J. Astrophys. Astr. (1999) **20**, 1–22

Understanding the Spectral Energy Distributions of the Galactic Star Forming Regions IRAS 18314–0720, 18355–0532 and 18316–0602

Bhaswati Mookerjea & S. K. Ghosh, *Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai (Bombay) 400 005, India*

Received 1998 August 5; accepted 1999 June 17

Abstract. Embedded Young Stellar Objects (YSO) in dense interstellar clouds are treated self-consistently to understand their spectral energy distributions (SED). Radiative transfer calculations in spherical geometry involving the dust as well as the gas component, have been carried out to explain observations covering a wide spectral range encompassing near-infrared to radio continuum wavelengths. Various geometric and physical details of the YSOs are determined from this modelling scheme.

In order to assess the effectiveness of this self-consistent scheme, three young Galactic star forming regions associated with IRAS 18314–0720, 18355–0532 and 18316–0602 have been modelled as test cases. They cover a large range of luminosity (≈ 40). The modelling of their SEDs has led to information about various details of these sources, e.g. embedded energy source, cloud structure and size, density distribution, composition and abundance of dust grains etc. In all three cases, the best fit model corresponds to the uniform density distribution. Two types of dust have been considered, viz., Draine & Lee (DL) and the Mezger, Mathis & Panagia (MMP). Models with MMP type dust explain the dust continuum and radio continuum emission from IRAS 18314–0720 and 18355–0532 self-consistently. These models predict much lower intensities for the fine structure lines of ionized heavy elements, than those observed for IRAS 18314–0720 and 18355–0532. This discrepancy has been resolved by invoking clumpiness in the interstellar medium. For IRAS 18316–0602, the model with DL type dust grains is preferred.

Key words. Infrared SED – HII regions – radiative transfer – IRAS 18314–0720 – IRAS 18355–0532 – IRAS 18316–0602.

1. Introduction

The formation and the initial stages of evolution of stars take place inside dense regions within molecular clouds, from which they are born. Hence, the star formation studies have necessarily to deal with the protostars/young stars in the environment of interstellar gas and dust from the parent cloud. The stars with sufficient supply of Lyman continuum photons (basically depending on their mass) create and maintain HII regions around them. The (ultra)/compact variety of HII regions in particular provides a natural test case for better understanding of medium/high mass star formation. They present two main advantages for such studies: being younger they

allow one to probe the physical processes closer in time to the formation of the embedded star, and secondly they have simpler geometrical shapes (e.g. with spherical/ cylindrical symmetry), thus providing opportunity for direct comparison between the observations and the predictions of detailed numerical models. In addition, being of the higher luminosity class (roughly between ZAMS O4 and B0.5), observationally they can be studied over larger distances (almost any part of the Galaxy) in the radio continuum as well as infrared/sub-mm. Available observational data regarding continuum emission from Galactic compact HII regions span a wide spectral range covering near-infrared (NIR) to sub-millimeter wavelengths. Whereas the atmospheric windows in NIR, sub-millimeter and a few restricted ones in the mid-IR (MIR) have been used extensively, the peak of the spectral energy distribution (SED) lies invariably in the far-IR (FIR) waveband which is inaccessible from the earth-based observatories. The work horse for the high angular resolution FIR measurements of Galactic star forming regions has been the Kuiper Airborne Observatory, occasionally supplemented by balloon-borne telescopes. Although the InfraRed Astronomy Satellite (IRAS) offers the only near all sky and most complete data base, its poorer angular resolution limits its usefulness, particularly for the study of compact HII regions. However, this situation is about to change drastically with the advent of the European Space Agency's Infrared Space Observatory (ISO) mission. ISO with improved photometric sensitivity as well as spectral and angular resolution, will open up enormous opportunity to perform detailed studies of Galactic star forming regions in the entire infrared waveband encompassing Near through Mid to Far Infrared. Most of the infrared continuum emission from the embedded YSOs originating from the interstellar dust component will be measured throughout the 2.5–200 μm region with greater photometric precision and with diffraction limited angular resolution, by the ISO instruments. In addition, the spectroscopic data will help understand the composition as well as physical parameters of the interstellar gas component.

Keeping the above in mind, the present study is a step in trying to establish a self-consistent yet simple scheme of modelling Galactic compact HII regions in spherically symmetric geometry. Here, by self-consistent we mean that the same geometric and physical configuration fits the observed data for the emission from the dust (most of the infrared, sub-mm, mm part of the SED) as well as the emission from the gas (radio continuum, fine structure line etc). The main aim is to extract maximum possible information regarding the geometric and physical details of the star forming region from the observed SED. In addition to predicting the continuum infrared emission from the dust, the gas component has also been integrated in a self consistent manner, thereby predicting the absolute and relative strengths of atomic/nebular lines as well as the radio continuum emission.

In order to assess the usefulness of the above scheme, three Galactic embedded YSOs (IRAS 18314–0720, 18355–0532 and 18316–0602; most likely in compact HII region phase, have been selected for modelling. These sources have been chosen based on availability of relevant observational data; and also to cover a wide range of total luminosity (a factor of ≈ 40). These three sources IRAS 18314–0720, 18355–0532 and 18316–0602 have luminosities 1.02×10^6, 1.4×10^5 and $2.5 \times 10^4 L_\odot$ respectively (which correspond to single ZAMS stars of type O4, O6.5 and B0). Although at this point of time the available amount of validated ISO data in public domain is rather limited, much more sophisticated application of this scheme is anticipated in the future.

The outline of this paper is as follows. Section 2 describes the modelling scheme including the treatments of radiative transfer in the dust and the gas components respectively. Section 3 presents the observational constraints of the three sources, and the results of modelling in the form of geometrical and physical information extracted about these sources. Conclusions are summarized in section 4.

2. The modelling scheme

A compact HII region is modelled as a spherically symmetric cloud (made of typical interstellar gas-dust material), powered by centrally embedded single or a cluster of zero age main sequence (ZAMS) star/(s). This cloud is assumed to be immersed in an isotropic radiation field (typical Interstellar Radiation Field, ISRF). The interstellar gas and the dust is assumed to follow the same radial density distribution law, but with the following difference—whereas the gas exists throughout the cloud (i.e. right from the stellar surface up to the outer boundary of the cloud, R_{max}; see Fig. 1), there is a natural lower limit to the inner boundary, R_{min}, for the dust distribution (i.e. a cavity in the dust cloud). This is because the dust grains are destroyed when exposed to excessive radiative heating. The gas to dust ratio, where they co-exist ($R_{min} < r < R_{max}$), is assumed to be constant. The position of the ionization front (R_{HII}, refer to Fig. 1) depends on the effective temperature and luminosity of

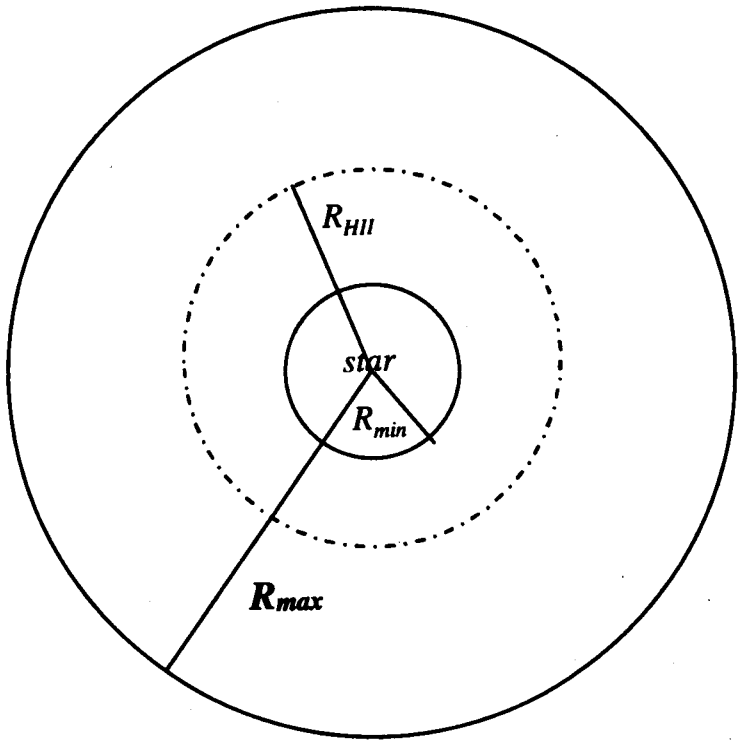

Figure 1. Schematic diagram of the model star forming region.

the exciting star, as well as the density of the gas. The case, $R_{\rm HII} < R_{\rm min}$ is also possible, if either the star is not hot enough and/or the density of gas around the star is quite high.

Modelling a specific compact HII region involves matching the predicted emergent spectral (continuum) shape with the measured SED and comparing the relative and absolute strengths of the atomic/nebular spectral lines from the gas component with spectroscopic observations (if available).

The following are the inputs for individual runs of the modelling scheme:

(i) the total luminosity;
(ii) the spectral shape of the radiation emerging from the embedded source/(s) (sensitive to the assumed Initial Mass Function and upper mass cut off, in case of a star cluster);
(iii) the gas to dust ratio;
(iv) the properties of the dust grains (of each type considered);
(v) the ISRF incident at the outer boundary;
(vi) the elemental abundances in the gas component (which is assumed to be uniform throughout the cloud).

The total luminosity of the embedded energy source/(s) is frozen at the value determined by integrating the observed continuum SED. The embedded energy source/(s) is varied between a single ZAMS star and a cluster of ZAMS stars. The canonical value of 100:1 for the gas to dust ratio by mass is used initially, but varied if no acceptable model can be constructed to fit all the data. Two types of dust have been considered here, that of Draine & Lee (1984; hereafter DL) and Mezger, Mathis & Panagia (1982; hereafter MMP). Within each type, dust consists of grains of different composition and size (see section 2.1.1). The ISRF has been taken from Mathis, Mezger & Panagia (1983), and has been held fixed for all model runs. The gas component is assumed to be consisting of either (**a**) only hydrogen (section 2.2.1); or (**b**) typical HII region abundance as listed by Ferland (1996) (section 2.2.2). In the latter case, only elements with abundance (relative to hydrogen) higher than 3.0×10^{-6} have been considered. In some model runs for a specific HII region, the elemental abundances used by earlier worker/(s) have been used for comparison. Although the elemental abundance is the same throughout the cloud, the ionization structure and the various level populations depend on several physical parameters including the local radiation field.

The following parameters are explored in order to get an acceptable fit to all the data:

(i) geometric details like $R_{\rm max}$ and $R_{\rm min}$ ($R_{\rm min}$ will not violate radiative destruction of grains);
(ii) radial density distribution law (only three power laws have been explored, viz., $n(r) \approx r^{0}, r^{-1}$ or r^{-2});
(iii) total radial optical depth due to the dust (inclusive of all constituents) at any selected wavelength;
(iv) the dust composition or relative fractions of different constituent grain types (see section 2.1.1).

The interstellar cloud is divided into 141 radial grid points. Near both the boundaries, these grid points are logarithmically spaced (in the rest of the cloud, a linear grid

has been used). The frequency grid consists of 89 points covering the wavelength range 944 A to 5000 μm.

2.1 *Radiative transport through the dust component (D1)*

The radiative transport through the dust component has been carried out by using a programme based on the code CSDUST3 (Egan, Leung & Spagna 1988). We have improvised this code by generalizing the boundary conditions leading to much better flexibility for modelling typical astrophysical sources. The moment equation of radiation transport and the equation of energy balance are solved simultaneously as a two-point boundary value problem in this programme. The effects of multiple scattering, absorption and re-emission of photons on the temperature of dust grains and the internal radiation field have been considered. In addition, the following details, viz., the radiation field anisotropy, linear anisotropic scattering and multi grain components are also included. The entire relevant spectral range covering right from the UV wavelengths to the millimeter region has been considered (the frequency grid consists of 89 points).

For preserving the energetics precisely and self-consistently, the total energy available for heating of the dust component includes all three components (all components being binned into the respectively relevant spectral grid elements):

(i) the star cluster/ZAMS stellar luminosity in photons below the Lyman limit ($\lambda > 912$ Å);
(ii) a part of the Lyman continuum luminosity of the embedded star, ($\lambda < 912$ Å), directly absorbed by the dust; and
(iii) a fraction of the same reprocessed by the gas.

The last contribution, viz., the reprocessed Lyman continuum photons, has been quantified by the prescription of Aller & Liller (1968) that each Lyman continuum photon emitted by the star ultimately leads to one Ly-α photon and one Balmer-α photon.

From the resulting dust temperature distribution in the cloud, the emergent intensities as a function of frequency at various impact parameters (depending on the radial grid) are calculated. Hereafter, the above modelling scheme is referred to as “D1”.

2.1.1 *Dust grains*

Two different approaches have been taken to deal with the dust grains in the present study. The first approach, referred to as “DL” (since largely based on grain properties of Draine & Lee (1984)), is summarized below. The physical properties of the grains, viz., absorption and scattering efficiencies, $Q_{abs}(a,v)$,$Q_{sca}(a,v)$, and the scattering anisotropy factor, $g(a, v)$, for all sizes (a) and frequencies (v) have been taken from the tables of B.T. Draine’s home page which are computed similar to Laor & Draine (1993). These have a finer grid of grain sizes than Draine & Lee (1984). Three types of most commonly accepted variety of interstellar dust have been included in this DL case, viz., (**i**) Graphite, (**ii**) Astronomical Silicate and (**iii**) Silicon Carbide (SiC). The relative abundances of these three types of grains are used as parameters to fit the observed SED well.

In the second approach, the dust grain properties proposed by Mezger, Mathis & Panagia (1982) have been used. This approach, hereafter "MMP", has been considered since it has been found very useful in explaining the SED of certain class of YSO's (Butner *et al.* 1994). This type of dust consists of graphite and silicate only, but their absorptive and scattering properties differ substantially from those for the DL case.

The size distribution of the dust grains is assumed in accordance with Mathis, Rumpl & Nordsieck (1977), to be a power law, viz., $n(a)\mathrm{d}a \sim a^{-m}\mathrm{d}a$, $a_{min} \leq a \leq a_{max}$ with $m = 3.5$. The lower and upper limits of the grain size distribution a_{min} and $a_{\max}$ have been chosen as recommended by Mathis, Mezger & Panagia (1983), to be 0.01 μm and 0.25 μm respectively.

2.2 *Radiative transport through the gas component*

Two independent approaches have been taken to deal with the radiative transfer through the gas component in spherical geometry. The first one takes a very simplistic view considering photoionization and recombination of hydrogen alone, neglecting other heavier elements, as well as the gas-dust coupling (where they co-exist). The treatment of this first approach has been entirely developed in the present study.

The second approach is more sophisticated and considers several prominent elements in the gas phase of the cloud. In addition to photoionization and recombination, other physical processes like collisional excitation and de-excitation, grain photoionization and gas-dust coupling are also included. This detailed modeling involves the use of the photoionization code CLOUDY (Ferland 1996), which has been supplemented with a software scheme developed in the present work, to make the model predictions more realistic and easy to compare with observations.

Whereas the first approach has been used to compute the expected radio continuum emission (without any assumption about optical thinness of the gas at radio wavelength), the second one is used for predicting the atomic/ionic nebular line emission strengths. Both these approaches are discussed next.

2.2.1 *The simple approach (G1)*

In the simple approach, since only hydrogen has been considered, the ionization structure of the gas can be specified by, R_{HII} , the location of the boundary of the HII region (see figure 1). R_{HII} has been determined by considering the radiative transfer of Lyman continuum photons from the embedded star cluster/ZAMS star through the cloud. The effect of the dust component in extinguishing the radiation field, has been considered whenever necessary (cases satisfying $R_{\min} < R_{\mathrm{HII}} < R_{\max}$). In addition, the radio continuum emission has been computed including the effect of appropriate radio optical depth (self absorption). Emission of relevant recombination lines from the ionized gas has been quantified for their role in the radiative transfer through the dust component. Hereafter, this simplistic modelling of the radiative transfer of UV and radio through the interstellar cloud will be referred to as "G1". Appendix 1 gives more details of this treatment.

2.2.2 *The detailed approach (G2)*

For the detailed approach to the radiative transfer through the gas component, the code CLOUDY (Ferland 1996), supplemented by a software scheme developed in the

present study, has been used. The supplementary scheme improves the modelling by (i) emulating the exact structure of the compact HII region; and (ii) including the absorption effects of the dust (present within the line emitting zones), on the emergent line intensities. This detailed approach self-consistently deals with almost all physical processes (radiative-collisional equilibrium) important in and around a photoionized nebula. It simultaneously looks for statistical and thermal equilibrium by solving the equations balancing ionization-neutralization processes and heating-cooling processes. It predicts physical conditions of the gas, e.g., ionization, level populations, temperature structure, and the emerging emission line spectrum. The gas component of the cloud has been considered with typical HII region abundance, as tabulated in Ferland (1996). This is an average of Baldwin *et al.* 1991, Osterbrock, Tran & Veilluex 1992, and Rubin *et al.* 1991, unless specified otherwise. Only the elements with abundance relative to hydrogen, higher than 3.0×10^{-6} have been used. This results into the following elements: H, He, C, N, O, Ne, Mg, Si, S and Ar. The grains of the Astronomical Silicate and Graphite types have been introduced at and beyond a radial distance from the exciting star such that they do not heat up above their sublimation temperature. The heating (photoelectric) and cooling (collisional) due to grains have been considered. The effect of a constant cosmic ray density on the gas is included (which affects energy deposition and ionization).

To be self-consistent with the radiative transfer treatment through the dust component (D1), the entire cloud is considered to be consisting of two spherical shells, the inner one made of gas alone and the outer one with gas and dust. The boundary between the two shells, $R_{\rm min}$, is taken from the corresponding best fit D1(DL) or D1(MMP) model. CLOUDY is run twice, the first time (RUN1) for the inner pure gas shell with the central energy source. The continuum emerging from RUN1 is used as input to the second run (RUN2) for the outer shell. The emerging line spectrum from RUN1 is transported to an outside observer, through the second (outer) shell by considering the extinction due to the entire dust column present there. For every spectral line considered, its emissivities from individual radial zones of RUN2 are transported through the corresponding remaining dust column densities within the outer shell. The emerging line luminosities from RUN1 and RUN2 are finally added to predict the total observable luminosity. This detailed modelling scheme will be referred to as "G2" in the later text.

A total of 27 most prominent spectral lines (from various ionization stages of the above mentioned 10 elements) have been considered. From an observational point of view, the reliable detectability of any spectral line will depend on experimental detail like: the instrument line function (spectral resolution); as well as the strength of the continuum in the immediate spectral neighbourhood of the line. An attempt has been made to predict line intensities for those lines which are detectable by the spectrometers onboard ISO (SWS & LWS).

A line has been defined to be "detectable" only if the expected power incident on the detector, due to the spectral line is more than 1% of the continuum (from the same astrophysical source, originating from the dust) in the corresponding resolution element. An instrumental resolution in the range 1000–20000 has been reported for ISO spectrometers depending on the wavelength (de Graauw *et al.* 1996; Swinyard *et al.* 1996). The lines are in general narrower than the resolution element if the widths are thermal. Only the lines turning out to be "detectable" according to the above criteria are presented with details.

3. Study of the galactic star forming regions: IRAS 183140720, 183550532 and 183160602

With the aim of extracting important geometrical and physical details of the galactic star forming regions—IRAS 18314–0720, 18355–0532 and 18316–0602, the modelling scheme described earlier, has been applied. These sources have been selected to cover a range of ≈ 40 in the total luminosity. In addition, they have adequate observational data necessary to constrain the modelling. In what follows, the observations available for these sources and the results of modelling them, are described.

