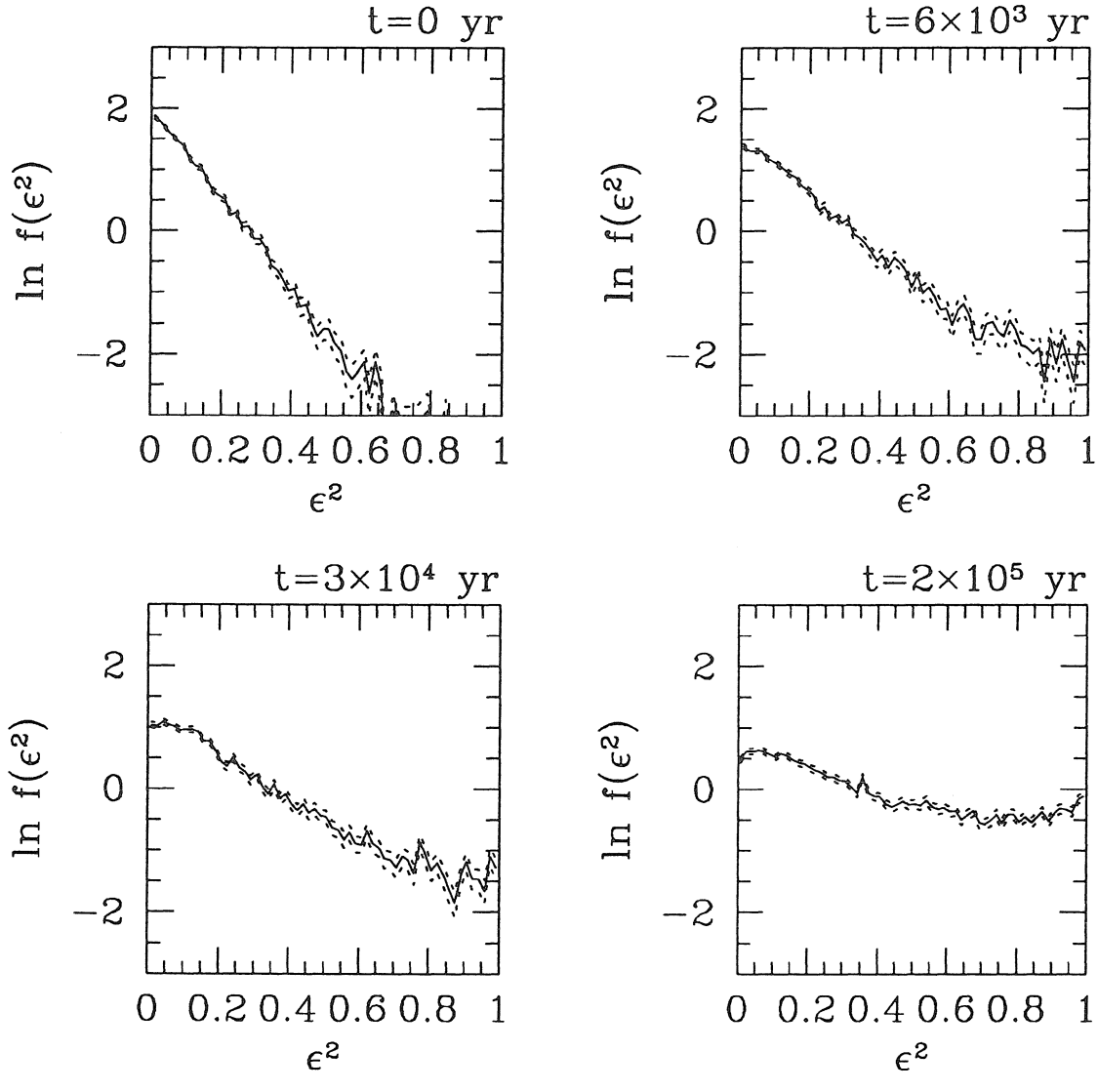

Fig. 5.27: Eccentricity distributions for model 5; the initial cluster in this model is more strongly anisotropic ($p = 10$) than the other models shown, with few very eccentric orbits. Qualitative evolution still proceeds towards smaller effective q-values, and the cluster is nearly isotropic by the time the simulation was stopped.