3.1 *IRAS 18314–0720*

The IRAS Point Source Catalog (hereafter, IRAS PSC) source 18314–0720 has flux densities of 156, 648, 4714 and 8089 Jansky in 12, 25, 60 and 100 μm bands respectively. Although this source did not appear in the original IRAS Low Resolution Spectra Atlas (8–22 μm; hereafter LRS), later analysis released the LRS spectrum of this source (Volk & Cohen 1989). Based on the LRS spectrum, forbidden lines of ions of neon and sulphur have been identified as well as possible detection of features due to Polycyclic Aromatic Hydrocarbons have been reported (Jourdain de Muizon *et al.* 1990). Recently, mid and far infrared spectroscopic detection of several ionic lines based on Kuiper Airborne Observatory measurements have become available (Afflerbach *et al.* 1997). IRAS 18314–0720 corresponds to the Revised Air Force Geophysical Laboratory source RAFGL 2190, which was detected in 4.2, 11, 20 and 27 μm bands. IRAS 18314–0720 was included in the IRAS colour selected sample of Chini *et al.* (1986) for study at 1.3 mm continuum and near infrared mapping (Chini *et al.* 1987), Bronfman *et al.* (1996) have detected *CS* emission at 98 GHz from IRAS 18314–0720. In a search for NH_3 and H_2O maser sources associated with this source, the former has been detected but not the latter (Churchwell *et al.* 1990).

The radio continuum emission from the HII region associated with IRAS 18314–0720 has been observed at various frequencies. The IRAS PSC associates 18314–0720 with radio continuum sources of Parkes and Bonn surveys of the Galactic plane at 5 GHz (Haynes *et al* 1979; Altenhoff *et al* 1979). Later surveys, some with higher angular resolution at 1.4, 5 and 10 GHz have detected this source (Handa *et al.* 1987; Becker *et al.* 1994; Griffith *et al.* 1995; Zoonematkermani *et al.* 1990). It has also been mapped with high angular resolution at 1.5, 4.9 and 15 GHz (Garay *et al.* 1993).

3.1.1 *Observational constraints*

Among all available observational data for IRAS 18314–0720, those most relevant for constraining the modelling are chosen based on their quality/sensitivity and the beam size (a smaller beam size is preferred since we are dealing with compact HII regions which are barely resolved in mid and far infrared wavebands). These include: IRAS PSC, LRS, 1.3 millimeter and the near infrared data for constructing the continuum SED. Observations at wavelengths longer than 1.3 mm have not been used to constrain the models because at these wavelengths the free-free emission from the gas will be a major contributor. Hence in order to compare with the dust continuum emission predicted by the models, one would need to subtract out the estimated free-free emission from the measurements. From the observations at 7 mm (Wood *et al.*

1988) it is clear that, the contamination due to the free-free emission is negligible at 1.3 mm. The infrared forbidden line measurements of ions though detected in LRS, they are only qualitative in nature, hence the only quantitative data from the Kuiper Observatory has been used in the present study. The distance to this source has been taken to be 9.4 kpc from Chan *et al.* (1996). Accordingly the total luminosity (from the observed SED) turns out to be $1.02 \times 10^6\ L_{\odot}$. The radio continuum measurements at 5 GHz (VLA) and 10 GHz (Nobeyama) by Garay *et al.* (1993) and Handa *et al.* (1987) respectively, have been used for model fitting.

3.1.2 *Results of modelling*

With the above observational constraints, the spherically symmetric radiative transfer model D1 has been run (with both DL and MMP type of dust) exploring various parameters described earlier.

The resulting best fit model for the DL dust case, corresponds to:

(i) a single ZAMS star of type 04 ($T_{\rm eff}$ = 50,000 K) as the embedded source;
(ii) a uniform density distribution (i.e. $n(r) = n_0$);
(iii) the radial optical depth at 100μm, $\tau_{100} = 0.1$;
(iv) the gas to dust ratio by mass, 100:1;
(v) the density $n_H = 1.15 \times 10^4 \rm cm^{-3}$;
(vi) $R_{\rm max}$ = 2.5pc;
(vii) $R_{\rm min}$ = 0.05 pc; and
(viii) the dust composition,silicate: graphite: silicon carbide in 7.2:45.3:47.5% proportion.

This D1(DL) model predictions fit the observed SED extremely well, which is shown in Fig. 2. The predicted radio continuum emission at 5 GHz (determined by G1(DL) ran), using these parameters, is only 324 mJy which is almost one tenth of the observed value of 3.51 Jy. The increase in the gas to dust ratio needed to bring the radio continuum emission close to the observed value would be very unphysical. The total cloud mass for this model turns out to be $1.85 \times 10^4 M_{\odot}$ implying the *L/M* ratio to be 55 ($L_{\odot}/M_{\odot}$).

On the other hand, the MMP dust case leads to the following best fit parameters:

(i) a single ZAMS star of type O4 ($T_{\rm eff}$ = 50,000K) as the embedded source;
(ii) a uniform density distribution (i.e. $n(r) = n_0$);
(iii) the radial optical depth at 100 μm, $\tau_{100} = 0.1$;
(iv) the gas to dust ratio by mass, 300:1;
(v) the density $n_H = 3.76 \times 10^4\ \rm cm^{-3}$;
(vi) $R_{\rm max}$ = 2.5 pc; (vii) $R_{\rm min}$ = 0.05 pc; and
(viii) the dust composition, silicate: graphite in 11.5 : 88.5% proportion.

The fit to the observed SED for this D1(MMP) model is also shown in Fig. 2. Although the fit is reasonably acceptable, the absorption feature at ≈ 10 μm predicted by this model is much narrower than the LRS measurements. However, the predictions for radio emission at 5 and 10 GHz in this case (G1(MMP)), viz., 3.56 Jy and 4.56 Jy respectively, match the observations very closely (3.51 Jy and 4.51 Jy). The total cloud mass in this case, turns out to be $5.78 \times 10^4 M_{\odot}$ implying an *L/M* ratio of 17.3 ($L_{\odot}/M_{\odot}$).

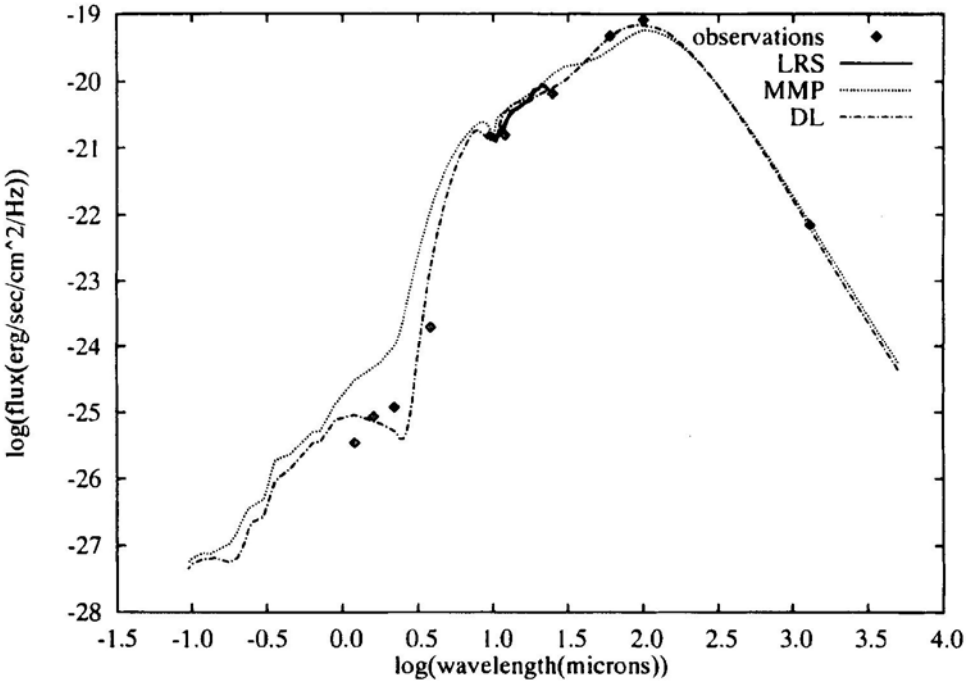

Figure 2. Spectral energy distribution for the source IRAS 18314–0720; the solid line represents LRS observations; the dashed line represents model with MMP grains; the dash-dot-dash line represents model with DL grains; and the symbol ◊ represents other observations (see text).

A comparison between the best fit parameters with DL and MMP dust, constrained by the observations of IRAS 18314–0720, bring out the following: most of the important parameters are identical including even the radial optical depth. This is quite reassuring to note that the geometry as well as the radial density distribution law are invariant to the type of dust (among DL and MMP) selected. Major differences exist for the radio continuum emission predicted and the dust grain composition. Considering the overall fit to the observed continuum SED from the dust and the radio continuum emission from the gas component, clearly the MMP is the favoured self-consistent model for IRAS 18314–0720. Next, we consider the additional information from the forbidden line emission of ionized heavy elements.

In order to predict the forbidden line emission from various ionized species of the gas component in IRAS 18314–0720, the G2 model calculations have been carried out. Runs have been made with the best fit parameters for both the models D1(DL) and D1(MMP). The resulting line luminosities are presented in Table 1. As described above (section 2.2.2) among the 27 relevant lines, only those are included which have intensities more that 1% of the neighbouring continua (for assumed instrumental spectral resolutions of the ISO spectrometers). The observed line strengths of IRAS 18314–0720, for the [SIII] (18.7 μm), [OIII] (51.8 and 88.4μm), and [NIII] (57.3 μm) lines as reported by Afflerbach *et al.* (1997; ACW), are listed in Table 2. The corresponding model predictions (in identical units) for both G2(DL) and G2(MMP) cases are presented in Table 2 for comparison. It is clear that both the models predict far less line emission (for all the four lines) compared to the observations. However, the DL

Table 1. Luminosities of detectable lines from the three sources predicted from their models.

Line (wavelength in μm)	IRAS 18314–0720		IRAS 18355–0532		IRAS 18316–0602	
	DL	MMP	DL	MMP	DL	MMP
	$L_\odot$	$L_\odot$	$L_\odot$	$L_\odot$	$L_\odot$	$L_\odot$
CII (157.7)	2.50	2.93	0.53	0.62	0.15	0.14
OI (145.6)	0.45	3.00	0.14	0.70	0.05	0.13
NI (121.8)	0.09	0.05	0.01	0.01	0.002	0.001
OIII (88.4)	31.3	28.75	1.62	1.76	–	–
OI (63.2)	2.60	4.90	0.88	1.14	0.02	0.20
NIII (57.2)	13.8	10.42	1.07	0.86	–	–
OIII (51.8)	179.3	109.5	10.3	7.18	–	–
NeIII (36.0)	19.3	10.18	1.37	0.72	–	–
SiII (34.8)	0.66	0.46	0.11	0.13	0.03	0.02
SIII (33.5)	10.9	1.73	1.05	0.22	0.03	–
OIV (25.9)	18.9	0.53	0.11	–	–	–
ArIII (21.8)	1.24	–	0.16	–	–	–
SIII (18.7)	19.6	0.48	2.38	0.07	0.06	–
NeIII (15.6)	124.8	15.55	9.70	1.37	–	–
ArV (13.1)	0.70	–	–	–	–	–
NeII (12.8)	1.37	–	0.64	0.12	0.17	0.006
ArIII (10.5)	18.9	–	0.75	–	–	–

Table 2. Comparison of model predictions and observed line intensities (Afflerbach *et al.* 1997) of IRAS 183140720.

Line (wavelength in μm)	Line intensities (10^{-18} watts cm^{-2})			
	G2(DL)	G2(MMP)	Observations (ACW)	G2(ACW)
SIII (18.7)	0.74	0.02	53.0 ± 4.2	47.00
OIII (51.8)	6.78	4.14	51.9 ± 3.8	80.21
NIII (57.2)	0.52	0.39	32.9 ± 2.4	16.55
OIII (88.4)	1.18	1.09	22.3 ± 1.6	26.67

case fares relatively better than the MMP case. ACW have explained these line emissions, originating from a region with electron density n_e = 825 cm^{-3} and $T_{\rm eff}$ = 35, 000 K. Their n_e is too low compared to our models. Perhaps that is the main reason for the discrepancy. In order to verify that, another G2 run is carried out, with the values of n_e, $T_{\rm eff}$ and the elemental abundances identical to those of ACW, but all other details identical to our D1(MMP) model. Results of this run, G2(ACW), are also presented in Table 2. The G2(ACW) predictions are very close to the observations for the [SIII] (18.7 μm) and [OIII] (88.4 μm) lines and within a factor of two for the rest.

This confirms that a lower value of n_e is mainly responsible for the higher observed line intensities in general. This is not unexpected since the collisional de-excitations will become important at higher densities thereby reducing the probability of radiative decays.

The above implies that although a uniform density self-consistent picture is able to explain the SED from the dust and the radio continuum emission from the gas, it fails to explain details of fine structure line strengths for ionized heavy elements. Whereas

the former suggests higher densities, the latter favours a lower one. The detection of molecular maser sources and CS line emission give additional support to the existence of dense medium predicted by our models. In actual source, the reality may lie somewhere in between, viz., a mixture of dense clumps in a thinner inter-clump medium. We explore the role of clumpiness on resolving the issue of fine structure line strengths. Consider the following simplistic scenario: clumps of constant density (ρ_1) immersed in the inter-clump (lower density of ρ_2) medium, with a volume filling factor of f . The clumps are of uniform size and are uniformly distributed throughout the interstellar cloud. This picture has three parameters, viz., ρ_1 , ρ_2 and f.

In this approach, the inter-clump medium (with density p_2, as in the model of ACW) will be mainly responsible for the fine structure line emission (collisional deexcitation will be more important in the clumps). On the other hand, the radio continuum emission will be dominated by the region with higher n_e^2, i.e. the clumps. In addition, if the fit to the observed continuum emission from the dust (SED) is to remain intact, the effective total optical depth (due to dust grains) has to match that from the radiative transfer model (DL or MMP). Hence, there are three constraints to the scenario of clumpy medium: ρ_2 dictated by the fine structure line data;$\langle n_e^2 \rangle$ from the radio continuum; and the dust optical depth from the continuum SED.

Assuming both the clump and the inter-clump medium to be optically thin in the radio; using the above three constraints; and DL type dust; we obtain the following parameters for IRAS 18314–0720. The clumps with a density of 1.34×10^5 cm^{-3} are embedded in the inter-clump medium (with a density of 825 cm^{-3}; same as in ACW), with a volume filling factor of 0.08. The detection of 98 GHz line from the CS molecule, whose excitation requires a critical density $\approx$ 3–4$\times 10^5$ cm^{-3}, further supports the above density of the clumps.

For the radiative transfer modelling of the continuum SED to remain valid, the individual clumps must be optically thin even at the lowest relevant wavelength (i.e. UV). This condition translates the clump diameter to be less than 10^{-3} parsec. For this size, the clumps are found to be optically thin at radio wavelengths as well, thus justifying our assumption of the same in calculating the density (ρ_1) and the volume filling factor of the clumps.

Thus, a self-consistent picture of IRAS 18314–0720 emerges with the DL model including clumpiness, which explains all three major types of observational constraints, viz., continuum SED, radio continuum and fine structure line emission from ionized heavy elements.

3.2 *IRAS 18355–0532*

The IRAS PSC source 18355–0532 has flux densities of 24.6, 209, 1127 and 1930 Jy in 12, 25, 60 and 100 μm bands respectively. This source is included in the original IRAS LRS Catalog. An inspection of the LRS spectrum led to the identification/possible identification of forbidden lines of neon and sulphur ions (Jourdain de Muizon *et al.* 1990). More detailed analysis has identified and quantified emission in a total of four lines from neon and sulphur (Simpson & Rubin 1990). IRAS 18355–0532 corresponds to the RAFGL catalogue source no. 2211, which was detected in 4.2, 11 and 20 μm bands. IRAS 18355–0532 was also included in the IRAS colour selected sample of Chini *et al.* (1986) for study at 1.3 mm continuum and near infrared mapping (Chini *et al.* 1987). Although the CS emission at 98 GHz has been detected

from IRAS 18355–0532 (Bronfman *et al.* 1996), the searches for H_2O (22.2 GHz) and methanol (6.6 GHz) maser emission have been unsuccessful (Codella *et al.* 1995; Schutte *et al.* 1993).

The radio continuum emission associated with IRAS 18355–0532 has been observed at various frequencies. The IRAS PSC associates 18355–0532 with radio continuum sources of Parkes and Bonn surveys of the Galactic plane at 5 GHz (Haynes *et al.* 1979, Altenhoff *et al.* 1979). Later surveys have also detected this source at 1.4, 5 and 10 GHz (Handa *et al.* 1987; Becker *et al.* 1994; Griffith *et al.* 1995; Zoonematkermani *et al.* 1990).

3.2.1 *Observational constraints*

The observations of IRAS 18355–0532 have been chosen to constrain its modeling in the same fashion as in the case of IRAS 18314–0720. The observed SED has been constructed by the following measurements: IRAS PSC, IRAS LRS, 1.3 millimeter and the near infrared data. The infrared forbidden line intensities have been taken from Simpson & Rubin (1990), who have analysed and quantified the IRAS LRS measurements. The distance adopted to this source has been taken to be 6.6 kpc from Chini *et al.* (1987). The total luminosity of IRAS 18355–0532 from the observed SED is estimated to be $1.21\times10^5 L_{\odot}$. The radio continuum measurements at 5 GHz (VLA; Becker *et al.* 1994) and 10 GHz (Nobeyama; Handa *et al.* 1987) have been used for model fitting.

3.2.2 *Results of modelling*

The best fit parameters for the radiative transfer modelling (D1(DL) case) of IRAS 18355–0532, are as follows:

(i) a single ZAMS star of type 06.5 ($T_{\rm eff}$ = 40,000 K) as the embedded source;
(ii) a uniform density distribution (i.e. $n(r) = n_0$);
(iii) the radial optical depth at 100 μm, $\tau_{100} = 0.1$;
(iv) the gas to dust ratio by mass, 100:1;
(v) the density $n_H = 1.71\times10^4$ cm^{-3};
(vi) $R_{\rm max}$ = 1.3pc;
(vii)$R_{\rm min}$ = 0.007 pc; and
(viii) the dust composition, silicate : graphite : silicon carbide in 6.0:46.4:47.6% proportion.

The fit of this D1(DL) model to the observed SED is shown in Fig. 3. This fit is very good for most of the spectral region, including the 10 μm feature. Although, the model curve passes closely to the 25 μm point of the IRAS PSC, it deviates from the longer wavelength region (≈ 15–22 μm) of the LRS spectrum. The radio continuum emission predicted by this model, at 5 GHz is only 51 mJy which is again one tenth of the observed value of 523 mJy. Like we stated earlier (in the case of IRAS 18314–0720), for IRAS 18355–0532 also, we avoid increasing the gas to dust ratio to bring the radio continuum emission closer to the measurements, since that would require a very unphysical value. The total cloud mass for this model turns out to be $4.93\times10\,M_{\odot}$ implying the L/M ratio to be 24.5 ($L_{\odot}/M_{\odot}$).