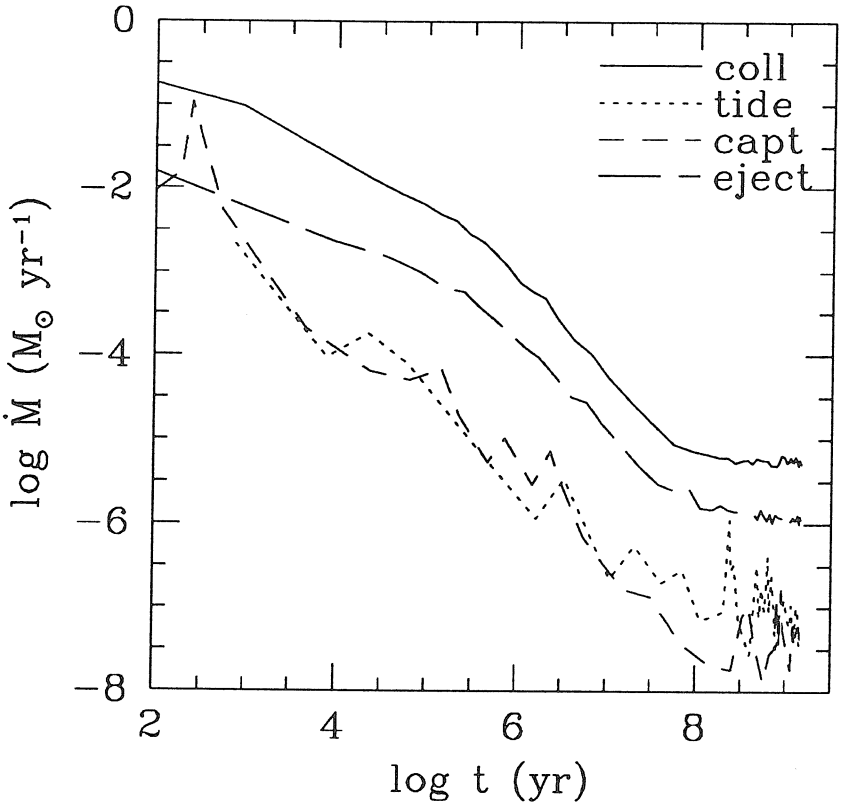

Fig. 5.28: Mass production rates for model 6. This model is similar to model 4 except that a range of stellar masses are used and this model begins with a moderate amount of anisotropy ($p = 5$), which results in a somewhat lower injection rate and equilibrium $\dot{M}_{\rm coll}$; qualitatively the results are similar (cf. Fig. 5.19).

orbits, which was generally small. In particular the spin of the hole did not qualitatively alter any of the results and no evidence of flattening due to the anisotropy of the Kerr metric (or anything else) was found. From a different perspective this is also perhaps to be expected. One reason is that the existence of a tidal radius, in some cases well outside the horizon, prevents stars from approaching the hole too closely (and surviving), where relativistic effects are largest. The main problem is probably one of mortality: whereas many relativistic influences such as precession are secular in nature and most noticeable when allowed to build up over time, in the present situation stars generally suffer complete disruption after only a few collisions, preventing the buildup of any such effects. This does not,

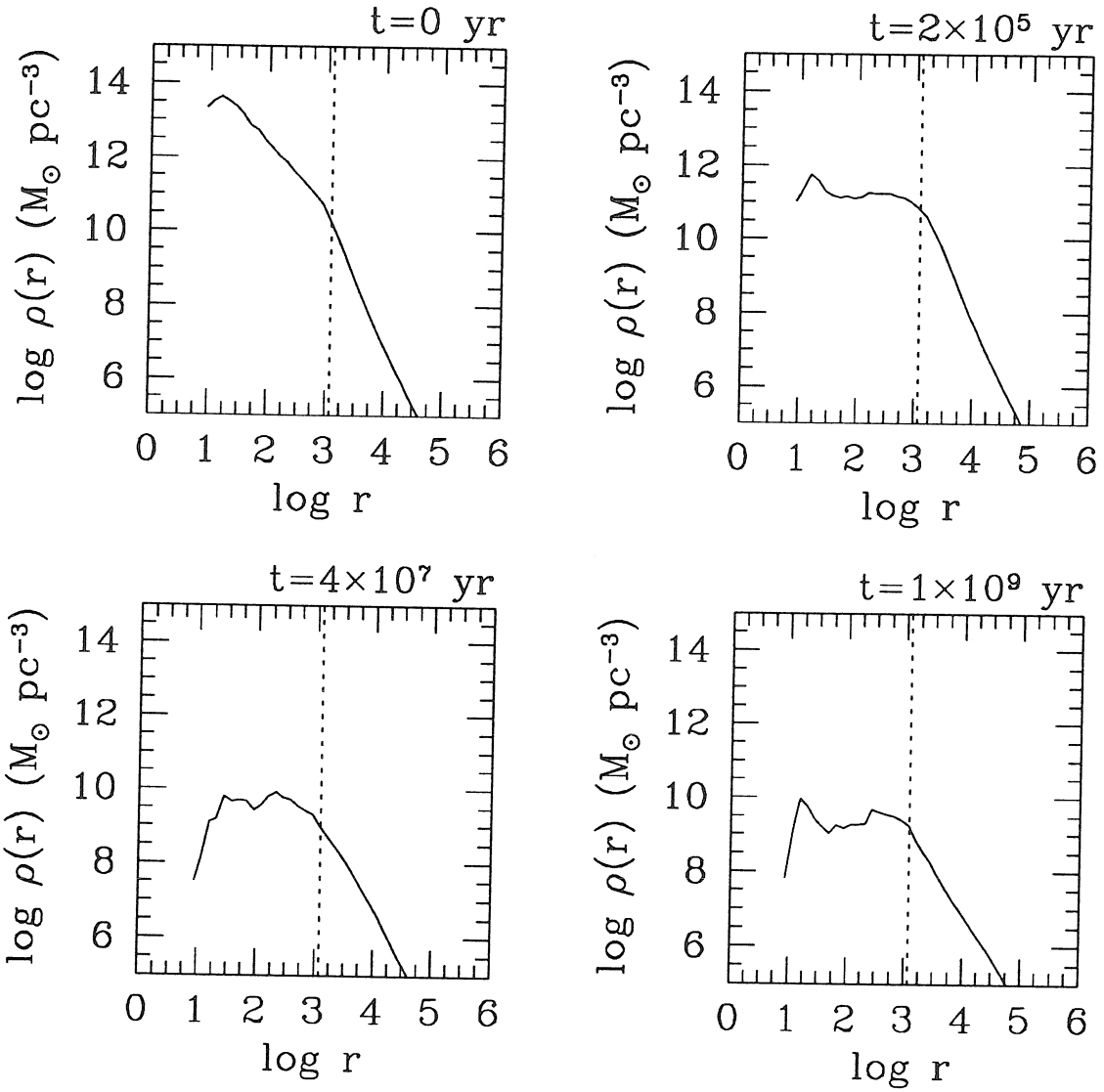

Fig. 5.29: Density evolution for model 6. The completeness limit is $r_{sc} \approx 1200 \approx 3 \times 10^{-3}$ pc.

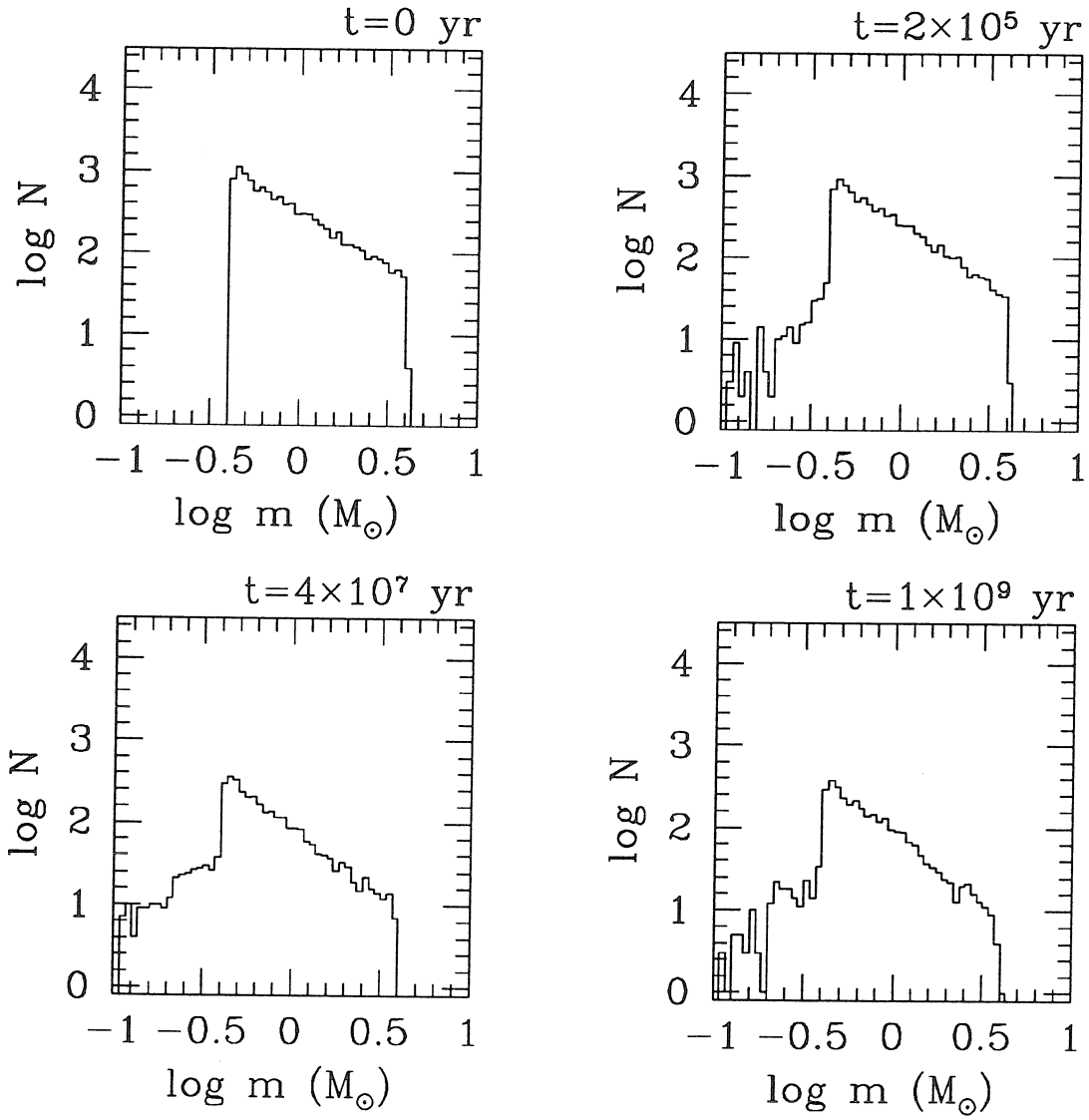

Fig. 5.30: Mass spectrum evolution for model 6.

TABLE 5.3

Model Parameters for Nearby Nuclei

galaxy	M_8	n_6	$\dot{M}_{\rm coll}$ ($M_\odot$ yr^{-1})
GC	0.02	0.3	6×10^{-7}
M31	0.4	0.3?	3×10^{-5}
M32	0.02	0.1	7×10^{-8}
M87	20	1?	3×10^{-4}

however, detract from the promise of the general methods of simulation presented here, as their versatility allows them to be adapted to a range of problems, and only a portion of the method's generality has been exploited in this application. The statistical collisions routine performed well, and its ability to dynamically keep pace with a changing and possibly very anisotropic cluster distribution was also only partially utilized, which simulations including an accretion disk, now in preparation, will make more complete use of. Finally, nearly 'immortal' systems, such as those undergoing relativistic relaxation instead of physical collisions, can in principle develop a long-term memory of previous interactions, allowing the growth of secular effects to take place, and the collision algorithm does not require major modification to simulate such processes.