The D1 (MMP) modelling leads to be the following best fit parameters for IRAS 18355–0532:

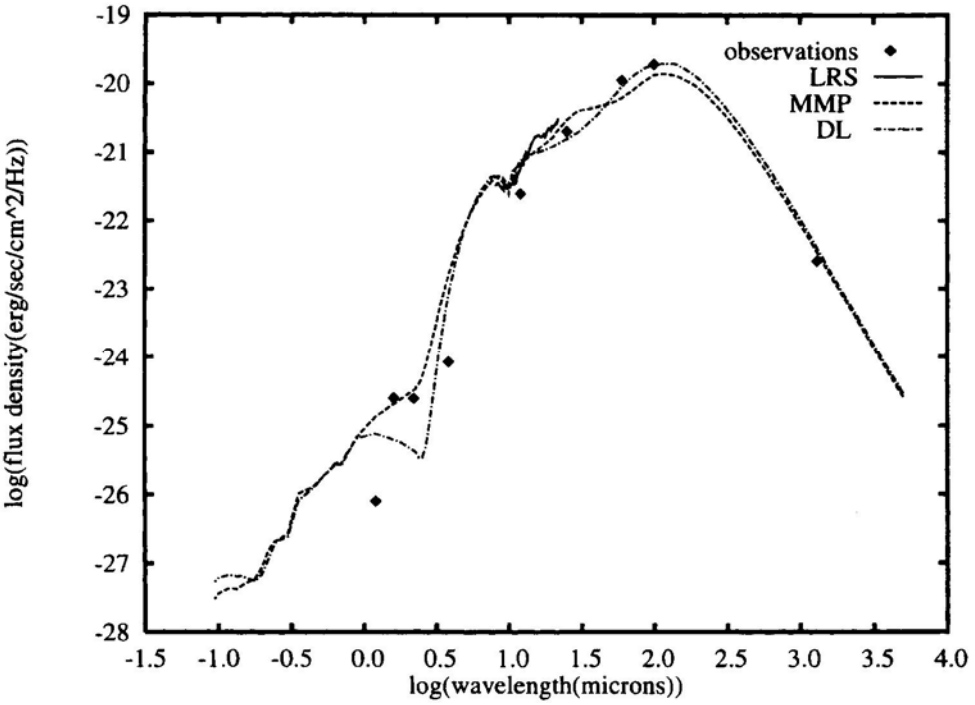

Figure 3. Spectral energy distribution for the source IRAS 18355–0532; the solid line represents LRS observations; the dashed line represents model with MMP grains; the dash-dot-dash line represents model with DL grains; and the symbol represents other observations (see text).

(i) a single ZAMS star of type 06.5 ($T_{\rm eff}$ = 40,000 K) as the embedded source;
(ii) a uniform density distribution (i.e. $n(r) = n_0$);
(iii) the radial optical depth at 100 μm, $\tau_{100} = 0.1$;
(iv) the gas to dust ratio by mass, 300:1;
(v) the density $n_H = 7.20 \times 10^4$ cm^{-3};
(vi) $R_{\rm max} = 1.3$pc;
(vii) $R_{\rm min} = 0.018$ pc; and
(viii) the dust composition, silicate: graphite in 12: 88% proportion.

The fit to the observed SED for D1(MMP) model is also shown in Fig. 3, which is reasonably good. This model fits the longer wavelength segment of the LRS spectrum quite well (which the DL case could not), but the 10 μm feature predicted is much narrower than the observed one. The fit to 60 and 100 μm IRAS PSC points is slightly poorer than the DL case. This G1(MMP) case is much more successful than G1(DL), in predicting the radio continuum emission. The predicted values for this case (G1(MMP)), are 872 mJy and 1.18 Jy at 5 and 10 GHz respectively, which are very close to the observed numbers (523 mJy and 1.61 Jy). The mass of the cloud associated with IRAS 18355–0532, for the parameters of this MMP model, is $1.56 \times 10^4 L_\odot$. This implies an L/M ratio of 7.8 ($L_\odot/M_\odot$)

Comparing the best fit parameters for IRAS 18355–0532, for the models with DL and MMP dust, one finds conclusions similar to the earlier case of IRAS 18314–0720.

Table 3. Comparison of model predictions and observed line intensities (Simpson & Rubin 1990) of IRAS 18355–0532.

Line (wavelength in μm)	Line intensities (10^{-18} watts cm^{-2})			
	G2(DL)	G2(MMP)	Observations (SR)	G2(SR)
SIV (10.5)	0.058	1.50×10^{-5}	0.5	3.09
NeII (12.8)	0.048	9.037×10^{-3}	14.0	2.55
NeIII (15.6)	0.745	0.105	8.1	6.91
SIII (18.7)	0.182	5.219×10^{-3}	14.0	8.35

Most important parameters are identical, e.g., radial density distribution and the radial optical depth. The differences exist for the predicted radio continuum and the dust grain composition. Once again, the MMP model is the favoured model for IRAS 18355–0532, since it self consistently fits the observed continuum SED as well as the radio continuum emission, reasonably well. Similar to the earlier source, here again we consider the data of fine structure lines of ionized heavy elements in IRAS 18355–0532.

The predictions from the G2 runs for IRAS 18355–0532, corresponding to the above mentioned best fit parameters of D1(DL) as well as D1(MMP) models are also presented in Table 1. Once again the line luminosities for only the detectable infrared forbidden lines are included. Simpson & Rubin (1990; SR) have carefully analysed the 822 μm IRAS LRS data for IRAS 18355–0532, as one member of a large sample. They have quantified the line intensities for [SIV] (10.5 μm), [Nell] (12.8 μm), [NeIII] (15.6 μm) and [SIII] (18.7 μm) lines. These are also presented in Table 3. The corresponding model predictions (in identical units) for both G2(DL) and G2(MMP) cases are listed in Table 3 for easy comparison. Both the models predict far less line emission (for all the four lines) compared to the observations. Just like in the earlier case of IRAS 18314–0720, for IRAS 18355–0532 also, the DL fares much better than the MMR

Since there are two pairs of lines from the same elements, viz., S and Ne, the line ratios will be less sensitive to the abundances. Whereas the measured intensity ratios between [NeII]/[NeIII] and [SIII]/[SIV] are 1.7 and 28 respectively, the same for the G2(DL) model are 0.06 and 3.1. It is interesting to note that, if in the same G2(DL) model, the $T_{\rm eff}$ is reduced to 28,000 K, then both the observed line ratios are reproduced.

SR have modelled the line emissions from IRAS 18355–0532, with electron density $n_e = 3.16 \times 10^3$ and $T_{\rm eff} = 38{,}500$ K. This n_e is very low compared to our models. In addition, their elemental abundances are different from ours. In order to verify the hypothesis that, the high value of n_e is responsible for the failure of our models to predict the line intensities, another G2 run is carried out with the abundances, n_e and $T_{\rm eff}$ values from SR, but all other details same as D1(MMP). Predictions of this model, G2(SR), are also presented in Table 3. The G2(SR) predictions are reasonably close to the observations for the [NeIII] (15.6 μm) and [SIII] (18.7 μm) lines but the other two line intensities are down by a factor of ≈ 6. This success of G2(SR) supports the above hypothesis about the value of n_e, which is not unexpected since the collisional de-excitations become less important at lower densities.

Once again, like in the case of IRAS 18314–0720, for IRAS 18355–0532 also, a uniform density self-consistent picture is able to explain the SED from dust and the

radio continuum emission from the gas, but fails to explain details of fine structure line strengths for ionized heavy elements. The detection of molecular maser sources and CS line further supports the existence of denser medium predicted by our self-consistent models. Since again, a lower value of n_e has been relatively more successful in predicting the forbidden line strengths, we propose the possible scenario of clumpiness in IRAS 18355–0532 too, for resolution of the above problem, like the earlier case of IRAS 18314–0720.

Using an identical approach of incorporating clumpiness in IRAS 18355–0532, as was used for the earlier source IRAS 18314–0720, a physically meaningful solution has been found corresponding to the DL scheme of modelling. This solution corresponds to the following parameters: $\rho_1 = 2.12 \times 10^5$ cm^{-3}; $\rho_2 = 3.16 \times 10^3$ cm^{-3} (same as in SR) and $f = 0.067$. The detection of 98 GHz line from the CS molecule further supports the above density inside the clumps. Once again, arguing from the point of validity of the DL modelling of the continuum SED, the upper limit on the diameter of the clumps is set to 5.2×10^{-4} parsec.

Thus, even for IRAS 18355–0532, a self-consistent picture emerges with the DL model including clumpiness, which explains all three major types of observational constraints.

3.3 *IRAS 18316–0602*

The IRAS PSC source 18316–0602 has flux densities of 22.8, 138, 958 and 2136 Jansky in 12, 25, 60 and 100 μm bands respectively. This source is included in the IRAS LRS Catalog. The LRS spectrum shows the 10 μm silicate feature, but no forbidden line or any feature due to the Polycyclic Aromatic Hydrocarbons (Jourdain de Muizon *et al.* 1990). However, recent ISO-SWS measurements of IRAS 18316–0602 do show a very rich spectrum full of various solid state molecular features (d'Hendecourt *et al.* 1996; Dartois *et al.* 1998). IRAS PSC associates 18316–0602 with RAFGL 7009S, which was detected at 4.2, 11, 20 and 27 μm. Sub-millimeter and millimeter waveband continuum observations of IRAS 18316–0602 have been carried out at 450, 800, 850 and 1100 μm by Jenness *et al.* (1995) and McCutcheon *et al.* (1995). The 2.6 mm CO line has also been detected from this source by McCutcheon *et al.* 1991. The CS and NH_3 emission have been detected from IRAS 18316–0602 (Bronfman *et al* 1996; Molinari *et al.* 1996). Searches for H_2O and methanol maser emission from this source have also been successful (Brand *et al.* 1994; van der Walt *et al.* 1995; Codella *et al.* 1996)

The radio continuum emission associated with IRAS 18316–0602 has been observed at 5 and 8 GHz (Jenness *et al.* 1995; McCutcheon *et al* 1991; Kurtz *et al.* 1994; Griffith *et al.* 1995).

3.3.1 *Observational constraints*

The observed SED for IRAS 18316–0602 has been generated from the following: IRAS PSC, IRAS LRS, selected continua from the ISO-SWS spectrum (3–8 μm) and all theavailable sub-mm/mm observations (450–1100 μm).Radio continuum data at 5 and 8 GHz have been used as constraints for the modelling of this source.

The distance to this source has been taken to be 3.3 kpc from Chan *et al.* (1996). The corresponding total luminosity is $2.54 \times 10^4 L_\odot$.

3.3.2 *Results of modelling*

The total luminosity of IRAS 18316–0602 corresponds to a single ZAMS star of type BO ($T_{\rm eff}$ = 30,900 K). However, since a single ZAMS star as the embedded source leads to a poor fit to the observed SED as well as the radio continuum data, various clusters of ZAMS stars with a given Initial Mass Function but a variable upper mass cut-off (M_u), have been tried. After exploring the parameter space for Dl(DL) modelling, the following best fit parameters have been determined for IRAS 18355–0532:

(i) a cluster of ZAMS stars with an Initial Mass Function of the form $N(M) \approx M^{-2.4}$, with the upper mass cut off, M_u, corresponding to the type B1,
(ii) a uniform density distribution (i.e. $n(r) = n_0$);
(iii) the radial optical depth at 100 μm, $\tau_{100} = 0.1$;
(iv) the gas to dust ratio by mass, 100:1;
(v) the density $n_H = 2.28 \times 10^4$ cm^{-3};
(vi) $R_{\rm min} = 0.0001$ pc; and
(vii) the dust composition, silicate: graphite: silicon carbide in 71.8:28.2:0.0% proportion.

The above parameters are also consistent with the column density (N_{H_2}) derived from CO measurements of McCutcheon *et al* 1991. The fit of this D1(DL) model to the observed SED is shown in Fig. 4. This fit is very good for most of the spectral

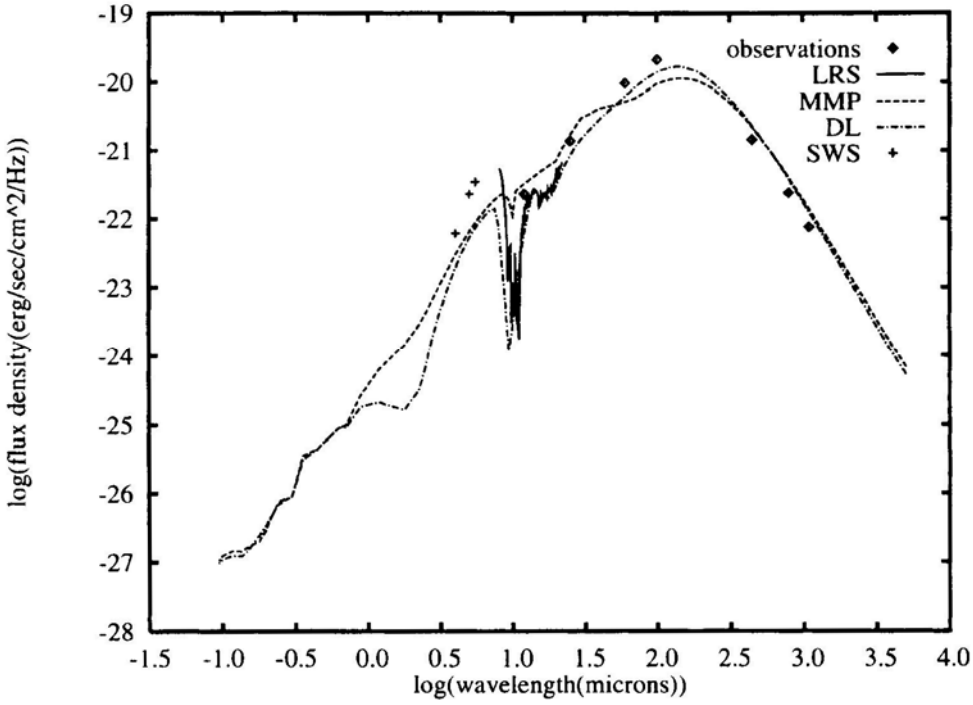

Figure 4. Spectral energy distribution for the source IRAS 18316–0602; the solid line represents LRS observations; the dashed line represents model with MMP grains; the dash-dot-dash line represents model with DL grains; the symbol + represents SWS observations; and the symbol represents other observations (see text).

region, particularly for the longer wavelengths of the LRS spectrum. The position of the predicted ≈ 10 μm feature is slightly to the shorter wavelength side compared to the observations (LRS). Any attempt to "align" this feature towards the longer wavelength by increasing SiC dust relative to the silicate dust, leads to very poor fit to the 15–22 μm part of the spectrum (LRS data). Incidentally, the SiC dominant dust composition corresponding to the best aligned feature, is – silicate: graphite: silicon carbide in 0.9:11.0: 88.1% proportion. Although the predicted SED by the DL model passes acceptably close to the ISO-SWS continuum at 4.0 μm, the fit is rather poor at 5.0 and 5.5 μm, the observed values being about 3 times the predictions. The radio continuum emission predicted corresponding to this model, at 5 GHz is about 2.16 mJy which is reasonably close to the observed value of 2.7 mJy. In fact a slight increase the gas to dust ratio can bring the model prediction exactly to the measured value the total cloud mass for this DL model turn out to be 3.88×10^3 $M_\odot$ implying the L/M ratio to be 6.55 ($L_\odot/M_\odot$).

The D1(MMP) modelling leads to the following best fit parameters for IRAS 18316–0602:

(i) a cluster of ZAMS stars with an Initial Mass Function of the form $N(M) \approx M^{-2.4}$, with the upper mass cut off, M_u, corresponding to the type B1
(ii) a uniform density distribution (i.e. $n(r) = n_0$);
(iii) the radial optical depth at 100 μm, $\tau_{100} = 0.1$;
(iv) the gas to dust ratio by mass, 300:1;
(v) the density $n_H = 7.74 \times 10^4 \mathrm{cm}^{-3}$;
(vi) $R_{\min} = 0.0005$ pc; and
(vii) the dust composition, silicate: graphite in 11: 89% proportion.

The fit to the observed SED of IRAS 18316–0602 by the D1(MMP) model is also shown in Fig. 4. Although this fit crudely represents the broad overall shape of the SED, it fails to reproduce many details, particularly near the ≈ 10 μm feature. The predicted feature in this MMP case is far too narrow as well as shallow compared to the LRS data. At far infrared wavelengths also (IRAS PSC 60 and 100 μm) the predictions are below the observations. The predicted radio emission at 5 and 8 GHz in this case (G1(MMP)) are 3.3 and 3.4 mJy respectively, which are in reasonable agreement with the observations (2.7 and 3.8 mJy; McCutcheon *et al.* 1991; Jenness *et al.* 1995). The cloud mass for this MMP case is 1.31×10^4 $L_\odot$, leading to a value for the L/M ratio of 1.9 ($L_\odot/M_\odot$).

The D1(DL) is clearly the preferred model for IRAS 18316–0602. The detection of maser sources associated with IRAS 18316–0602 is consistent with best fit model gas densities.

The model G2 prediction (corresponding to the best fit parameters of D1(DL) as well as D1(MMP) cases) for all the detectable infrared forbidden line emission from IRAS 18316–0602 have been presented in Table 1, alongwith other two sources. Although IRAS 18316–0602 has been studied by the ISO-SWS spectrometer covering the entire wavelength range of 2.5–45 μm, the observations/data binning have been carried out at a low resolution of 300–500 (d'Hendecourt *et al.* 1996; Dartois *et al.* 1998). The few selected narrow wavelength regions covered at higher resolution (1500–2000; Dartois *et al.* 1998) do not cover the lines predicted to be detectable by our models (see Table 1). In addition, the detectability criterion used by us uses ISO-SWS observation modes with much higher resolution (in the 12–45 μm

region) than these reported spectra (i.e. even if the predicted lines were covered, they would not have been detectable at this intermediate resolution observational mode employed by Dartois *et al.* 1998).

For IRAS 18316–0602 too, like in the cases of IRAS 18314–0720 and 18355–0532, a uniform density self-consistent picture is able to explain the SED from dust and the radio continuum emission from the gas. Unfortunately, no measurement exists to date, for any infrared fine structure line for this source. In case, IRAS 18316–0602 has been observed using the ISO-LWS spectrometer (in high resolution configuration), then our models can be qualified further.

4. Conclusions

A simple yet self-consistent approach towards explaining observed spectral energy distribution of interstellar clouds with embedded YSO's/compact HII regions has been described. The radiation of the embedded source/(s) is transported through the dust and the gas components by different schemes in spherical geometry. Two kinds of dust have been considered (Draine & Lee (DL); and Mezger, Mathis & Panagia (MMP)), each with its own variable composition. Here, by self-consistent one means that the same geometric and physical configuration fits the observed data for the emission from the dust (most of the infrared, sub-mm, mm part of the SED) as well as the emission from the gas (radio continuum).

The effectiveness of this scheme has been assessed by applying it to three Galactic star forming regions associated with IRAS 18314–0720, 18355–0532 and 18316–0602. They cover a range of about 40 in luminosity of the embedded source/(s). Relevant observational data for these sources have been modelled to extract information about their physical size, density distribution law, total optical depth and dust composition. Interestingly, in all these three cases, the best fit models correspond to the uniform density distribution (for either DL or MMP dust). Similar conclusion about constant density envelopes, has been drawn recently by Faison *et al.* (1998) for a sample of 10 Galactic compact HII regions.

For both IRAS 18314–0720 and 18355–0532, the MMP dust models are the favoured models, since they not only give reasonably acceptable fits to the continuum SED, but also explain the radio continuum data.

Even though SED and radio continuum observations have been well explained by the above modelling, they predict much lower intensities for fine structure lines of ionized heavy elements, wherever measurements are available (IRAS 18314–0720 and 18355–0532). This discrepancy has been resolved by invoking clumpiness in the interstellar medium. Two phase (clump/inter-clump) models with DL type dust, have been successfully constructed for IRAS 18314–0720 and 18355–0532.

In the case of IRAS 18316–0602, DL is the preferred model which gives a very good fit to the observed SED, as well as predicts radio continuum emission which is consistent with the measurements.

Acknowledgements

It is a pleasure to thank Gary Ferland for his help on several occasions regarding the code CLOUDY; and D. Narasimha for clarifying certain doubts about radiative

transfer. Members of Infrared Astronomy Group are thanked for their comments. The authors thank the anonymous referee for the comments which improved the conclusions regarding clumpiness.

APPENDIX 1

A.1 *Simple approach to radiative transfer through the gas (G1)*

In this simple approach, the extent of the ionized region (in spherical geometry) is determined by transporting the Lyman continuum photons (from the centrally located star/star cluster) through the cloud including the effect of the dust, where they can exist (as determined by their sublimation). The gas component of the cloud is assumed to consist of only hydrogen. Next, the radio continuum emission emerging from the cloud is calculated by transporting the radio photons (free-free emission throughout the ionized medium of the cloud), through the entire cloud without making any approximation about the optical thinness of the gas (i.e. self absorption is treated appropriately). The gas to dust coupling has been neglected.