Putting aside these issues, overall the results presented here are in line with previous investigations of the role of collisions in galactic nuclei. Specifically, these results continue to support the notion that collisions will dominate over tidal processes whenever mass liberation rates approach values capable of sustaining AGN-like luminosities. In a related note, it was found here that the collision-induced tidal disruption rate, which is due to post-collision orbits being such as to cross their new tidal radius, was often larger than the rate produced by relaxation; in effect, collisions can serve as a mechanism to refill the loss cone on timescales characteristic of the collisions instead of the much longer relaxation timescales. It was also found that the rate of ejection of stars from the core normally exceeded

the combined tidal disruption rate, which effectively removes potential energy stores from the black hole's vicinity. On the issue of how in detail will collisions modify the local distribution function of the cluster, these simulations indicate that the main effect is the creation of an essentially constant density core in the collision-dominated cusp around the MBH. This is flatter than the $\propto r^{-1/2}$ power law normally found in corresponding Fokker-Planck calculations (e.g., Duncan & Shapiro 1983; Murphy, Cohn, & Durisen 1989). Although it is unclear to us what the cause of this discrepancy is, it is possible that the small numbers of stars or simplified treatment of collisions used in the Fokker-Planck treatments are becoming inadequate at these radii; this matter requires further investigation. We note, however, that observations of the Galactic Center (GC) favor a flat inner core to the nuclear star cluster instead of a rising cusp (Genzel et al. 1994), although in their discussion of collisional effects in the GC, Genzel et al. do state that an $r^{-3/2}$ form cannot be ruled out with present data. To better put the GC and several other galactic nuclei into context with the simulations, we list in Table 5.3 the approximate model parameters and estimated steady state mass loss rate in the associated model for the GC, M31, M32, and M87.

In velocity space the main influence appears to be a trend towards an increasingly anisotropic core dominated by nearly radial orbits. Results for the modification of the stellar mass function by collisions indicate that a substantial collision-resistant (due to their small cross-sections) low mass tail can be produced, which could in essence 'hide' a significant amount of material in low mass objects. We note that this does not occur if one assumes that stars colliding at high speeds always suffer complete disruption, as has sometimes been assumed—on the contrary, what was found here (cf. Fig. 5.7*b*) is that most collisions, which as a matter of geometry tend to be at grazing incidence rather than nearly head-on, produced rather small fractional mass loss even when a head-on collision results in complete disruption. An issue related to these simulations but outside the scope of this paper concerns the detailed physics of stars colliding at moderately relativistic velocities, as was sometimes the case here (cf. Fig 5.7*a*); namely, there could

be qualitative changes in the manner of shock propagation and subsequent mass loss, for instance, or novel effects on collision-induced nuclear reaction processes with unexpected consequences. As the existence of dense nuclear star clusters and massive black holes together essentially guarantees that such ultra-high velocity collisions will occur, serious examination of this question would appear to be interesting beyond the realm of pure physics.

ACKNOWLEDGEMENTS

Special thanks go to Melvyn Davies, who computed the SPH collision results used in the analysis, and to both him and Roger Blandford for discussions and suggestions on the manuscript. This work was supported by NSF grant AST92-23370.

APPENDIX A

KEPLER'S EQUATION IN THE KERR METRIC

This appendix derives a generalized form of Kepler's Equation valid for quasi-periodic geodesics in the Kerr metric to $O(\delta^{5/2})$, where $\delta = 1/[a(1-e^2)]$ (the inverse of the semi-latus rectum) is assumed to be a small parameter; also given is a compact, exact method for deriving a geodesic's constants of motion given its orbital elements. (Note that we will change notation slightly and denote the eccentricity of an orbit by the usual e instead of ϵ, which was done earlier to avoid all possible confusion with $e = 2.71828\ldots$, which will not be a problem here.) To our knowledge these results have not been published elsewhere. The results given in §A3 should be considered a generalization of the work of Darwin (1961; see also Darwin 1959), who introduced the relativistic anomaly χ, which we will make use of, and essentially derived the modified form of Kepler's Equation valid to $O(\delta)$ in the Schwarzschild geometry, although he did not actually write the result in such a form or refer to it by that name.

A1. EQUATIONS OF MOTION IN THE KERR METRIC

The calculations explained below were done in a modified Boyer-Lindquist coordinate system with coordinates (t, u, μ, ϕ) that are related to the standard Boyer-Lindquist coordinates (t, r, θ, ϕ) by $u \equiv 1/r$ and $\mu \equiv \cos\theta$. As shown in Rauch & Blandford (1994), which also used these coordinates and which should be consulted for further information, the geodesic equations of motion in the modified coordinates take the form

$$\begin{aligned}
p^t &\equiv \frac{dt}{d\lambda'} = \rho^{-2}\left\{-j\left[j(1-\mu^2)-\ell\right] + \frac{(u^{-2}+j^2)}{\Delta}(u^{-2}+j^2-j\ell)\right\} \\
p^\mu &\equiv \frac{d\mu}{d\lambda'} = S_\mu \rho^{-2}\sqrt{M} \\
p^u &\equiv \frac{du}{d\lambda'} = S_u \rho^{-2}\sqrt{U} \\
p^\phi &\equiv \frac{d\phi}{d\lambda'} = \rho^{-2}\left\{-j + \frac{\ell}{1-\mu^2} + \frac{j}{\Delta}(u^{-2}+j^2-j\ell)\right\},
\end{aligned} \tag{5.9}$$

where $U = (1-\gamma^{-2})+2\gamma^{-2}u+\left[j^2(1-\gamma^{-2})-q^2-\ell^2\right]u^2+2\left[(j-\ell)^2+q^2\right]u^3-j^2q^2u^4$, $M = q^2-(\tilde{\jmath}^2+q^2+\ell^2)\mu^2+\tilde{\jmath}^2\mu^4$, $\tilde{\jmath}^2 = j^2(\gamma^{-2}-1)$, $\rho^2 = u^{-2}+j^2\mu^2$, $\Delta = u^{-2}-2/u+j^2$, and $S_u, S_\mu = \pm 1$ are arbitrary signs. The (dimensionless) constants of motion are $\ell \equiv L_z/E$, $q^2 \equiv \mathcal{Q}/E^2$, and $\gamma^{-2} \equiv m^2/E^2$, where L_z, E, m, and Q, the angular momentum component, energy at infinity, rest mass, and Carter's constant, respectively, are the usual dimensional constants of motion. As before, the angular momentum of the hole itself is j. The normalization of the affine parameter λ' is such that $p_t = -1$ and $p_\nu p^\nu = -\gamma^{-2}$ for all particles. All the formulae given in the following two sections were derived by suitable expansion and/or manipulation of the preceding equations of motion.

A2. ORBITAL ELEMENTS AND CONSTANTS OF MOTION

Presented in this section is a simple, exact procedure for computing the constants of motion ℓ, q^2, and γ^{-2} for a generic quasi-periodic orbit, given the corresponding orbital elements a, e, and i; formal definitions of the orbital elements themselves are also given. First recall that quasi-periodic bound orbits occur when the quartic U, whose roots determine the radial turning points of the orbit, possesses four positive real roots (neglecting the special case $j^2q^2 = 0$) and the particle's radial motion is trapped between the two smallest roots (smallest in u coordinates), say $u_{\max}$ and $u_{\min} < u_{\max}$; all such orbits have $\gamma^{-2} > 1$ and $q^2 \geq 0$. For these orbits the polynomial M can be factored into $M = \tilde{\jmath}^2(\mu_-^2-\mu^2)(\mu_+^2-\mu^2)$, where $0 < \mu_+ < 1 < \mu_-$; thus the μ coordinate is trapped between $\pm\mu_+$. We define the orbital elements a and e the same way they would be defined in the non-relativistic case, namely by setting $u_{\max} = 1/[a(1-e)]$ and $u_{\min} = 1/[a(1+e)]$. We see two reasonable ways in which to define the inclination of the orbit; if $j = 0$ the orbits are strictly planar and the definitions become equivalent (and agree with intuition), but for $j > 0$ they are slightly different. The two definitions are $\tan i = \ell^{-1}\sqrt{q^2}$ and $\sin i = \mu_+$, with the additional condition that $i > \pi/2$ when $\ell < 0$. Numerically the two definitions agree to several digits even for highly relativistic orbits ($a(1-e) \sim 10$), so for our purposes it should not matter which one

is used. In this paper the former expression was used as the definition of i because of its simpler relation to the constants of motion.

Having formally defined the orbital elements, the method for computing the constants of motion from these elements—assuming the orbit they describe actually exists—can be derived in the following manner. Define the constant $b = (\ell^2 + q^2)^{1/2}$, which implies that $\ell = b\cos i$ and $q^2 = b^2 \sin^2 i$. By writing U in the form $U = -j^2 q^2 (u - u_1)(u - u_2)(u - u_{\max})(u - u_{\min})$ (where $u_{\max}(a, e)$ and $u_{\min}(a, e)$ are known), expanding, equating the resulting coefficients with the original expression for U given in §A1, and replacing ℓ and q^2 with their expressions in terms of b and i, one obtains a system of equations from which u_1, u_2, and γ^{-2} can be eliminated, resulting in the end in a quadratic equation in b with coefficients depending only on a, e, i, and j, easily solved for b, from which ℓ, q^2, and γ^{-2} can then be computed. Note that the quadratic in b would have been a *quartic* if not for a fortuitous cancellation. The equation determining b is

$$B_2 b^2 - 2B_1 b + B_0 = 0, \tag{5.10}$$

where

$$\begin{aligned}
B_0 &= 1 + j^2\delta^2\left[2(1+e^2) - 4\delta(1-e^2) + j^2\delta^2(1-e^2)^2\right], \\
B_1 &= j\delta^2\cos i\left[(3+e^2) - 4\delta(1-e^2) + j^2\delta^2(1-e^2)^2\right], \\
B_2 &= -\delta\Big\{\left[1-2\delta(1-e)\right]\left[1-2\delta(1+e)\right] + j^2\delta^2\left[-\delta(1-e^2)^2\right. \\
&\qquad \left. + \sin^2 i\big(2(1+e^2) - \delta(3-2e^2-e^4) + j^2\delta^2(1-e^2)^2\big)\right]\Big\},
\end{aligned}$$

$\delta \equiv 1/[a(1-e^2)]$, and the positive root is to be taken. This procedure will always succeed even if the input orbit is invalid (e.g., if its traversal would require crossing the horizon), but in this latter case the values of a and e derived from the actual roots of U will not be the same as those originally given the procedure, so a consistency check should be made for orbits passing close to the horizon ($\delta \gtrsim 0.1$). For $\delta \ll 1$ we have

$$\begin{aligned}
b = \delta^{-1/2}\big\{1 + 2\delta - j\cos i(3+e^2)\delta^{3/2} + \left[2(2+e^2) + j^2\cos^2 i(1+e^2)\right]\delta^2 \\
- 8j\cos i(1+e^2)\delta^{5/2} + O(\delta^3)\big\}.
\end{aligned}$$

The expressions for the constants of motion are $\ell = b\cos i$, $q^2 = b^2 \sin^2 i$, and

$$\gamma^{-2} = \frac{1 - \delta^3(1-e^2)^2\left[(1 - j^2\delta)q^2 + (j-\ell)^2\right]}{1 - \delta(1-e^2)}. \tag{5.11}$$

For $\delta \ll 1$ we have

$$\gamma^{-2} = 1 + (1-e^2)\delta + (1-e^2)^2\delta^2 + (1-e^2)^2\left[1 - e^2 - (j-\ell)^2 - q^2\right]\delta^3 + O(\delta^4).$$

A3. THE MODIFIED KEPLER'S EQUATION

Following Darwin (1961), define a relativistic anomaly χ which is related to the particle's radial coordinate u by

$$u(\chi) = \delta(1 + e\cos\chi), \tag{5.12}$$

so that periapse and apoapse always occur at $\chi = 0$ and $\chi = \pi$, respectively, regardless of the amount of precession the orbit is undergoing. For nearly Newtonian orbits, $\delta \ll 1$, χ will be almost the same as the true anomaly ν familiar from celestial mechanics, and hence its time evolution will be closely approximated by Kepler's Equation,

$$\mathcal{M} = \psi - e\sin\psi, \tag{5.13}$$

where the mean anomaly $\mathcal{M} = 2\pi t/t_{\rm orb} = t/a^{3/2}$ and the eccentric anomaly ψ is related to the true anomaly by $(1 + e\cos\nu)(1 - e\cos\psi) = 1 - e^2$. Clearly, by suitably expanding the equations of motion in powers of $\delta \ll 1$, the relativistic corrections to Kepler's Equation can be found; here they are given to $O(\delta^{5/2})$.