A. 1.1 *Extent of the ionized region*

The size of the HII region has been calculated by considering photoionization and recombination of hydrogen, along with the absorption due to the dust grains. The presence of the dust reduces the size of the ionized region, ($R_{\rm HII}$), compared to that of pure gas Stromgren sphere considerably, depending on the density and the gas to dust ratio. The dust grains can exist in principle, only beyond a radial distance, say $r_{\rm subl}$, depending on its sublimation temperature and the radiation field due to the central source. In practice, the actual distance beyond which the dust exists, say $r_{\rm fit,}$ is determined by the model fitting of the observed SED, by radiative transport calculations through the dust (D1(DL) or D1(MMP)). The $r_{\rm fit}$ is often much larger than $r_{\rm subl}$

Hence, whether one encounters a dusty Stromgren sphere or not, is determined by the type of the star/integrated spectrum of the cluster; radial density distribution around the central star; and $r_{\rm fit}$ We call it Case A, if the ionized region extends into the region where gas and dust coexist. The other case of entire ionized region devoid of any dust grains is termed Case B. So for Case B, the extent of the HII region can be obtained by solving the equation,

$$-\mathrm{d}N(r) = 4\pi\alpha_2 r^2 n_e^2(r)\mathrm{d}r \tag{1}$$

where, N is the number of Lyman continuum photons, α_2 is the recombination coefficient for hydrogen (for recombinations to all states except the ground state) and n_e is the number density of electrons or H^+ ions (for a pure HII region), which in our case is the gas number density (n_g) and may be given by,

$$n_g = n_0^{'}\left(\frac{R_{\rm min}}{r}\right)^m, \quad m = 0, 1, 2. \tag{2}$$

For m = 0,1,2 equation (1) can be solved easily by using the boundary condition

$$\text{at} \quad r = r_*, \quad N(0) = N_{\rm Lyc} \tag{3}$$

where, $N_{\rm Lyc}$ is the total number of Lyman continuum photons emitted per second by the embedded exciting star/star cluster and r_* is an effective stellar radius (with volume equal to the sum of that of all the stars of the embedded cluster; as it turns out, results are extremely insensitive to r_*).

In case A however, the ionizing (Lyman continuum) photons experience further attenuation due to direct absorption by the dust, so the modified radiation transfer equation would be,

$$-{\rm d}N(r) = 4\pi r^2 \alpha_2 n_e^2 {\rm d}r + N(r)\tau_{\rm Lyc}{\rm d}r \tag{4}$$

where $\tau_{\rm Lyc}$ refers to the optical depth of dust at $\lambda < 912$ Å. We solve the above equation, using the boundary conditions,

$$\text{at} \quad r = R_{\rm min}, \quad N(R_{\rm min}) = N_1 \tag{5}$$

where N_1 is determined by using equation (1) and

$$\text{at} \quad r = R_{\rm HII}, \quad N(R_{\rm HII}) = 0 \tag{6}$$

A. 1.2 *Calculation of continuum emission*

With $R_{\rm HII}$ properly determined, the radio continuum emission which occurs due to the free-free emission from the ions and electrons can be calculated by using the formula (Spitzer 1978),

$$J_\nu = \int_{r_*}^{R_{\rm HII}} 4\pi r^2 (4\pi\epsilon_\nu) e^{-\int_r^{R_{\rm HII}} \kappa_\nu {\rm d}r} {\rm d}r \tag{7}$$

where, the coefficients of emission ε_ν and absorption k_ν are respectively given by,

$$\epsilon_\nu({\rm erg/cm^3/sec/sr/Hz}) = 5.44 \times 10^{-39} g_{ff} Z_i^2 n_e n_i T^{-0.5} e^{-h\nu/kT} \tag{8}$$

$$\kappa_\nu(1/{\rm cm}) = 0.1731(1 + 0.130\log(T^{3/2}\nu^{-1}))Z_i^2 n_e n_i T^{-3/2}\nu^{-2} \tag{9}$$

with, Gaunt factor (*gff)* given by,

$$g_{ff} = 9.77(1 + 0.130\log(T^{3/2}\nu^{-1})). \tag{10}$$

An electron temperature of 8000 K has been assumed for all calculations. The radio continuum emission is computed at different frequencies depending on the availability of measurements for the particular astrophysical source under study. The frequencies are typically between 5 and 10 GHz.

References

Afflerbach, A., Churchwell, E., Werner, M. W. 1997, *Astrophys. J.,* **478**, 190 (ACW).

Aller, L. H., Liller, W. 1968, *Stars and Stellar Systems,* **Vol. VII**, 483.

Altenhoff, W. J., Downes, D., Pauls, T., Schraml, J. 1979, *Astr Astrophys. (Suppl)*, **35**, 23.

Baldwin, J., Ferland, G. J., Martin, P. G. *et al.* 1991, *Astrophys. J.,* **374**, 580.

Becker, R. H., White, R. L., Helfand, D. J., Zooneraatkermani, S. 1994, *Astrophys. J. Suppl*., **91**, 347.

Brand, J.,Cesaroni, R., Caselli, P. *et al.* 1994, *Astr. Astrophys.* (*Suppl*)*,* **103**, 541.

Bronfman, L.,Nyman,L.-A., May,J.1996, *Astr. Astrophys. (Suppl)*,**115**, 81.
Butner, H. M., Evans, N. J., Lester, D. F, Levreault, R. M., Strom, S. E. 1994, *Astrophys. J.,* **420**,326.
Chan, S. J., Henning, T., Schreyer, K. 1996, *Astr. Astrophys. (Suppl),* **115**, 285.
Chini, R., Kreysa, E., Mezger, P. G., Gemund, H. P. 1986, *Astr. Astrophys.,* **154**, L8.
Chini, R., Krugel, E., Wargau, W. 1987, *Astr. Astrophys.,* **181**, 378.
Churchwell, E.,Walmsley, C. M.,Cesaroni, R.1990, *Astr. Astrophys. (Suppl),***83**, 119.
Codella, C, Felli, M., Natale, V. 1996, *Astr. Astrophys.*, **311**, 971.
Codella, C, Palumbo, G. G. C., Pareschi, G., Scappini, R, Caselli, P., Attolini, M. R. 1995, *Mon. Not. R. Astr. Soc.,* **276**, 57.
Dartois,E., d'Hendecourt, L.,Boulanger, R *et al.* 1998, *Astr. Astrophys.*, **331**, 651.
de GraauwTh., Haser,L.N., Beintema,D. A. *et al.* 1996, *Astr. Astrophys.*, **315**, L49.
D'Hendecourt, L., Jourdain de Muizon, M., Dartois, E. *et al.* 1996, *Astr. Astrophys.*, **315**, L365.
Draine, B. T., Lee, H. M., 1984, *Astrophys. J.*, **285**, 89 (DL).
Egan, M. P., Leung, C. M., Spagna, G. R1988, *Computer Physics Communications*,**48**,271.
Faison, M., Churchwell, E., Hofner, P., Hackwell, J., Lynch, D. K., Sussell, R. W. 1998, *Astrophys. J.*, **500**, 280.
Ferland, G. J. 1996, *Hazy*, a brief introduction to CLOUDY, Univ. of Kentucky, Dept. of Phys. and Astron. Internal Reports.
Garay, G., Rodriguez, L. R, Moran, J. M., Churchwell, E. 1993, *Astrophys. J.*, **418**, 368.
Griffith, M. R., Wright, A. E., Burke, B. F, Ekers, R. D. 1995, *Astrophys. J. (Suppl),* **97**, 347.
Handa, T., Sofue, Y, Nakai, N., Hirabayashi, H., Inoue, M. 1987, *PASJ*, **39**, 709.
Haynes, R. R, Caswell, J. L., Simons, L. W. J. 1979, *Aust. J. Phys. Astrophys. (Suppl),* No. 48.
Jenness, T., Scott, P. F, Padman, R. 1995, *Mon. Not. R. Astr. Soc*, **276**, 1024.
Jourdain de Muizon, M., Cox, P., Lequeux, J. 1990, *Astr. Astrophys. (Suppl)*, **83**, 337.
Kurtz, S., Churchwell, E., Wood, D. O. S. 1994, *Astrophys. J. (Suppl)*, **91**, 659.
Laor, A., Draine, B. T. 1993, *Astrophys. J.*, **402**, 441.
Mathis, J.S.,Mezger, P. G.,Panagia, N. 1983, *Astr. Astrophys.*, **128**, 212.
Mathis, J. S., Rumpl, W., Nordsieck, K. H. 1977, *Astrophys. J.*, **217**, 425.
McCutcheon, W. H., Dewdney, P. E., Purton, R., Sato, T. 1991, *Astr. J.*, **101**, 1435.
McCutcheon, W. H., Sato, T., Purton, C. R., Matthews, H. E., Dewdney, P. E. 1995, *Astr. J.*, **110**, 1762.
Mezger, P. G., Mathis, J. S., Panagia, N. 1982, *Astr. Astrophys.*, **105**, 372 (MMP).
Molinari, S., Brand, J., Cesaroni, R., Palla, R 1996, *Astr. Astrophys.*, **308**, 573.
Osterbrock, D.E.,Tran, H. D.,Veilluex, S.1992, *Astrophys. J.*, **389**, 305.
Rubin, R. H., Simpson, J. R., Haas, M. R., Erickson, E. F. 1991, *Astrophys. J.*, **374**, 564.
Schutte, A. J., van der Walt, D. J., Gaylard, M. J., Macleod, G. C. 1993, *Mon. Not. R. Astr. Soc*, **261***,* 783.
Simpson, J. P., Rubin, R. H. 1990, *Astrophys. J*, **354**, 165 (SR).
Spitzer, L. 1978, *Physical Processes in the Interstellar Medium*, p.57.
Swinyard, B. M., Clegg, P. E., Ade, P. A. R. *et al.* 1996, *Astr. Astrophys.*, **315**, L43.
van der Walt, D.J., Gaylard, M. J., Macleod, G.C.1995, *Astr. Astrophys. (Suppl)*, **110**, 81.
Volk, K., Cohen, M. 1989, *Astr. J.,* **98**, 931.
Wood, D. O. S., Handa, T., Fukui, Y, Churchwell, E., Sofue, Y., Iwata, T. 1988, *Astrophys. J.*, **326**, 884.
Zoonematkermani, S., Helfand, D. J., Becker, R. H., White, R. L., Perley, R. A. 1990, *Astrophys. J. (Suppl)*, **74**, 181.

J. Astrophys. Astr. (1999) **20**, 23-35

Infrared Emission from Interstellar Dust Cloud with Two Embedded Sources: IRAS 19181 + 1349

A. D. Karnik & S. K. Ghosh, *Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai (Bombay) 400 005, India.*

Received 1999 June 18; accepted 1999 June 30

Abstract. Mid- and far-infrared maps of many Galactic star forming regions show multiple peaks in close proximity, implying more than one embedded energy source. With the aim of understanding such interstellar clouds better, the present study models the case of two embedded sources. A radiative transfer scheme has been developed to deal with a uniform density dust cloud in a cylindrical geometry, which includes isotropic scattering in addition to the emission and absorption processes. This scheme has been applied to the Galactic star forming region associated with IRAS 19181 + 1349, which shows observational evidence for two embedded energy sources. Two independent modelling approaches have been adopted, viz., to fit the observed spectral energy distribution (SED) best; or to fit the various radial profiles best, as a function of wavelength. Both the models imply remarkably similar physical parameters.

Key words. Interstellar clouds—infrared SED—IRAS 19181 + 1349.

1. Introduction

Galactic star forming regions mostly comprise of Young Stellar Objects (YSOs)/ protostars still buried inside/in the vicinity of the parent interstellar cloud from which they are formed. Hence the study of YSOs leads to understanding of the interstellar medium in the close neighborhood of the starbirth. Early evolution of star forming regions is obviously more important from the point of understanding the star formation process itself. Typically the photons from the embedded energy source, protostar/ZAMS star, get reprocessed by the dust component of the interstellar medium in the immediate neighborhood. The dust grains absorb/scatter the incident radiation, depending on their dielectric properties. The dust grains acquire an equilibrium temperature based on the local radiation field which depends on the distance from the energy source and secondary radiation from other grains. The reradiation from the grains in the outer regions, is what is observable. Hence, the emerging observable spectrum can in principle be connected to the spectrum of the embedded protostar/ZAMS star through detailed radiation transfer provided some details about the geometry are known.

In order to study the earliest stage of star formation, a large amount of observational effort is directed towards far infrared / sub-mm observations of prospective young star forming regions. Many Galactic star forming regions have been mapped with near diffraction limited angular resolutions, using Kuiper Airborne Observatory,

Infrared Space Observatory (ISO), James Clerk Maxwell Telescope etc. in recent times. In many cases, the maps of continuum emission resolve several closeby individual intensity peaks as evident from the morphology of their isophots.

The present study is a step in the direction of extracting maximum possible information about the geometrical and physical details of the source by comparing radiative transfer models with observations in cases where two nearby sources are resolved. Here, the "nearby" implies the interference in energetics of each of the two resolved sources by the other one. The heating of dust in a cloud with multiple sources has been studied by Rouan (1979) under certain simplification. However, the problem of more than one embedded source has rarely been addressed quantitatively. The existing observational data showing clear evidence of resolved multiple embedded sources justify the need to explore geometries dealing with more than one embedded source. Ghosh & Tandon (1985) attempted to study the case of two embedded sources with many simplifying assumptions. They neglected some basic phenomena like scattering which limited its applicability to $\lambda \geq 50$ μm only. The present work is an extension of this earlier attempt by including the effect of isotropic scattering.

In section 2 the problem has been formulated and the radiation transfer scheme is described. In section 3, we model observations of IRAS 19181 + 1349 which shows evidence of having two embedded sources. The results of our modelling are then discussed.

2. Model formulation

The primary aim of the present model is to reproduce the infrared emission from a star forming cloud with two embedded sources, keeping the computational complexities at a minimum level.

The interstellar cloud is assumed to be of cylindrical shape. As a starting point, a uniform density of the cloud has been assumed. The line joining the two embedded ZAMS stellar/ protostellar energy sources defines the symmetry axis of the problem. Around each of the sources there will be a dust free cavity (Fig. 1). The existence of such a cavity is widely accepted due to evaporation of the dust grains in the intense radiation field. In addition, radiation pressure effects on the dust grains may also play a role in deciding the cavity size. The radiation transfer is carried out through the dust component alone. Dust grains with a continuous size distribution have been considered, and their composition is a parameter of modelling. Three types of grains, viz., Graphite, Astronomical Silicate and Silicon Carbide have been invoked since their existence is generally accepted. All properties of the dust grains, viz., absorption and scattering coefficients as a function of wavelength, for various sizes and all three types of grains, have been taken from Laor & Draine (1993). The size distribution of all the three types of grains has been assumed to be power law ($n(a)\mathrm{d}a \sim a^{-3.5}\mathrm{d}a$) as per Mathis, Rumpl & Nordsieck (1977). The wavelength grid used 89 points covering from the Lyman continuum limit to the millimeter wavelengths.

The geometrical parameters relevant to the model (Fig. 1) include the radius of the cylinder (R_{cyl}), radii of dust free cavities near the two sources (R_{c1}, R_{c2}), and distance between two exciting sources (D). Other physical parameters are the composition of the dust (relative abundances of three types of dust components) and the dust density expressed in optical depth at 100 μm (τ_{100}). The optical depth at any other wavelength

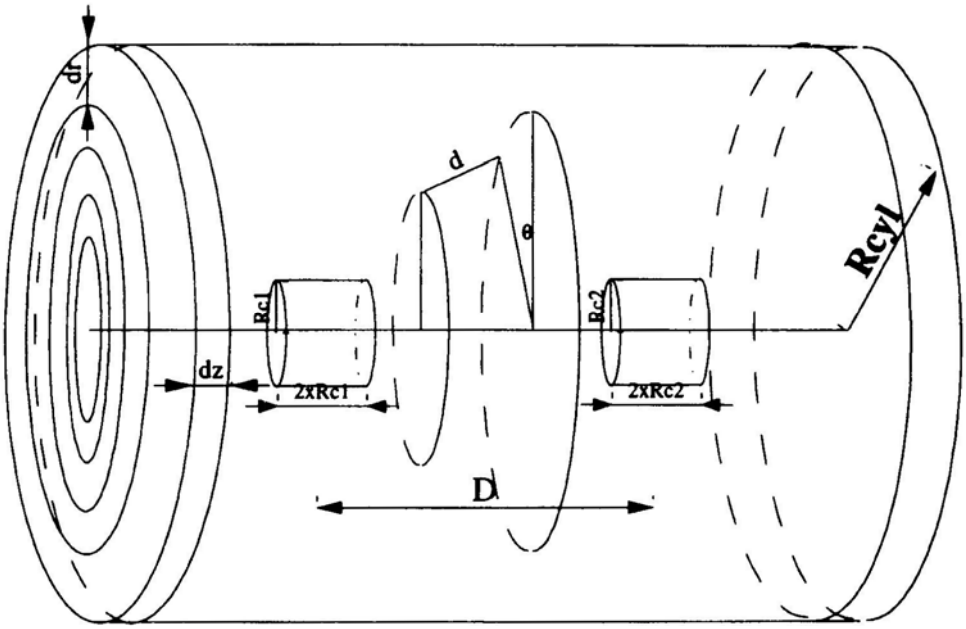

Figure 1. The geometry of the cylindrical cloud considered in the present study. The embedded energy sources lie along the axis of the cylinder (at the centers of the two small cylinders) separated by distance *D*. The two small cylinders represent the dust free cavities near the respective sources. The *z*-axis is defined to be along the symmetry axis of the cylinder. The cloud is divided into n_z disks along z-axis, and each disk into n_r concentric rings.

is uniquely connected to τ_{100} via the dust properties and composition assumed in the particular model. The cylindrical interstellar cloud is divided into n_z identical discs. Each of these discs is further divided into n_r annular rings. The n_z is chosen such that each disc is optically thin even at UV, along the *z*-axis. Along the radial direction (total number of grid points being n_r), a two stage grid has been employed which is initially nearly logarithmically spaced (near the symmetry axis) and linearly spaced in the outer regions of the cylinder. This scheme of radial grid has been arrived at by keeping the optical depth related inaccuracies under check, for the entire wavelength region considered for the radiation transfer. Both the near logarithmic and the linear grids are matched by ensuring that radial cell size, $\delta r(n_r)$, is a smooth function of n_r. For modelling attempts of IRAS 19181 + 1349 a grid of 600 points in axial direction and of 25 points in radial direction were employed.

The relevant calculations are represented in equations (l)–(7). For clarity the frequency suffix has been dropped from all the terms, though calculations are performed for each frequency grid point. The quantities in angled brackets are averages over the dust size distribution. The code is simplified and optimized in many ways in view of the memory requirements and speed. Initially, factors totally dependent on the geometry of the problem and which do not change in each iteration are calculated. These include the optical depth terms and the geometric integrals involved in computations of radiation received by a unit volume element of the ring *i* from the ring *j* (equation (1)). The geometric symmetry is such that such terms only depend on the axial separation between the two rings and their respective radii. The

total flux absorbed and scattered by unit volume in each ring due to radiation from other rings is then calculated (equation (2)). Also, the radiation received by a unit volume element of the ring i from embedded sources depends only on the geometry and the optical depth per unit length of the particular model and hence is fixed over iterations (equation (3)). The equilibrium temperature of the dust grains (on the median circle) for any particular annular ring is calculated using an iterative scheme by equating the power radiated by the dust (equation (6)) to the power absorbed (equation(7)). The latter is contributed by the embedded exciting sources (attenuated by the line of sight dust) as well as secondary emission and scattering from dust grains in all other annular rings (equation (4)). This simplifies the calculations leading to a set of coupled equations with only two parameters per ring, temperature (T_i) and F^i changing from iteration to iteration thus greatly reducing the memory requirements.

$$F_r^{i,j} = \sum_{\theta} \frac{e^{-\langle \pi a^2 Q_{\text{ext}} \rangle n_d d_{i,j,\theta}}}{d_{i,j,\theta}^2} \left[\langle \pi a^2 Q_{\text{abs}} \rangle n_d B(T_j) + \frac{\langle \pi a^2 Q_{\text{scat}} \rangle F^j}{4\pi \langle \pi a^2 Q_{\text{ext}} \rangle} \right] \Delta r_j \, \Delta z \, \Delta\theta r_j, \tag{1}$$

$$F_r^i = \sum_{j=1, j\neq i}^{n_r \times n_z} \langle \pi a^2 Q_{\text{ext}} \rangle n_d F_r^{i,j}, \tag{2}$$

$$F_s^i = \sum_{s=1}^{2} \langle \pi a^2 Q_{\text{ext}} \rangle \frac{\pi R_s^2 F_s}{r_{i,s}^2} e^{-\langle \pi a^2 Q_{\text{ext}} \rangle n_d (r_{i,s} - r_{c,s})}, \tag{3}$$

$$F^i = F_r^i + F_s^i, \tag{4}$$

$$F_e^i = \langle \pi a^2 Q_{\text{abs}} \rangle 4\pi B(T_i), \tag{5}$$

$$P_{\text{Emitted}} = \int F_e^i d\nu, \tag{6}$$

$$P_{\text{Absorbed}} = \int \frac{\langle \pi a^2 Q_{\text{abs}} \rangle}{\langle \pi a^2 Q_{\text{ext}} \rangle} (F_r^i + F_s^i) d\nu, \tag{7}$$

Where

$d^2 \equiv (z_i - z_j)^2 + (r_i)^2 + (r_j)^2 - 2 r_i r_j \cos(\theta)$
θ is azimuthal angular difference between two volume unit under consideration.
$r_{i,s}^2 \equiv (z_i - z_s)^2 + r_i^2$
$r_{c,s}^2 \equiv$ dust cavity radius for source s.
$Q_{\text{ext}} \equiv Q_{\text{abs}} + Q_{\text{scat}}$
$n_d \equiv$ number density of dust grains.
$T_i \equiv$ Temperature of dust in ring i.
$F_s \equiv$ surface flux spectrum for source s.
$F_r^i \equiv$ Flux absorbed and scattered by unit volume of ring i, due to other rings.
$F_s^i \equiv$ Flux absorbed and scattered by unit volume of ring i, due to sources.
$F^i \equiv$ Total flux absorbed and scattered by unit volume of ring i.
$B(T) \equiv$ Planck function.

The dust temperature for each annular ring, leading to the temperature distribution throughout the cloud, is determined iteratively. Initially (the very first iteration), only

the two embedded sources power the heating of the grains. From the second iteration onwards the effects of secondary heating and scattering are taken into account. In this iterative procedure the temperature of each annular ring is gradually updated in each iteration satisfying the condition $P_{\text{Emitted}} = P_{\text{Absorbed}}$. The iterations are continued till the fractional changes in absorbed power for each annular ring, between successive iterations, reduces below the convergence criteria.

The emergent intensity distribution, as seen by a distant observer is predicted by integrating the emitted and scattered radiation along relevant lines of sight and taking account of extinction due to the line of sight optical depth. This spatial intensity distribution at any selected wavelength is convolved with the relevant instrumental beam profile (PSF) for direct comparison with observations.

Before applying the scheme developed above to any astrophysical source it is necessary to verify its reliability and quantify its accuracy. For this, we simulate the case of a single embedded source by "dimming" one of the two sources thereby keeping our original code intact during the test runs. This simulates a single exciting source embedded on the symmetry axis of the uniform density cylindrically shaped cloud. The size of the cylindrical cloud has been chosen such that the results from other codes using spherically symmetric geometry could be compared effectively. We have used the well established code CSDUST3 (Egan, Leung & Spagna 1988) for such a comparison.

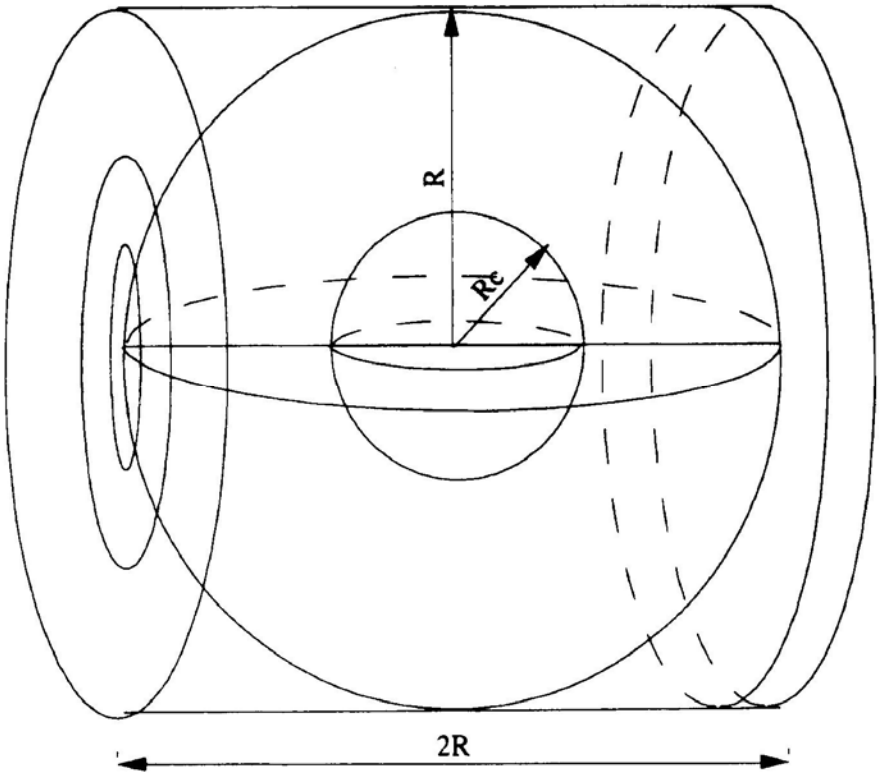

Figure 2. The geometrical dimensions of the spherical (CSDUST3) and the cylindrical modelling schemes under comparison.

The radius of the cylindrical cloud used for comparing our code with the spherically symmetric code was identical to that of the "equivalent" spherical cloud. The length of the cylinder (along z-axis) is twice this radius. Fig. 2 is a schematic of the cylindrical as well as the spherical models.

The assumed radius of the cylinder, as well as the sphere, was 2 pc. The radius of the dust free cavity near the source was adopted to be .01 pc. In order to make the comparison possible and effective, all model parameters were made identical for both the codes. These include: dust size distribution; dust number density (and hence total optical depth along the line of sight between the embedded source and the distant observer). A typical dust composition has been assumed consisting of 50 % Graphite and 50 % Astronomical Silicate. The embedded energy source was assumed to be a single ZAMS 05 star with luminosity $9.0 \times 10^5 L_{\odot}$. Some parameters were explored (e.g. optical depth) to study regions of validity with specified accuracy, by changing them identically for both the schemes. A large range in the optical depth τ_{100},viz., 1.0×10^{-3} to 0.1 was covered during the test runs.

The comparison of the emergent spectral energy distributions (SEDs) from both the schemes for various optical depths are shown in Fig. 3. It may be seen that the SEDs match quantitatively over a wide range of optical depths from mid-infrared to millimeter wavelengths. There are some differences at near infrared wavelengths; this is to be expected from the differences in the cell sizes and the geometry. However, since the main motive of the present study is to interpret measurements in the wavebands beyond the mid-IR, our code may be considered to be satisfactory. From this comparative study we conclude that our code is accurate up to an optical depth corresponding to $\tau_{100} \approx 0.06$ ($\tau_{1\ \mu m} \approx 6$), for the present choice of grid points. This limit already corresponds to much denser clouds than generally found in the Galactic star forming regions.

3. IRAS 19181 +1349

The Galactic star forming region IRAS 19181 + 1349 is an IRAS Point Source Catalog (IRAS PSC) source associated with the radio source G48.60 + 00. The presence of a radio continuum suggests ongoing high mass star formation in the region. This source has been resolved in two components in the 210 μm map (Fig. 4b) generated from the observations using the TIFR 1-meter balloon borne telescope (Karnik *et al.* 1999). Although a single IRAS PSC source is associated with this star forming region, the HIRES processing of the IRAS survey data has led to the resolution of these two sources in the 12 and 25 μm band maps. The map at 12 μm is shown in Fig. 4(a). We denote FIR peak with higher R.A. as S1 and that of lower R.A. S2, respectively, hereafter. Zoonematkermani *et al.* (1990) have reported four compact radio sources 48.603 + 0.026, 48.609 + 0.027, 48.606 + 0.023 and 48.592 + 0.044 in this region. The first three of these closely match S1 (positionally) while the fourth one matches S2. Kurtz, Churchwell & Wood (1994) have also reported three ultracompact HII regions which are positionally close to S1, indicating ongoing high mass star formation in a very early evolutionary phase. This source has also been observed using the ISOCAM instrument onboard Infrared Space Observatory (ISO) using seven filters centred at four PAH features and 3 continuua (3.30, 3.72, 6.00, 6.75, 7.75, 9.63, 11.37 μm; Verma *et al.* 1999). The ISOCAM images at all the bands

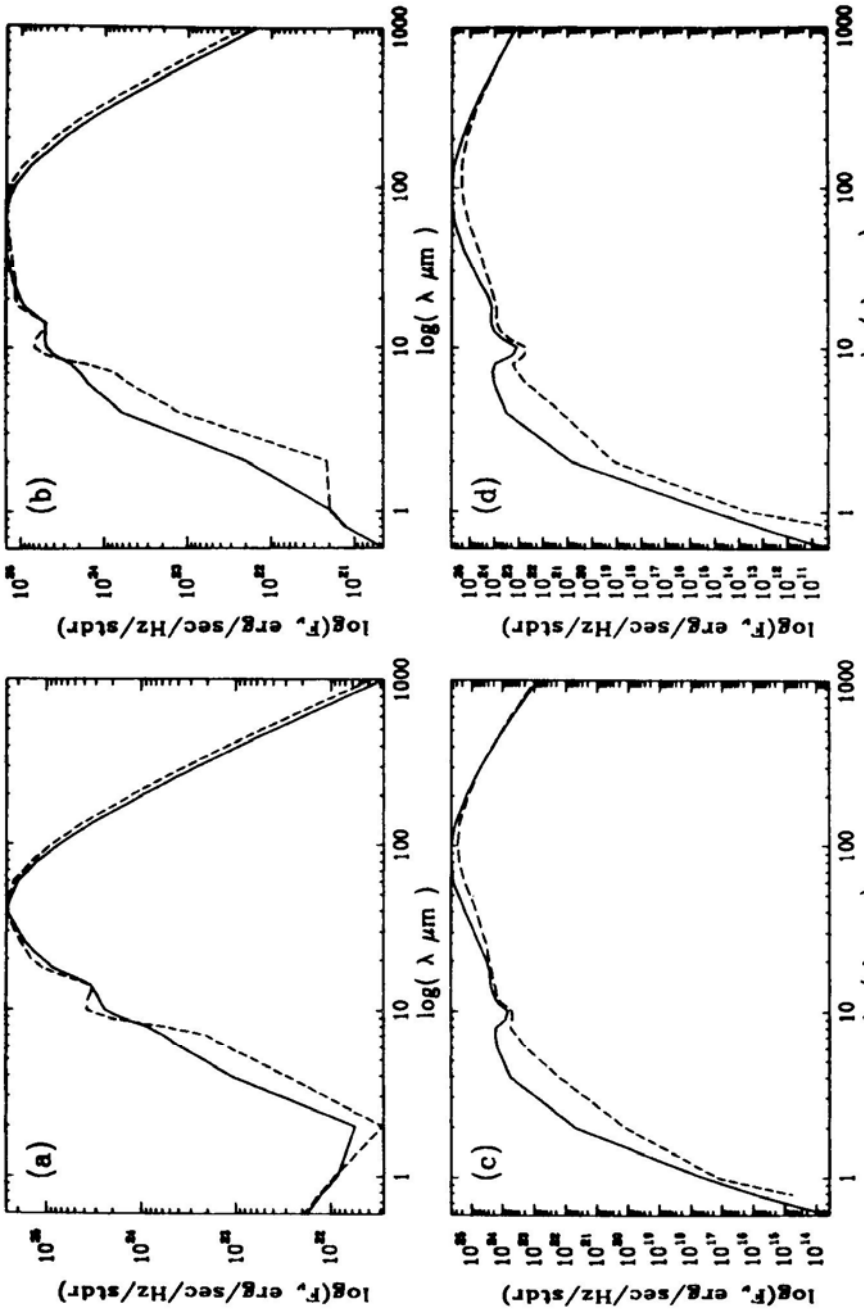

Figure 3. The comparison of predicted SEDs from the two schemes. The continuous lines represent CSDUST3 while the dashed lines refer to our scheme. The SEDs correspond to τ_{100} of – (**a**) 0.001, (**b**) .01, (**c**) .06 and (**d**) 0.1.

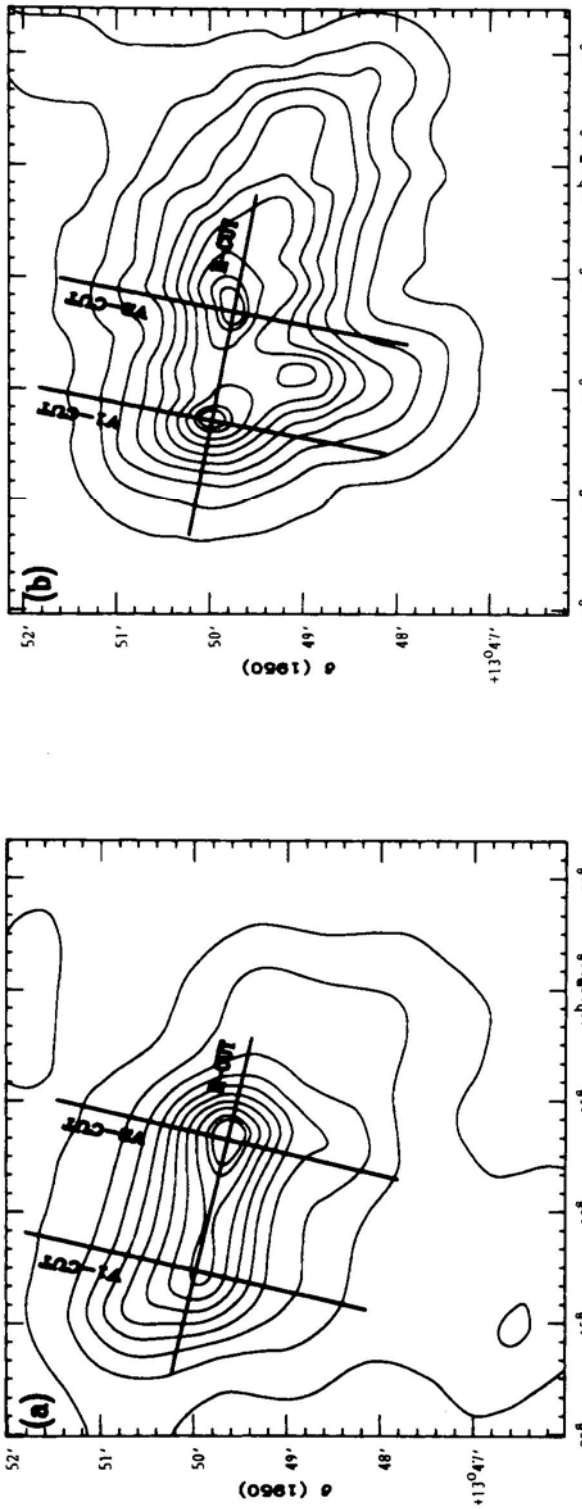

Figure 4. The infrared maps of IRAS 19181 + 1349 at (**a**) 12 μm and (**b**) 210 μm. Contours are drawn at 10, 20, 30, 40, 50, 60, 70, 80, 90, 95 per cent of peak intensities 26 and 303 Jy/(arc min)^2 respectively. The straight lines show the positions of radial and axial cuts compared with the model in Fig.6.

show two extended prominent complexes with multiple peaks. Hence from radio as well as mid to far infrared observations the nature of IRAS 19181 + 1349 is established to be of double embedded source/cluster type.

This seems to be an ideal astrophysical example of interstellar cloud with two embedded sources. We attempt to model this source with our scheme to extract important physical parameters about the embedded energy sources as well as the intervening interstellar medium in IRAS 19181 + 1349. Next we describe the results obtained from such a study.

The source IRAS 19181 + 1349 was considered as a cylindrical dust cloud with two protostellar/ZAMS stellar sources embedded along the axis of the cylinder. The size of the cloud and the dust density were used as free parameters. The sum of the luminosities of the two embedded sources is determined by integrating the observed spectral energy distribution (SED). The observations used for this integration includes the four IRAS bands and the two TIFR bands. This total luminosity is treated as an observational constraint in the modelling. Further observational constraints include: the shape of the SED which was obtained from HIRES-IRAS, TIFR and ISO observations, and the structural morphology of IRAS 19181+1349 as reflected by the isophote contours of the high angular resolution maps at 12 and 210 μm. As our code does not deal with the physics of PAH emission at present, only the data in ISOCAM filters sampling the continuua have been used to constrain the model.

Two completely independent approaches have been followed in modelling IRAS 19181 + 1349 using our scheme. In each approach, all the model parameters are floated to obtain the best fit model. The parameters fine tuned to achieve the best fit model are: luminosities of individual embedded sources; geometrical size details of the cylindrical cloud (including the size of the cavity); and the dust density/optical depth. The main aim of the first approach is to optimize the fit to the observed SED (hereafter $M_{\rm SED}$)· The second approach optimizes the one dimensional radial intensity profiles at selected wavelengths covering mid to far infrared bands (hereafter $M_{\rm RC}$). The radial profiles are taken along geometrically interesting axes, viz., along the line joining the two embedded sources and along lines perpendicular to the cylinder axis and passing through either of the two sources (see Fig. 4). Whereas the first approach gives precedence to the overall energetics the latter gives more importance to the structural details in the isophotes, particularly close to the embedded sources and the region between them. The actual reality may lie somewhere in between these two approaches.

The axes where radial cuts have been taken are also displayed on the 12 and 210 μm maps in Fig. 4. The fits to the observed SED for both $M_{\rm SED}$ as well as $M_{\rm RC}$ are presented in Fig. 5. Comparison of the radial cuts at 12 and 210 μm along the three axes between the observed maps and the best fit $M_{\rm RC}$ model is shown in Fig. 6.