It is convenient to define a variable ζ according to

$$\zeta(\chi) = \int_{u(\chi)}^{u_{\rm max}} \left[j^2\mu_+^2(\gamma^{-2} - 1)\right]^{1/2} \frac{du'}{\sqrt{U(u')}}, \tag{5.14}$$

which upon expansion in δ becomes

$$\begin{aligned}\zeta(\chi) = \big\{1 + 3\delta - 6j\cos i\,\delta^{3/2} + [54 + 3e^2 + j^2\cos^2 i\,(14 + e^2) - j^2(2 + e^2)]\delta^2/4 \\ - j\cos i\,(42 + 3e^2)\delta^{5/2}\big\}\chi + \big\{1 + [9 + j^2(2\cos^2 i - 1) - 2j\cos i\,\delta^{1/2} \\ + e\cos\chi(3 - j^2\sin^2 i)/4]\delta - j\cos i\,(32 + 3e\cos\chi)\delta^{3/2}\big\}\delta e\sin\chi + O(\delta^3).\end{aligned}$$

By inverting the integral relation $\int^{\mu} d\mu'/\sqrt{M} = \int^{u} du'/\sqrt{U}$ to find μ, one obtains the *exact* solution

$$\mu(\zeta) = \mu_+ \,\mathrm{sn}\big(\zeta_0 + \zeta \,\big|\, (\mu_+/\mu_-)^2\big), \tag{5.15}$$

where $\mathrm{sn}(x \,|\, m)$ is a Jacobian elliptic function with parameter m and the phase ζ_0 is akin to the argument of periapse. Note that since

$$\mu_+ = \sin i \left[1 - j^2 \cos^2 i\,(1-e^2)\delta^2/2\right] + O(\delta^3),$$
$$\left(\frac{\mu_+}{\mu_-}\right)^2 = j^2 \sin^2 i\,(1-e^2)\delta^2 + O(\delta^3),$$

we have $\mu(\zeta) = \sin i\, \sin(\zeta_0 + \zeta) + O(\sin i\, \delta^2)$. Using this lower order expansion for μ when calculating the expansion for $\phi(\chi) = \int_0^{\chi} (d\phi/du)(du/d\chi) d\chi'$ gives

$$\phi(\chi) = \phi_0 + \tan^{-1}(\cos i\, \tan\zeta) + \Delta\phi(\chi) + O(\sin i\, \delta^2), \tag{5.16}$$

where ϕ_0 is akin to the argument of the ascending node and

$$\Delta\phi(\chi) = \Big\{ \big[4 + (2+e^2)\big(-j\cos i\, \delta^{1/2} + (6+j^2)\delta\big)\big]\chi$$
$$+ \big[4 + (4 + e\cos\chi)\big(-j\cos i\, \delta^{1/2} + (6+j^2)\delta\big)\big] e\sin\chi \Big\} \frac{j}{2}\, \delta^{3/2} + O(\delta^3).$$

In the mean plane of the orbit, which for a given periapse-to-periapse orbital segment was defined as the plane spanned by the position and velocity vectors at apoapse, the effective true anomaly is $\nu(\chi) \approx \zeta(\chi) + \cos i\, \Delta\phi(\chi)$, and thus the precession in the mean orbital plane per orbit is given by

$$\Delta\nu_{\mathrm{p}} = \nu(2\pi) - 2\pi = 2\pi(3\delta - 4j\cos i\, \delta^{3/2}) + O(\delta^2). \tag{5.17}$$

The first term, $6\pi\delta = 6\pi/[a(1-e^2)]$, is the well-known lowest order Schwarzschild value of the precession and the second represents the lowest order Kerr correction.

The concept of a (fixed) mean plane was found to be useful down to radii $r \sim 10$ (for $j \sim 1$), below which the perpendicular deviation of the true orbit from the mean plane over a single orbital period can be of order the radius itself. Since the errors in all of these formulae are secular in nature, to avoid accumulation

of errors, the collision finder algorithm (§5.3.3) used the accurate Kerr geodesic routines to reset ζ_0 and ϕ_0 to their optimum values for each orbital segment.

Finally, the modified form of Kepler's Equation follows from expansion and integration of $dt/d\psi = (dt/du)(du/d\chi)(d\chi/d\psi)$, where ψ is now taken to be the eccentric anomaly corresponding to χ, i.e., $(1+e\cos\chi)(1-e\cos\psi) = 1-e^2$. Note that since the μ^2 term in dt/du enters at $O(\delta^2)$, a solution accurate to $O(\delta^{5/2})$ need only use $\mu(\chi) \approx \sin i \, \sin(\zeta_0+\chi)$ during the derivation. The result for $t(\chi)$ is

$$\begin{aligned}\frac{t(\chi)-t_0}{a^{3/2}} = &\Big\{1+3(1-e^2)\delta\big[1-j\cos i\,\delta^{1/2}[1+(7-e^2)\delta] \\ &\quad +\big(2+\tfrac{1}{2}j^2\cos^2 i\big)\delta\big]\Big\}\psi \\ &-\Big\{1+(1-e^2)\delta\big[-j\cos i\,\delta^{1/2}[1+2(2-e\cos\chi)\delta] \\ &\quad +\big(2+\tfrac{1}{2}j^2\cos^2 i\big)\delta\big]\Big\}e\sin\psi \\ &+\frac{1}{2}\Big\{[15-24j\cos i\,\delta^{1/2}]\chi - j^2\sin^2 i\,\cos(2\zeta_0+\chi)\,\sin\chi\Big\}(1-e^2)^{3/2}\delta^2.\end{aligned} \tag{5.18}$$