The $M_{\rm SED}$ approach fits the observed SED very well right through near-IR to sub-mm, though there is some discrepancy in the far-IR (Fig. 5). On the other hand, fit from the $M_{\rm RC}$ approach is reasonable for $\lambda \geq 25$ μm only. The radial cuts at 210 μm fit the observed data very well. However, at 12 μm although the model predictions qualitatively agree with the data, there are discrepancies viz., the extent of emission along vertical axis and the relative contrast of the minima between the two sources. From the above it appears that the $M_{\rm RC}$ approach is very sensitive to the exact distribution of sources, specially at shorter wavelengths (which traces much hotter

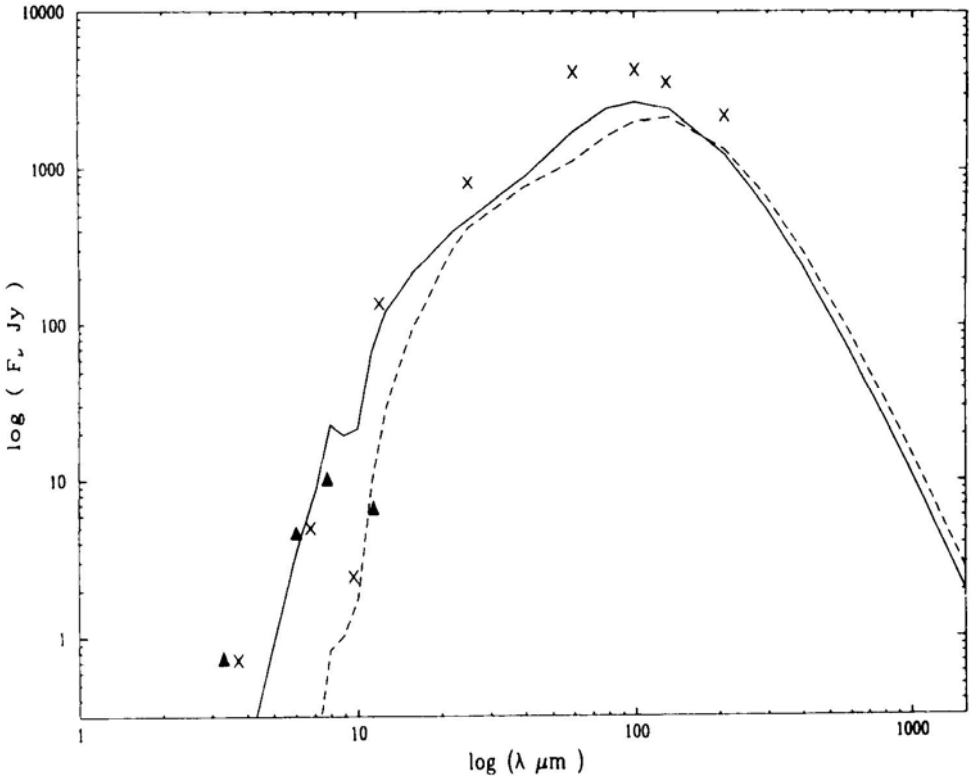

Figure 5. The SEDs for the source IRAS 19181 + 1349 predicted by the radiative transfer models. The solid and the dashed lines represent best fit models from the M_{SED} and M_{RC} approaches, respectively. The crosses represent HIRES (processed IRAS), T1FR and ISOCAM continuum flux densities. The solid triangles represent ISOCAM flux densities measured through filters centered on the PAH features.

dust and hence physically closer to the exciting source). One possibility for these discrepancies is that our model is too simplified than the actual reality in IRAS 19181 + 1349. For example there may be a distribution of sources along the line joining S1 and S2 and away from it, but with more concentration near S1 and S2. This way the discrepancy for M_{RC} in the mid-IR part of SED will also be explained. Assuming a constant dust density in the cloud may also be a major reason for the discrepancy.

The results of modelling IRAS 19181 + 1349 in the form of best fit model parameters as found under both the approaches (M_{SED} and M_{RC}) are presented in Table 1. The value of τ_{100} found from both the approaches are well within the range of validity of the code, i.e. much less than 0.06. The luminosities of both the sources are similar and the dust is found to be predominantly (80%) Silicate with no Silicon Carbide, the rest (20%) being Graphite. It is remarkable that almost all the important parameters are identical as found from both the approaches. The only difference is seen at the physical sizes of the cavities. The remarkable similar physical parameters obtained from both the approaches gives good confidence on the derived astrophysical parameters for the Galactic star forming region IRAS 19181 + 1349.

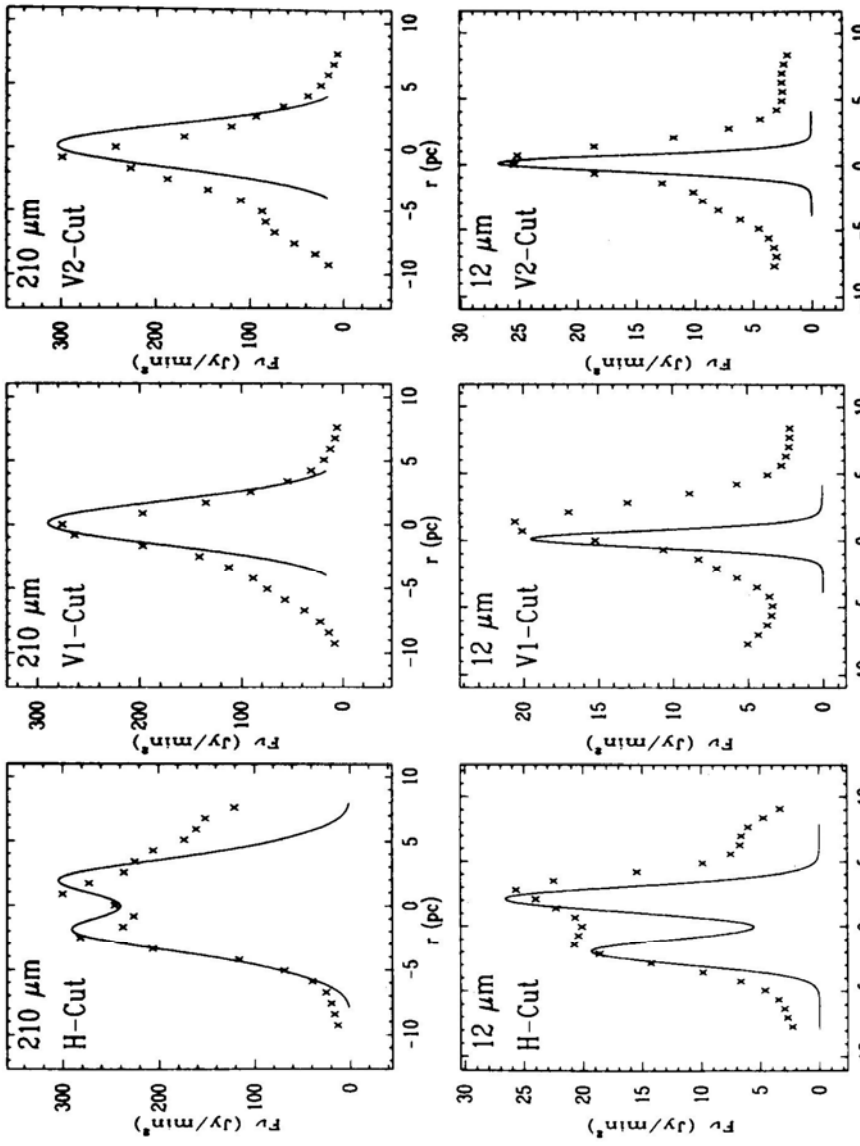

Figure 6. The axial and radial cuts predicted by the radiative transfer model (M_{RC}) (continuous lines), compared with the observations (crosses), at 12 and 210 μm. The spatial positions of the cuts are displayed in Fig. 4.

Table 1. Best fit model parameters for M_{RC} and M_{SED} schemes.

Parameter	M_{RC}	M_{SED}
R_c (pc)(Source 1)	0.11	0.03
R_c (pc)(Source 2)	0.09	0.01
$L_{source1}$ ($L_\odot$)	6.2×10^5	6.2×10^5
$L_{source2}$ ($L_\odot$)	6.3×10^5	6.3×10^5
$\tau_{100\,cyl}$ (/pc)	0.024	0.028
R_{cyl} (pc)	2.1	1.8
Graphite : Silicate : SiC	20 : 80 : 0	20 : 80 : 0

This scheme will be useful to model other similar double sources. A huge database from ISO (SEDs as well as images leading to radial cuts, covering 2.5 to 200 μm of the spectrum) will soon become available for such modelling.

4. Summary

High angular resolution mid/far infrared maps of many Galactic star forming regions show evidence for multiple embedded energy sources in the corresponding interstellar clouds. With the aim of studying such star forming regions with two embedded sources, a scheme has been developed to carry out radiative transfer calculations in a uniform density dust cloud of cylindrical geometry, and which includes the effect of isotropic scattering in addition to the absorption and emission processes. In addition to the luminosities of the two embedded energy sources, the cylindrical cloud size, separation between the two sources, dust density and composition are the parameters for the modelling. The accuracy of our scheme has been tested by comparing the results with a well established 1-D code.

An attempt was made to model the Galactic star forming region associated with IRAS 19181 + 1349 which shows double peaks in the mid- and far-infrared maps using the scheme described here. Two independent approaches were employed to find the best fit models for IRAS 19181 + 1349. Whereas the first (M_{SED}) approach aims to fit the observed SED best; the latter (M_{RC}) aims to fit the radial intensity profiles (along a few important axes) at mid-to far-infrared wavebands. Interestingly most of the crucial model parameters like luminosities, effective temperatures, dust composition, optical depth etc. turn out to be identical under both the approaches.

Acknowledgements

It is a pleasure to thank the members of the Infrared Astronomy Group of T.I.F.R. for their encouragement.

References

Egan, M. P., Leung, C.M.,SpagnaJr, G. F.,1988, *Computer Physics Communications,* **48**, 271.
Ghosh, S. K., Tandon, S. N., 1985, *Mon. Not. R. Astr. Soc.,* **215**, 315.
Karnik, A. D., Ghosh, S. K., Rengarajan, T. N., Tandon, S. N., Verma, R. P, 1999, *Bull. Astr. Soc. India*, **27**, 167.

Kurtz, S., Churchwell E., Wood, D. O. S. 1994, *Astrophys. J. Suppl.*, **91**, 659.
Laor, A., Draine, B. T., 1993, *Astrophys. J.*, **402**, 441.
Mathis, J. S., Rumpl, W., Nordsieck, K. H. 1977, *Astrophys. J.*, **217**, 425.
Rouan, D.,1979, *Astr. Astrophys.*,**79**, 102.
Verma, R. R, Ghosh, S. K., Karnik, A. D., Mookerjea, B., Rengarajan, T. N. 1999, *Bull. Astr. Soc. India,* **27**, 159.
Zoonematkermani, S., Helfand, D. J., Becker, R. H., White, R. L., Perley, R. A., 1990, *Astrophys. J. Suppl*, **74**, 181.

J. Astrophys. Astr. (1999) **20**, 37–50

Determination of Linear Polarization and Faraday Rotation of Pulsar Signals from Spectral Intensity Modulation

P. S. Ramkumar & A. A. Deshpande , *Raman Research Institute*, *C.V Raman Avenue*, *Bangalore 560 080, India.*

Received 1999 March 19; accepted 1999 June 24

Abstract. Most of the known pulsars are sources of highly linearly polarized radiation. Faraday rotation in the intervening medium rotates the plane of the linear polarization as the signals propagate through the medium. The Rotation Measure (RM), which quantifies the amount of such rotation as a function of wavelength, is useful in studying the properties of the medium and in recovering the intrinsic polarization characteristics of the pulsar signal. Conventional methods for polarization measurements use telescopes equipped with dual orthogonally polarized feeds that allow estimation of all 4 Stokes parameters. Some telescopes (such as the Ooty Radio Telescope) that offer high sensitivity for pulsar observations may however be receptive to only a single linear polarization. In such a case, the apparent spectral intensity modulation, resulting from differential Faraday rotation of the linearly polarized signal component within the observing bandwidth, can be exploited to estimate the RM as well as to study the linear polarization properties of the source. In this paper, we present two improved procedures by which these observables can be estimated reliably from the intensity modulation over large bandwidths, particularly at low radio frequencies. We also highlight some other applications where such measurements and procedures would be useful.

Key words. Stars: neutron—pulsars—interstellar medium: Faraday rotation—telescope: polarization.

1. Introduction

Pulsar signals are generally weak (with average flux densities ranging from a few milli-Jansky to a few Jansky), with a high degree of linear polarization. The position angle of the linearly polarized component changes as a function of longitude within the pulse in a manner that depends on the geometry of the spin axis and magnetic poles of the pulsar relative to the observer's line-of-sight (Radhakrishnan & Cooke 1969). At a given longitude, the average polarization seems to have a good long-term stability in most cases, but the apparent plane of linear polarization rotates as function of frequency across the band due to Faraday rotation in the intervening medium. The extent of the rotation θ is given by

$$\theta = \mathrm{RM}\,\lambda^2 \tag{1}$$

Where λ is the wavelength of observation and RM is the Rotation Measure ($= \int_0^D neB_{\|}dl$ where n_e is the electron density, $B_{\|}$ is the line-of-sight component of the magnetic field and D is the distance to the source). It is clear from this equation that by measuring the polarization position angles at different frequencies spanning a sufficiently wide band one can estimate the Rotation Measure (RM). Usually, polarization measurements use antennas with dual orthogonally polarized feeds. By measuring the Stokes parameters across the spectrum or at several suitably separated frequencies, the amount of Faraday rotation is determined. At low radio frequencies, the Faraday rotation of the position angle becomes large enough to be able to measure, the differential rotation within even moderate bandwidths. In addition, the pulsar signal is generally stronger at lower frequencies, although at very low frequencies the strong galactic background radiation seriously affects the sensitivity of these measurements. Thus such measurements require a suitable observing frequency, large bandwidths and sensitive telescopes to achieve the required accuracy. The number of pulsars for which RM has been estimated is only about 260 (Taylor *et al.* 1993) out of about 1000 known pulsars. Most of these 260 pulsars are strong sources and hence relatively easy to study also for their polarization characteristics. To extend such studies to weaker pulsars and at low radio-frequencies, large telescopes operating at suitable frequencies are required. Some telescopes may have large collecting areas, but are receptive to only a single linear polarization. In such cases, an indirect way exploiting the effect of Faraday rotation can be used for studying the linear polarization properties of continuum sources. The basic principle involved in this type of measurements is outlined below.

For simplifying the following discussion, we assume that the RM contribution of the interstellar medium and the ionosphere are constant over the period of observation and consider the time-varying effects of these media later. When a 100% linearly polarized wave is incident on a linearly polarized antenna, the amount of power received by the antenna depends, among other things, on the angle ζ between the directions of polarization of the wave and the antenna feed as

$$P_{\text{received}} = \frac{P_{\text{lin}}}{2}[1 + \cos(2\zeta')] \tag{2}$$

where P_{lin} is the linearly-polarized incident power. The incident polarization position angle changes at different frequencies within the observed bandwidth due to the Faraday rotation in the medium between telescope and the source. Then, even if the power radiated by the pulsar at all frequencies remained the same, the power received by a linearly polarized antenna would show a modulation across the band as shown in Fig. 1. The sampled version of the power spectrum is given by

$$P(f_L + i\Delta f) = A_0 + A_1 \cos\left\{2\left(\zeta + \left(\frac{\text{RM}c^2}{(f_L + i\Delta f)^2}\right)\right)\right\} \quad \text{for } i = 0, 1, \ldots, N \tag{3}$$

where c is velocity of light, f_L is the lower edge-frequency of the spectrum, Δf is the bandwidth of each frequency channel, i is the channel number, A_0 is the average power, A_1 is the linearly polarized power, ζ is the intrinsic position angle of radiation relative to the antenna polarization angle and the second term in the argument is the total Faraday rotation ($\theta_i = \text{RM}\lambda_i^2$). The following can be determined from the modulation:

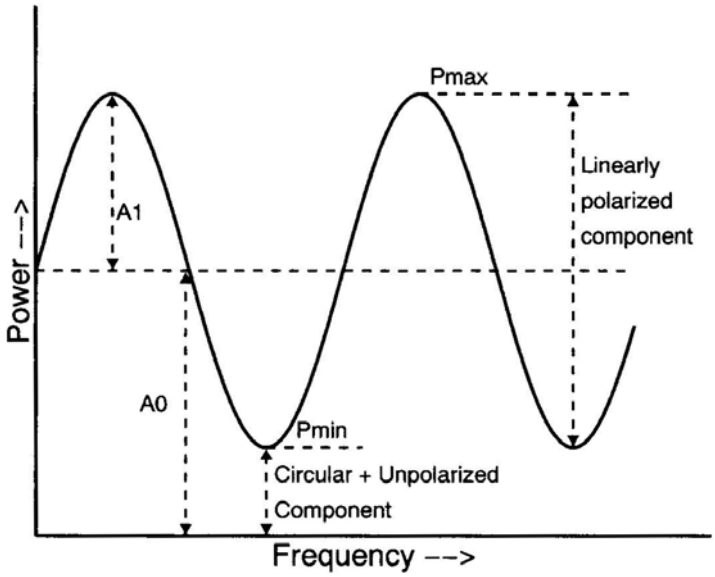

Figure 1. The modulation due to Faraday rotation in the spectrum of the power received by a single linearly polarized antenna.

(**1**) The degree of linear polarization, d_1, can be determined by simply measuring the depth of modulation, as

$$d_1 = \frac{P_{\max} - P_{\min}}{P_{\max} + P_{\min}} = \frac{(A_1)}{(A_0)} \tag{4}$$

where $P_{\max}$ and $P_{\min}$ are the maximum and minimum values of the apparent spectral power contribution, respectively.

(**2**) Byobserving with a bandwidth B over which the modulation completes X_0 cycles, the Rotation Measure can be estimated as

$$\mathrm{RM} = \frac{X_0 \pi}{c^2}\left[\frac{1}{f_L^2} - \frac{1}{(f_L + B)^2}\right]^{-1}. \tag{5}$$

(**3**) The phase(ϕ) of the modulation pattern (i.e. the argument of the cos term in equation (3)) depends directly on the intrinsic position angle of the radiation. Therefore, at a fixed frequency, the variation of ϕ as a function of the pulse longitude can be used to trace the intrinsic sweep of the position angle during the rotation of the pulsar. Using the ϕ value at a reference frequency, say f_L, ζ can be estimated as

$$\zeta = \frac{\phi_0}{2} - \left(\frac{\mathrm{RM}c^2}{f_L^2}\right) \tag{6}$$

where $\phi = \phi_0$ at $f = f_L$.

(4) Knowing the dispersion measure (DM in pc cm^{-3}) and the rotation measure (RM in rad m^{-2}), the mean line-of-sight component of the magnetic field (weighted by the electron density) can then be estimated from the relation

$$\langle B_{\parallel} \rangle = 1.235 \left[\frac{\mathrm{RM}}{\mathrm{DM}}\right] (\mu\,\mathrm{Gauss}). \tag{7}$$

The basic method outlined above has been used earlier (Sulemanova *et al.* 1988) for polarization measurements of 18 pulsars. However, they model the modulation phase to vary linearly with the radio frequency which does not account properly for the non-linear dependence of ϕ (see equation 3), particularly at low radio-frequencies and over large fractional bandwidths.

A modified, more direct method to estimate $\langle B_{\parallel} \rangle$ along the directions to pulsars has also been used by Smirnova & Boriakoff (1997). In this method, the data are not dedispersed and the spectral modulations translated to an equivalent temporal modulation (through the dispersion law) are monitored across the pulse longitude. This treatment exploits the similarity between the non-linear frequency dependences of Faraday rotation and dispersion making the temporal modulation phase a linear function of time. Despite the elegance and simplicity of this procedure, it unfortunately suffers from several disadvantages. The depth of the modulation decreases with large position angle sweeps across the pulse, and with smoothing by the finite pulse width. The method is not applicable for 'continuous' sources, and for pulsed sources it has a poorer signal-to-noise ratio than potentially available. And finally, the modelling of various effects in this work is less than satisfactory.

In this paper we explore two different approaches (which are presented in the next two sections). We also present some test observations and the results obtained using these two approaches. In the last section, we compare the two approaches and discuss their limitations and advantages.

2. Autocorrelation (ACF) domain approach

We begin by noting that it is difficult to detect weak modulation across the band directly from the power spectrum. If the modulation was a simple sinusoid (as in Fig. 1), the domain best suited for studying its parameters would be the 'auto-correlation' domain. In this hypothetical case, it would correspond to a narrow feature in the auto-correlation function obtained through a Fourier Transform of the power spectrum. The 'lag' associated with the feature would be directly proportional to the RM. The relative amplitude of the feature (with respect to the 'zerolag' auto-correlation) would correspond to the fractional linear polarization while the associated phase would depend on the Faraday rotation as well as on the intrinsic position angle. In the method outlined below, we exploit the simplicity of analysis in studying the modulation feature in the auto-correlation domain.

Step 1: Linearization of modulation phase: The argument of the cosine term in equation (3) has a non-linear (inverse square) dependence on the frequency of observation, and consequently on the frequency channel index in the case of a conventional spectrometer with uniform channel spacing. The spectrum can, however, be resampled in a suitable (non-uniform) manner such that the argument varies linearly with the new pseudo-frequency indices, making the modulation appear as a pure sinusoidal wave as a function of the new ordinate. This *linearization* simplifies the analysis and enables the use of linear methods, such as Fourier transforms, to detect and interpret the possible periodic feature directly in terms of the RM, ζ and % linear

polarization. The relation between the new (pseudo-frequency) channel index, j, and the original (true frequency) index i. is given by

$$j = \frac{\alpha\left(\frac{1}{(f_L+i\Delta f)^2} - \frac{1}{f_L^2}\right)}{\left(\frac{1}{(f_L+\Delta f)^2} - \frac{1}{f_L^2}\right)} \tag{8}$$

where the value of α can be chosen to match the new and the original modulation rate at a desired reference frequency (for example, $\alpha = 1$ will give a match at the lower-edge frequencies).

The range of j is that implied by the range of i. For integer values of j, the corresponding i values are not integers in general. Therefore, a suitably interpolated spectral contribution from the original spectrum is to be obtained for a given j. Alternatively, one may use all the samples in the original spectrum by stepping uniformly in i, where for each i, the spectral contribution is suitably shared by new spectral channels, j & j + 1. Then, the share in each of the new channels should be noted, so that the linearized data may then be normalized by the respective counts. In both cases, linear interpolation would suffice provided the phase rotation between two adjacent channels is small (say, less than a radian). It is important to note that this linearization procedure is independent of the rotation measure and depends only on the nature of the non-linearity (see equation 8).

Step 2: Fourier transformation: The 'linearized' spectrum at each longitude is (inverse) Fourier transformed separately to obtain the corresponding ACFs to allow a detailed estimation of the modulation parameters. The magnitude of the ACF is scanned to find the location corresponding to the modulation feature, and the corresponding frequency is used to estimate the RM using equation (5).

Step 3: Estimation of parameters: The peak magnitude of the ACF feature (corresponding to the Faraday modulation) gives the value of A_1, while that at 'zero lag' (the first point in the ACF) provides an estimate of A_0 .The ratio of A_1 to A_0 yields the corresponding fractional linear polarization. An estimate of A_0 is also available directly as simply the mean power in the spectrum. Removal of this mean value from the RF power spectrum before computing the ACF can significantly reduce the side-lobe leakage of the 'zero-lag' component in to the modulation feature. This is useful particularly when the differential Faraday rotation across the band is not large. The modulation phase ϕ_0, is the phase at the peak of the ACF. Then, from equation (6), the intrinsic position angle ζ is estimated at different longitudes. The longitude corresponding to the centroid of the pulse in the A_0 profile is taken as a reference longitude. The final results include the fractional polarization and position angle as functions of the longitude. The uncertainty in estimates of the parameters A_0 and A_1 is given by the rms value of noise in the ACF excluding the two discrete features. When the signal-to-noise ratio (SNR) is large, the formal statistical uncertainty in the modulation feature phase (expressed in radians) is simply the reciprocal of the SNR at the peak of the modulation feature in the ACF. However, a given phase value can result from a wide range of combinations of RM and ζ values, making it difficult to decouple the uncertainties in the two. If the value of ζ is known *apriori*, then the RM estimate can be refined using the observed phase information, provided

the possible 2π ambiguity is resolved. Otherwise, only the modulation 'frequency' (rather than the modulation 'phase') can be used to estimate the RM as mentioned above. The uncertainty in RM estimated in this manner can be related to the signal-to-noise ratio as shown below.

Effect of nonintegral number of modulation cycles: If the number of modulation cycles within the bandwidth is not an integer, then the modulation feature contribution will not be centered on one of the sampled points in ACF with nominal delay resolution (1/B). This is generally the case, necessitating finer sampling of the ACF to avoid appreciable additional errors in the estimation of the location and other parameters of the ACF feature. The required over-sampling is achieved by artificially extending the spectral span by suitably zero-padding the trailing edge of the measured spectrum or by a direct sinc-interpolation of the ACF. The SNR of the ACF feature dictates the optimum over-sampling factor and in turn implies the uncertainty in the RM estimation. It is easy to see that the uncertainty σ_x in estimation of the location x_0 of the feature is given by $\mathrm{sinc}(\sigma_x) = (1 - 1/\mathrm{SNR})$ where x is in units of the nominal delay-resolution (i.e. 1/B). Also, the optimum over-sampling factor is then simply $\sim 1/\sigma_x$ From the estimates of x_0 and σ_x the RM and the corresponding uncertainty can be estimated as

$$\mathrm{RM} \pm \sigma_{\mathrm{RM}} = \frac{\pi}{c^2}\left[\frac{1}{f_L^2} - \frac{1}{(f_L + B)^2}\right]^{-1} (x_0 \pm \sigma_x). \tag{9}$$

An improved RM estimation is possible by using a suitably weighted sum of the ACFs (magnitudes) across the longitude range of the pulse.

Effect of scintillation: The intensity scintillations produced due to the interstellar medium result in superposed random modulations in the RF power spectra, on the scale of the associated decorrelation bandwidth. Correspondingly, the Faraday modulation feature in the ACF is convolved with a "scintillation ACF feature", resulting in reduction of the contrast of the feature of interest and thereby increasing the uncertainty in the RM estimate. This effect is expected to be small for data averaged over spans much longer than the decorrelation time-scales of the scintillations or when the decorrelation bandwidth is much wider than the width of the band observed.

3. NonLinear LeastSquare (NLS) fitting approach

In this approach, an equivalent least-squares fit solution is sought through 'matched filtering' and the best fit values of A_0, A_1, ζ and RM are obtaiIned. In doing so, we will restrict the 'grid' search to RM only, and use a simple procedure to estimate (rather solve for) the other three parameters, for each of the RM values. Thus, for each of the trial RM value, a model spectrum is obtained and compared with observation and the RM corresponding to the best match is sought. For the purpose of the following discussion, we rewrite the equation (3) for the model spectrum, P_m, as

$$P_m = A_0 + A_1[\cos(h)\cos(2\zeta) - \sin(h)\sin(2\zeta)] = \sum_{j=1}^{3} C_j B_j \tag{10}$$

where B_j and C_j are the three basis functions and the corresponding coefficients respectively, and the observed pattern as $P_{obs} = P_m + n$, n being the random noise term. Here, the coefficients, $C_1 = A_0$; $C_2 = A_1 \cos(2\zeta)$ and $C_3 = -A_1 \sin(2\zeta)$, contain the parameters we wish to solve for. The basis functions, namely, $B_1 = 1$; $B_2 = \cos(h)$ and $B_3 = \sin(h)$, can be assumed to be mutually orthogonal functions in principle, except when RM = 0. This orthogonality can be exploited to estimate the coefficients by matched filtering (cross-correlating the P_{obs}, the observed spectral pattern, with the corresponding basis functions). Assuming that we have sampled versions of the relevant patterns/functions, the cross-correlation will estimate some measures, say X_i, such that

$$X_i = \frac{1}{N}\sum_{k=1}^{N} P_{obs}\, B_i = \sum_{j=1}^{3} C_j Y_{ij} \quad \text{for } (i, j = 1, 2, 3). \tag{11}$$

Where Y_{ij} is the cross-correlation between the basis function B_i & B_j computed over the N sampled points and is formally defined as $Y_{ij} = \frac{1}{N}\sum_{k=1}^{N} B_i B_j$ It is easy to see that $Y_{ij} = Y_{ji}$ and ideally, $Y_{ij} = 0$ when $i \neq j$ as a result of orthogonality. In practice, however, given the available span and the sampling of the basis functions, $Y_{ij} = 0$ are non-zero even when $i \neq j$. This is no different from the 'side-lobe leakage' that one refers to in Fourier transforms, for example. However, given the basis functions in h for an assumed RM (and hence Y_{ij}), the coefficients C_i can be solved for in a straight forward way from the above set of equations for X_i. Note that the same formulation would result from conditions for minimization of mean square deviations (of P_m from P_{obs}) with respect to the parameters A_0, A_1 and ζ.

Using these parameter values, a model spectrum is computed and its mean square deviation $\overline{e^2}$ from the observed data is obtained. This procedure is repeated for several trial values of RM in fine enough steps. The best estimate of RM and the other 3 parameters corresponds to the fit with minimum $\overline{e^2}$. The $\overline{e^2}$ value depends on the correctness of the model, as well as on the other sources of uncertainty in the observed pattern as already discussed. The variation of e^2 as a function of changes in the model-parameters (such as RM, A_0, etc.) can be used to estimate the uncertainty in the parameter values. The minimum detectable change in $\overline{e^2}$, i.e. $\Delta_{\overline{e^2}} \cong \overline{e^2}/N_{dof}$ can be attributed to the uncertainties in the parameter values which can be derived. In the present case, the degrees of freedom (N_{dof}) are equal to $N - 4$. The minimum detectable change in the mean-square-error can be expressed in terms of the variance associated with the estimation of individual parameters, to the first order, as

$$\overline{\Delta_e^2} = \frac{1}{N}\sum_{k=1}^{N}\left[\left(\frac{\partial P_m}{\partial A_0}\right)^2 \cdot \Delta_{A_0}^2 + \left(\frac{\partial P_m}{\partial A_1}\right)^2 \cdot \Delta_{A_1}^2 + \left(\frac{\partial P_m}{\partial \mathrm{RM}}\right)^2 \cdot \Delta_{\mathrm{RM}}^2 + \left(\frac{\partial P_m}{\partial \zeta}\right)^2 \cdot \Delta_{\zeta}^2\right]. \tag{12}$$

Ideally, the covariances of the parameters should also be considered (e.g., for the pair RM and ζ), but are ignored here for the sake of simplicity. The entire error on the left-hand side may be associated to one parameter at a time, to get the worst-case formal statistical uncertainty in that parameter. Also, the noise statistics are assumed to be same for all the frequency channels as is usually the case. The estimation of uncertainty in RM does not include the effect of the error in ζ and assumes that there is no ambiguity in the modulation phase in multiples of 2π. In such a case, the maximum error in RM (corresponding to a phase error of $\pm\pi$) is $(\pi c^2 / f_L^2)$

The method outlined above is extended in a straight-forward way to a combined fit over data for a range of pulse longitudes, the only common parameter being the RM.

4. Tests and Results

The processing methods discussed above were tested first using simulated data and then applied to data from pulsar observations. As a first trial, the data on PSR 0740-28, a pulsar with reasonably large RM ($\cong$ 150 rad m^{-2}) and pulse strength ($S \cong 300$ mJy), were obtained using the Ooty Radio Telescope with the pulsar search preprocessor (Ramkumar *et al.* 1994). The spectral data for ~10 minutes (sampled every 0.5 msec) from 256 frequency channels covering a band of 8 MHz (around 327 MHz) were used. For each of the spectral channels, only the deviations from their long term mean power were recorded after 1-bit quantization. The data were aligned by correcting for the dispersion delay gradient across the band, and folded over the pulsar period to improve the signal-to-noise ratio. The folded profiles of all channels were arranged in the form of a time-frequency matrix. Fig. 2 displays a 3-D plot of intensity as a function of frequency and pulse longitude. The spectral channel gain calibration was done using estimates of off-pulse rms deviations. The data were then analyzed using the ACF method. Fig. 3 shows the position angle, total (solid line) and linearly polarized intensity (dashed line) as a function of the pulse longitude. For comparison, Fig. 3(c) displays the observation at 631 MHz by McCulloch *et al.* (1978), made with the 64-m telescope at Parkes, Australia. Comparison of the modulation pattern observed on three consecutive days (at the longitude of the peak of the pulse)

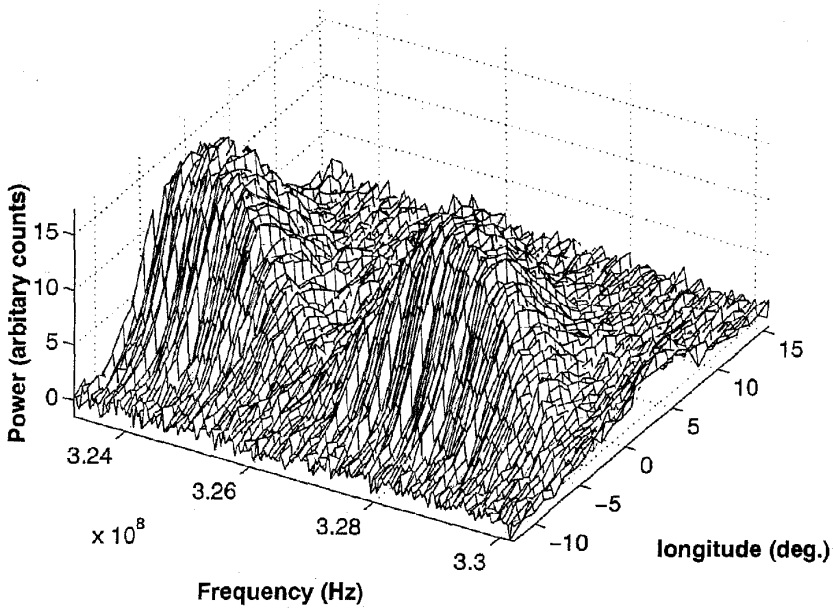

Figure 2. Pulse intensity as a function of frequency and longitude (relative to the pulse centroid) for PSR 0740-28, observed on 19-03-94 at ORT using the Pulsar Search Preprocessor.

indicates an apparent change in RM of about 0·5 rad/m^2 from day to day (shown in Fig. (4)). The rate of change is too fast to be associated with the contribution of the interstellar medium, and is more likely to be due to changes in the RM of the ionosphere. This method and initial results were discussed by Ramkumar & Deshpande (1994).

The tests were repeated on data from subsequent observations of the same pulsar using both the ACF and NLS procedures. The results from the two methods are compared in Fig. 5, where Fig. 5(a) shows the average A_0 components estimated by

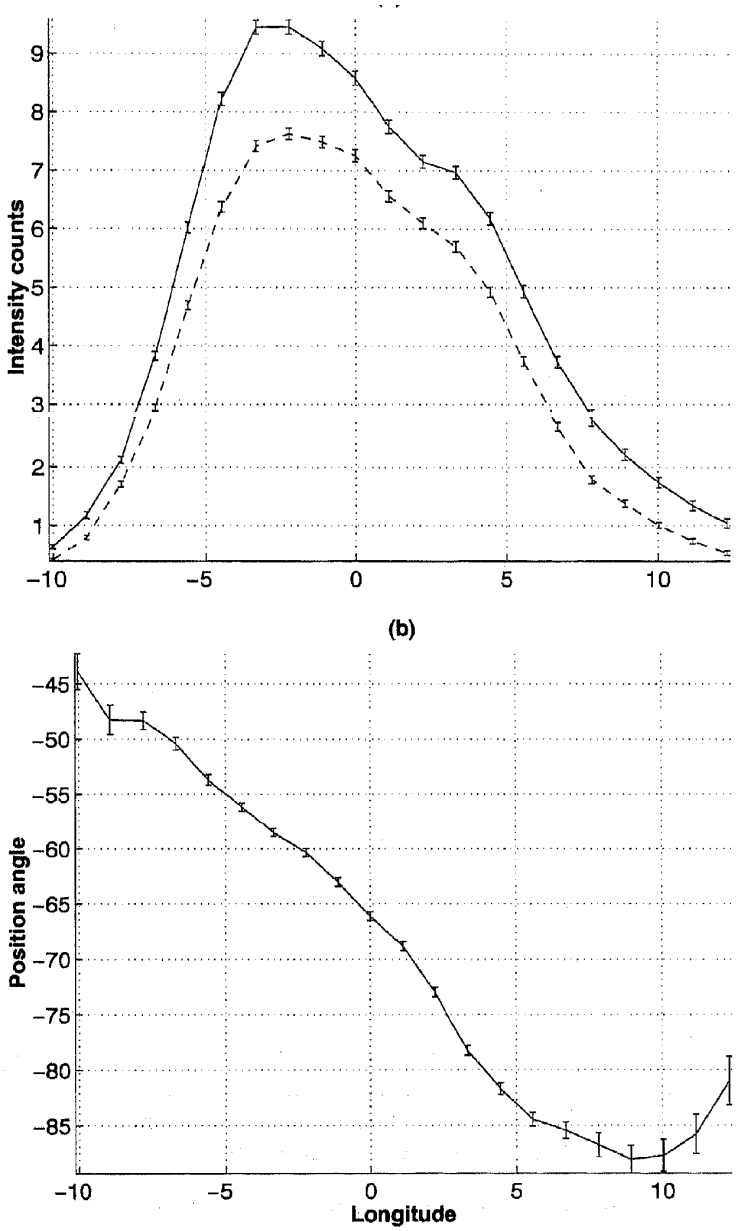

Figure 3. (*Continued*)

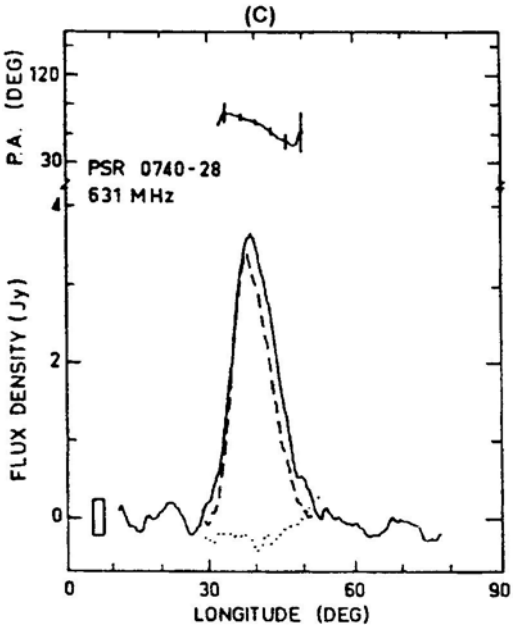

Figure 3(a,b&c). The estimated Position Angle (**a**) and Intensity (**b**) profiles of pulsar PSR 0740-28 from the observations on 19-03-94. Panel (**c)** shows the corresponding profiles at 631 MHz obtained by McCulloch *et al.* (1978) using dual-polarization data.

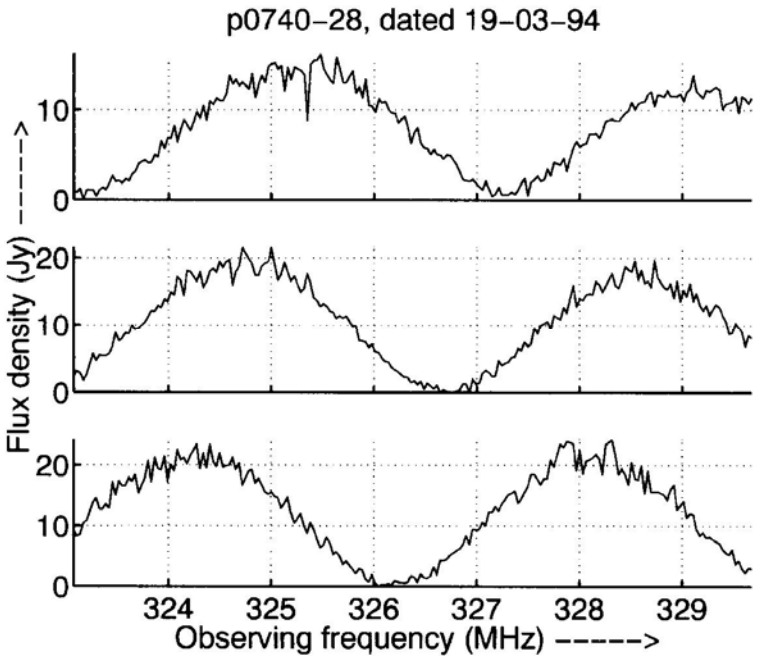

Figure 4. Average power spectra showing modulations due to Faraday rotation observed on three consecutive days (corresponding to the same nominal reference longitude). The observed differences in the modulation phase are possibly due to ionospheric RM changes.

the two methods, Fig. 5(b, c) show the corresponding fractional linear polarization d_L, and the position angle patterns respectively. The estimated value of RM (which also includes the ionosphere contribution) is 152·5 rad/m^2 and 153.5 rad/m^2 (with corresponding statistical uncertainties of 0.007 and 4.35) in the NLS and ACF methods, respectively, as compared to 152 rad/m^2 (excluding the ionospheric contribution) quoted by Hamilton & Lyne (1987).