The value of the orbital period is simply $t_{\rm orb} = t(2\pi) - t_0$ which in addition to a is seen to depend on both e and ζ_0. By rewriting the expansion using $\mathcal{M} = 2\pi t/t_{\rm orb}$, one arrives at the modified Kepler's Equation with errors $O(\delta^3)$,

$$\mathcal{M} = \psi - \tilde{e}\sin\psi + \Delta\psi, \tag{5.19}$$

$$\begin{aligned}\tilde{e} = &\Big\{1-(1-e^2)\delta\Big[3-2j\cos i\,\delta^{1/2}\big\{1+\big(1+6e^2+6(1-e^2)^{1/2}+e\cos\chi\big)\delta\big\} \\ &\quad +\tfrac{1}{2}\big(-10+18e^2+15(1-e^2)^{1/2}+2j^2\cos^2 i\big)\Big]\Big\}e,\end{aligned}$$

$$\Delta\psi = \frac{1}{2}\Big\{[15-24j\cos i\,\delta^{1/2}](\chi-\psi) - j^2\sin^2 i\,\cos(2\zeta_0+\chi)\,\sin\chi\Big\}(1-e^2)^{3/2}\delta^2.$$

Because $0 \le \tilde{e} \le e$ and $\Delta\psi \sim O(\delta^2)$, the inverse solution $\psi(\mathcal{M})$ can be computed very economically by first finding the solution ψ_0 corresponding to $\Delta\psi \to 0$ and then correcting ψ_0 treating $\Delta\psi$ as a small perturbation, $\psi \approx \psi_0 - \Delta\psi_0/(1-\tilde{e}\cos\psi_0)$. It was found that the relative error in the solution obtained in this way was normally no greater than that inherent in the expansion itself, except perhaps at larger distances, where the absolute error was still quite small. It was

also found empirically that excluding the term involving ζ_0 in $\Delta\psi$ and $\cos\chi$ in $\tilde{e}$ did not noticeably worsen the error in ψ compared to its true value, allowing an additional gain in efficiency. This allowed the inverse solution $\psi(\mathcal{M})$ to be computed with little more work than is needed in the Newtonian case, greatly improving the performance of the collision finding procedure.

APPENDIX B

COLLISION PRODUCTS FITTING FORMULAE

The outcome of a collision between a pair of stars, characterized by the mass loss, angular deflection, and change in speed and radius suffered by each, was computed (except for the new radius, see below) from fitting formulae derived from the results of SPH simulations of high velocity stellar collisions performed for this work by Davies (1994). The fits to each of the derived quantities depended on three dimensionless variables describing the circumstances of the collision: the mass ratio, q, the relative velocity of the stars, $\beta_{\rm rel}$, and the impact parameter in units of the sum of the stellar radii, $\tilde{b} = b/(R_{\star,1} + R_{\star,2})$, each of which ranges from 0 to 1. The relative speed as well as the vector changes in each star's velocity were computed in the orthonormal center-of-mass frame, each star's 4-velocity being transformed from the Boyer-Lindquist coordinate frame to the center-of-mass frame and back using the expressions appropriate for the Kerr metric at the point of collision.

Useful fitting functions could not be obtained for the change in radii because of noise limitations in the hydrodynamic results as well as physical uncertainties about the post-encounter equilibrium state of the star. Additionally, the post-encounter radius is likely to vary (shrink) with time as the star sheds any excess thermal energy acquired during the collision, which will occur on a Kelvin-Helmholtz timescale ($\sim 10^7$ years for solar type stars). Because of these inherent uncertainties, the new radius was crudely approximated as being equal to that of a main sequence star having the same mass as the surviving star. Although the actual surviving fragment will probably be somewhat larger than this initially, those outer layers will also be less dense and more loosely bound than an unperturbed main-sequence envelope, leading to less deflection and mass loss during glancing collisions. Hence it may not be so unreasonable to use the smaller radius in subsequent calculations, since the collisions resulting from the additional cross

section presented by the outer layers might do rather little further damage to the star anyway.

Most of the SPH simulations used $q = 1$ to determine basic dependence of the derived values on $\beta_{\rm rel}$ and $\tilde{b}$; estimation of the q dependence of the results was based on a smaller number of simulations in which models originally using $q = 1$ were redone using $q = 1/2$ and $q = 1/4$. The stars were taken to be $n = 3$ polytropes. The fitting formulae are (it is assumed $q = m_2/m_1 < 1$)

$$\frac{\delta m}{m} = \frac{\delta m_1 + \delta m_2}{m_1 + m_2} \approx q^{0.2}\beta_{\rm rel}^{1.25} e^{P_1(\tilde{b})}, \quad \frac{\delta m_1/m_1}{\delta m_2/m_2} \approx q^{0.9},$$
$$\delta\theta_1 \approx \delta\theta_2 \approx \left[\pi - 2/\left(P_2(\tilde{b}) + P_3(\tilde{b})\,\beta_{\rm rel}^{-2}\right)\right] q^{-0.3},$$
$$\frac{\delta v_1}{v_1} \approx \frac{\delta v_2}{v_2} \approx \beta_{\rm rel}^{-2.75} e^{P_4(\tilde{b})} + \beta_{\rm rel}^{-0.5} e^{P_5(\tilde{b})},$$

where $P_1(\tilde{b}) = 5.5530 - 2.8607\tilde{b} - 10.5255\tilde{b}^2$, $P_2(\tilde{b}) = 0.57366 + 0.18251\tilde{b} - 0.13155\tilde{b}^2$, $P_3(\tilde{b}) = 10^{-6}(3.057 - 4.825\tilde{b} + 2.616\tilde{b}^2)$, $P_4(\tilde{b}) = -16.138 - 7.26\tilde{b}$, and $P_5(\tilde{b}) = 0.417 - 15.96\tilde{b}$. These fits reproduce the SPH results to within errors $\sim 10\%$; the true accuracy of the formulae as applied in the simulations described in this paper is not so easily quantified, most notably because of the need to extrapolate the results to relativistic impact velocities $\beta_{\rm rel} \sim 0.1 - 0.5$, the maximum velocity for which SPH results were obtained being $\beta_{\rm rel} \sim 0.05$. However, since it is clearly not possible to compute detailed encounter models for each of the thousands of collisions occurring over the course of the simulations, some standardized strategy is necessary, and the use of hydrodynamics simulations to provide a quantitative foundation for calculating the collision products allowed as much microphysics as practically feasible to be injected into this part of the simulations.

REFERENCES

Artymowicz, P., Lin, D. N. C., & Wampler, E. J. 1993, ApJ, 409, 592

Bacon, R., Emsellem, E., Monnet, G., & Nieto, J. L. 1994, A&A, 281, 691

Bahcall, J. N., & Wolf, R. A. 1976, ApJ, 209, 214

Bardeen, J. M. 1970, Nature, 226, 64

Bardeen, J. M., & Petterson, J. A. 1975, ApJ, 195, L65

Benz, W., & Hills, J. G. 1987, ApJ, 323, 614

Blandford, R. D. 1987, in Hawking, S. W., Israel, W., eds, 300 Years of Gravitation (Cambridge: Cambridge Univ. Press), p. 277

Blandford, R. D., & Znajek, R. L. 1977, MNRAS, 179, 433

Colgate, S. A. 1967, ApJ, 150, 163

Darwin, C. 1959, Proc. Roy. Soc. Lond. A, 249, 180

———. 1961, Proc. Roy. Soc. Lond. A, 263, 39

David, L. P., & Durisen, R. H. 1989, ApJ, 346, 618

David, L. P., Durisen, R. H., & Cohn, H. N. 1987a, ApJ, 313, 556

———. 1987b, ApJ, 316, 505

Davies, M. B. 1994, in preparation

Davies, M. B., Benz, W., & Hills, J. G. 1991, ApJ, 381, 449

———. 1992, ApJ, 401, 246

Duncan, M. J., & Shapiro, S. L. 1983, ApJ, 268, 565

Frank, J., & Rees, M. J. 1976, MNRAS, 176, 633

Genzel, R., Hollenbach, D., & Townes, C. H. 1994, Rep. Prog. Phys, 57, 417

Harms, R. J. et al. 1994, ApJ, in press

Hills, J. G. 1975, Nature, 254, 295

Hoyle, F., Fowler, W. A., Burbidge, G. R., & Burbidge, E. M. 1964, ApJ, 139, 909

Karas, V., & Vokrouhlický, D., 1993, MNRAS, 265, 365

Kormendy, J. 1994, in Harris A., Genzel, R., eds, The Nuclei of Normal Galaxies: Lessons from the Galactic Center

Kumar, S., & Pringle, J. E. 1985, MNRAS, 213, 435

Lai, D., Rasio, F. A., & Shaprio, S. L. 1993, ApJ, 412, 593

Lauer, T. R. 1989, in Merritt, D., ed, Dynamics of Dense Stellar Systems (Cambridge: Cambridge Univ. Press), p. 3

Lauer, T. R., Faber, S. M., Currie, D. G., & Ewald, S. P. 1992, AJ, 104, 552

Lee, H. M., & Nelson, L. A. 1988, ApJ, 334, 688

Lee, M. H. 1993, ApJ, 418, 147

Lightman, A. P., & Shapiro, S. L. 1977, ApJ, 211, 244

Misner, C. W., Thorne, K. S., & Wheeler, J. A. 1973, Gravitation (San Francisco: W. H. Freeman Co.), Chapter 33

Murphy, B. W., Cohn, H. N., & Durisen, R. H. 1989, in Merritt, D., ed, Dynamics of Dense Stellar Systems (Cambridge: Cambridge Univ. Press)

Murphy, B. W., Cohn, H. N., & Durisen, R. H. 1991, 370, 60

Penrose, R. 1969, Rev. Nuovo. Cimento, 1, 252

Pringle, J. E. 1992, MNRAS, 258, 811

Quinlan, G. D., & Shapiro, S. L. 1989, ApJ, 343, 725

———. 1990, ApJ, 356, 483

Rauch, K. P. 1994, MNRAS, submitted

Rauch, K. P., & Blandford, R. D. 1994, ApJ, 421, 46

Rees, M. J. 1984, ARA&A, 22, 471

Richstone, D. O., Bower, G., & Dressler, A., 1990, ApJ, 353, 118

Sanders, R. H. 1970, ApJ, 162, 791

Spitzer, L., & Saslaw, W. C. 1966, ApJ, 143, 400

Spitzer, L., & Stone, M. E. 1967, ApJ, 147, 519

Syer, D., Clarke, C. J., & Rees, M. J. 1991, MNRAS, 250, 505

Thorne, K. S. 1974, ApJ, 191, 507