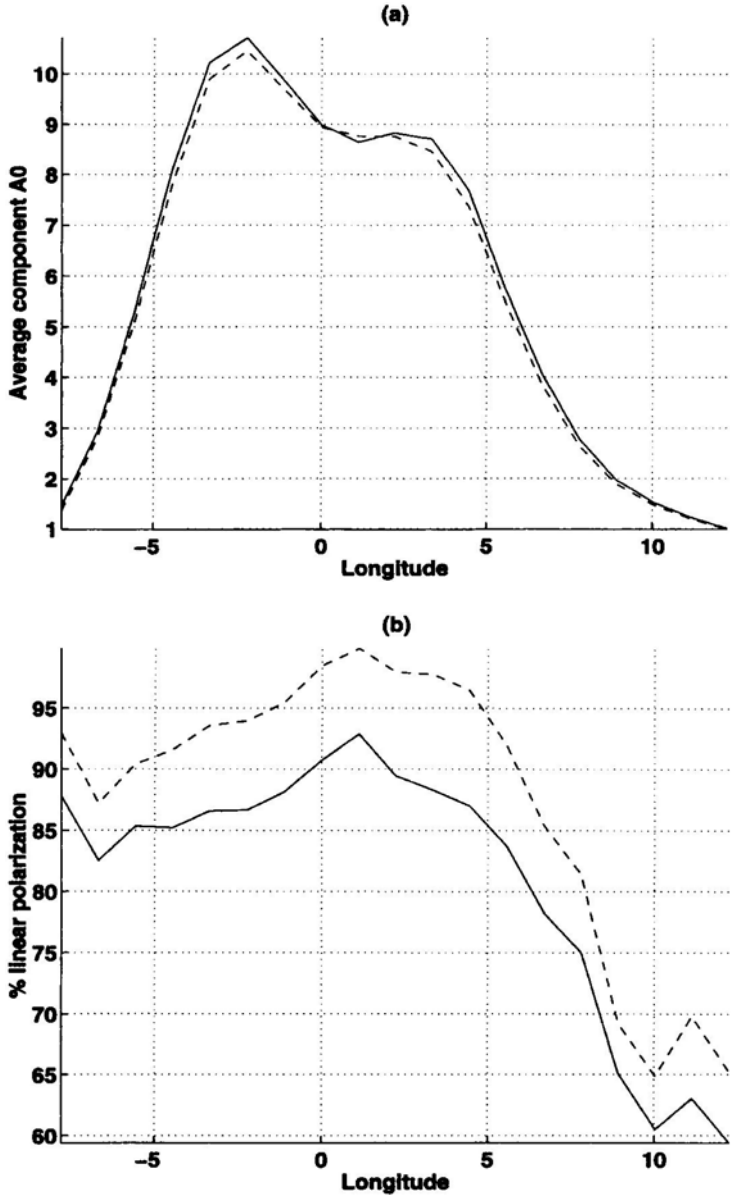

Figure 5. (*Continued*)

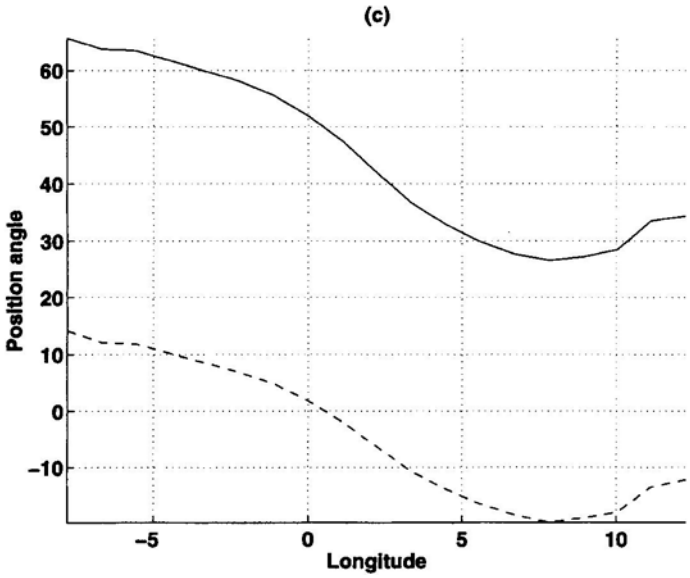

Figure 5 (a,b&c). Profiles corresponding to the best-fit parameters from the two estimation procedures applied to the data of PSR 0740-28 (observed on 15-7-1997). The solid and the dashed lines show the results from the NLS and the ACF methods respectively.

5. Discussion

In the RM determination using the basic method described here, the bandwidth and the number of spectral channels set limits to the range for measurable RM at a given operating frequency. The RM should be large enough to produce at least one cycle of intensity modulation across the band, while it should be less than a value at which one cycle of modulation spans only two frequency channels. In the ACF method, the accuracy in estimation of the parameters is also limited by the fact that "linear" interpolation was used in sharing the power of original samples to those in the linearized domain. This limits the modulation frequency that can be resampled properly, and thereby implies an upper limit for RM up to which good estimation can be made given the bandwidth, operating frequency and SNR. However, a higher order interpolation would greatly reduce this problem. In the second method, there is no such restriction since the pattern non-linearity is implicit and hence the performance is robust. As such, these methods are well suited for observations of high RM pulsars or observations over relatively large bandwidths, where simple sinusoidal approximations to modulation phase may lead to significant errors in the estimation of RM. The ACF method does not need any initial guess of RM, while the NLS method is based on a grid search over a range of RM values. In practice, it may be better to initially use the auto-correlation domain processing to arrive at an estimate of RM, and then use the non-linear fit method to refine the estimate. This approach will save the number of computations substantially, for measurements demanding high accuracy.

The ultimate uncertainty limit for RM and linear polarization measurements is set by the signal-to-noise ratio of the data particularly that of the linearly polarized

component. In the ACF method, the SNR can be enhanced further by averaging the magnitude squares of the ACFs at different longitudes thus ignoring the phase differences and using appropriate weights based on the pulse shape. The pulse longitude resolution can be suitably optimized based on the sweep rate of polarization angle within the pulse. Also, the data time span should be short compared to the typical time scales for apparent changes in the RM contributed by the ionosphere, so as to keep the depolarization due to integration well below that implied by the required RM accuracy. However, to smooth-out the undesirable modulation due to interstellar scintillations, it is desirable to average data over spans much longer than the de-correlation time scales of scintillation.

The observed modulation phase, as already noted, can be attributed to a range of combinations of ζ and RM values. The ability to *distinguish* between relative contributions from RM and ζ terms improves as the bandwidth increases or operating frequency decreases, reducing the range of degenerate combinations of ζ and RM. Since the estimated value of RM is a "weak" function of the reference modulation phase, the estimation accuracy is intrinsically higher for RM measurements compared to those of ζ. The estimation of the intrinsic position angle of the radiation can show large changes due to even a small change in the RM estimate. On the other hand, the estimation accuracy of RM and ζ is much higher in differential measurements, where any 'changes' in the modulation phase are interpreted as changes in only one of the two parameters (i.e. RM or ζ). Thus, the sweep of intrinsic position angle across the pulse (where RM is assumed constant) and the possible variation in RM with time (where the source position angle is constant, a fair assumption in most cases) can both be measured with high accuracy. For a given signal-to-noise ratio, the non-linear least-squares fit method has better performance than the ACF method. This is because the former method uses the complete information (amplitude and phase) of the signal to fit for RM, while in the ACF method only the amplitude information is used for the RM determination.

A comparative analysis of such observations made on a given suitable pulsar on short time spans should provide useful information about any ionospheric RM change as a function of hour-angle and time in general. As such changes are expected to be small, they would be noticeable first in the variation of the reference phase of the modulation. This information should help us in modelling the changes in the ionospheric RM reliably. The basic technique and the estimation procedures, discussed here in the context of pulsars, are also applicable to continuum sources that do not have pulsed radiation.

The 'linearization' technique suggested in the context of the ACF method has very useful applications in many other situations. For example, this linearization approach when applied to pulsar search data over wide bandwidths, would allow the use of Taylor's dedispersion algorithm meant for linear dispersion delay gradients.

Acknowledgements

We thank V. Radhakrishnan for fruitful discussions and his many useful comments on the manuscript. We are also thankful to K. Kishan Rao for useful discussions and for providing support during visits by one of us (PSR) to the Regional Engineering College, Warangal.

References

Hamilton, P. A., Lyne, A. 1987, *Mon. Not. R. Astr. Soc*, **224**, 1023.
McCulloch, P. M., Hamilton, P. A., Manchester, R. N., Abies, J. G. 1978, *Mon. Not. R. Astr. Soc*, **183**, 645.
Radhakrishnan, V., Cooke, D. J., 1969, *Astrophys. J. Lett.*, **3**, 225.
Ramkumar, P. S., Prabu, T., Madhu Girimaji, Markendeyulu, G. 1994, *J. Astrophys. Astr.*, **15**, 343.
Ramkumar, P. S., Deshpande, A. A. 1994, Proc. of the 16th meeting of Astronomical Society of India, Pune 1995, (ed.) V. K. Kapahi, *Bull. Astr. Soc. India*, **23(4)**, 475.
Smirnova, T. V., Boriakoff, V. 1997, *Astr. Astrophys.*, **321**, 305.
Suleimanova, S. A., Volodin, Yu. V., Shitov, Yu. P. 1988, *Sov. Astron.,* **32(2)**, 177.
Taylor, J. H., Manchester, R. N., Lyne, A. G. 1993, *Astrophys. J. Suppl Ser.*, **88**, 529.

J. Astrophys. Astr. (1999) **20**, 51–65

Density and Temperature Diagnostics of Solar Emission Lines from NeV/MgV and SiVII/MgVII Ions

Anita Mohan & B.N. Dwivedi, *Department of Applied Physics, Institute of Technology, Banaras Hindu University, Varanasi 221 005, India.*

P. K. Raju, *Indian Institute of Astrophysics, Bangalore 560 034, India.*

Received 1998 June 5; accepted 1999 June 2

Abstract. We present NeV/MgV and SiVII/MgVII theoretical line intensity ratios as a function of electron density N_e and temperature T_e . These are shown in the form of ratio-ratio diagrams, which should in principle allow both N_e and T_e to be deduced for the emitting region of the solar plasma. We apply these diagnostics in the solar atmosphere, and discuss the available observations made from space. In most cases, however, we deduce N_e and T_e from the computed absolute line intensities in a spherically symmetric model atmosphere of the Sun. Possible future applications of this investigation to spectral data from the Coronal Diagnostic Spectrometer (CDS) on the Solar and Heliospheric Observatory (SOHO) are briefly discussed.

Key words. Solar atmosphere—EUV diagnostics—emission lines—spectroscopic diagnostics.

1. Introduction

Line ratios involving transitions in the ultraviolet (UV) and extreme-ultraviolet (EUV) regions of the spectrum frequently provide excellent temperature and density diagnostics for the emitting or absorbing plasmas. Over the past twenty five years or so, many such diagnostics have been developed for application to astronomical spectra, such as those of the solar transition region/corona and stellar observations from balloon, rocket and satellite-borne experiments (cf., Dwivedi 1994; Mason and Monsignori Fossi 1994). High-quality EUV data obtained from the spacecraft SOHO, provide the motivation for diagnostic applications to the analysis and interpretation of such data.

The usual procedure has been to look for line intensity ratios which are sensitive either to electron density N_e or electron temperature T_e . In various investigations it has been noticed that electron pressure within the chromosphere-corona transition region, and the corona, is either constant or varies slowly with height in the transition region and to some extent in the corona. The observed intensity of a particular line is due to several emitting layers. Each layer would have different electron density and temperature values but electron pressure would be nearly the same in all the emitting layers. The comparison of theoretical ratios with the observed values would then give the effective values of electron density and temperature within the emission regions.

It is, therefore, physically meaningful to study the variation of intensity ratios with electron density (and thus temperature) at constant pressure. This approach has been applied for spectroscopic diagnostics of several solar ions including NeV/MgV and SiVII/MgVII (cf., Dwivedi, Mohan & Raju 1997 and references cited therein).

Under the conditions that obtain in the Sun, the line intensity ratios are thus clearly sensitive to variations in both the electron temperature and density. Hence, in principle, they should only be used to determine N_e or T_e when the other plasma parameter has been independently estimated. In view of this, we investigate the problem from another standpoint. In what follows, we plot several ratio-ratio diagrams, such as log R_1 vs log R_2 and so on, for a grid of (log $N_{e,,}$ log T_e) values appropriate to the solar transition region. Using these figures it is possible to simultaneously determine both the electron temperature and density from the measured/computed values of the ratios. Such a technique has also been used by Keenan *et al.* (1995).

The ions NeV, MgV and SiVII, MgVII have their respective ionic concentrations maximum at about the same temperatures of 2.8×10^5 K and 6.3×10^5 K, respectively. Moreover, their ionization equilibrium curves overlap around the respective temperatures for maximum ionic concentrations. Therefore, these ionic pairs could be used for the electron density and temperature diagnostics of the relevant portions of the chromosphere-corona transition region and also to estimate their relative element abundances. We have computed theoretical line intensities for several NeV, MgV, SiVII and MgVII lines using a model solar atmosphere by Elzner (1976) and assuming values of 3.5×10^{-5}, 3.7×10^{-5} and 3.9×10^{-5} for the elemental abundances (relative to hydrogen) of Ne, Mg and Si, respectively (Meyer 1985). Theoretical intensities have been compared with the available observed quiet-Sun intensities for these lines. However, for want of data, we make use of these computed line intensities to deduce electron density and temperature from the ratio-ratio approach while emphasizing their applications with the SOHO data when available.

In section 2, we briefly describe the line emissivity. Atomic data are discussed in section 3. Observed and computed line intensities are considered in section 4. Electron density and temperature diagnostic aspects are examined in section 5. We make concluding remarks in the last section.

2. Line emissivity

The volume emission coefficient in a radiative transition from the upper level j to a lower level i for an optically thin spectral line is given by

$$\epsilon(\lambda_{ij}) = N_j A_{ji} \frac{hc}{4\pi\lambda_{ij}} \; (j > i) \; \mathrm{ergs\,cm^{-3}\,s^{-1}\,sr^{-1}} \tag{1}$$

where λ_{ij} is the wavelength for the transition $i \rightarrow j$, h is Planck's constant, c the velocity of light and A_{ji} is the spontaneous transition probability. The number density N_j of the emitting level of the ionic species can be parametrised as:

$$N_j(X^{+p}) = \frac{N_j(X^{+p})}{N(X^{+p})} \cdot \frac{N(X^{+p})}{N(X)} \cdot \frac{N(X)}{N(\mathrm{H})} \cdot \frac{N(\mathrm{H})}{N_e} \cdot N_e \tag{2}$$

where X^{+p} denotes the p^{th} ionization stage of the element X, $N_j(X^{+p})$ /$N(X^{+p})$ is the population of level j relative to the total population of the ion X^{+p}, $N(X^{+P})$ /$N(X)$ is

the ionization ratio of the ion X^{+p} . $N(X)$ /$N(H)$ is the abundance of the element X relative to hydrogen which may or may not be constant in the solar plasma. We have assumed $N(H)$ /N_e = 0.8 for the fully ionized plasma. The emissivity can now be expressed as:

$$\epsilon(\lambda_{ij}) = \frac{1.265 \times 10^{-9}}{\lambda_{ij}} \cdot A_{ji} \cdot \frac{N_j(X^{+p})}{N(X^{+p})} \cdot \frac{N(X^{+p})}{N(X)} \cdot \frac{N(X)}{N(\mathrm{H})} \cdot N_e. \tag{3}$$

We denote the line intensity as $I(\lambda_{ij})$ which is the intensity integrated over the line of sight. In the case of two lines emitted from the same ion, the intensity ratio can be expressed as

$$\frac{I(\lambda_{ij})}{I(\lambda_{kl})} = \frac{A_{ji}}{A_{lk}} \frac{\lambda_{kl}}{\lambda_{ij}} \frac{N_j(X^{+p})}{N_l(X^{+p})}. \tag{4}$$

The intensity ratio for the lines emitted from the same volume element but from different elements X and Y is then given by:

$$R = \frac{I(\lambda_{ij})}{I(\lambda_{kl})} = \frac{A_{ji}}{A_{lk}} \cdot r(j,l,X,Y) \cdot s(X,Y) \cdot N(X,Y) \tag{5}$$

Where

$$r(j,l,X,Y) = \frac{N_j(X^{+p})}{N(X^{+p})} \Big/ \frac{N_l(Y^{+q})}{N(Y^{+q})},$$

$$s(X,Y) = \frac{N(X^{+p})}{N(X)} \Big/ \frac{N(Y^{+q})}{N(Y)},$$

and

$$N(X,Y) = \frac{N(X)}{N(\mathrm{H})} \Big/ \frac{N(Y)}{N(\mathrm{H})}.$$

In the second case the intensity ratio depends on the relative ionic concentrations of the elements and their relative element abundances. The ionization equilibrium curves for NeV, MgV and SiVII, MgVII overlap around their respective temperatures for maximum ionic concentrations, respectively. We, therefore, assume that lines from NeV-MgV originate from the same emitting layers; similarly for the SiVII-MgVII ions. This assumption then justifies the use of equation (5) for line intensity ratios. We have solved the steady state equations for the various atomic levels to obtain N_j(NeV)/ N(NeV), N_j(MgV)/N(MgV), N_j(SiVII)/N(SiVII) and N_j(MgVII)/N(MgVII) as a function of electron density and temperature. We have considered the first 15 atomic levels for NeV, MgVII and the first 9 atomic levels for MgV, SiVII ions.

3. Atomic data

The atomic data needed to compute line intensities are the following: (**i**) wavelengths, (**ii**) radiative transition probabilities, and (**iii**) collision strengths. The wavelengths have been taken from Kelly & Palumbo (1973). The wavelengths for a few transitions which are not listed by Kelly & Palumbo have been estimated from the term values

Table 1. Line intensities of Ne V lines (N (Ne) / N (H) = 3.5 ×10^{-5}).

Transition $2s2p^3 \rightarrow 2s^22p^2$	Wavelength (Å)	Intensities (ergs cm^{-2} s^{-1} sr^{-1})	
		Computed (Vernazza & Reeves 1978)	Observed
$^3S^0_1 - {}^3P_0$	357.95	0.77	–
$^3S^0_1 - {}^3P_1$	358.48	2.32	–
$^3S^0_1 - {}^3P_2$	359.39	3.87	102.16
$^1P^0_1 - {}^1D_2$	365.61	2.73	136.11
$^1D^0_2 - {}^1D_2$	416.20	5.67	33.73
$^1P^0_1 - {}^1S_0$	416.82	0.62	–
$^3P^0_1 - {}^3P_0$	480.41	0.95	4.83
$^3P^0_0 - {}^3P_1$	481.28	1.02	–
$^3P^0_1 - {}^3P_1$	481.36	0.76	–
$^3P^0_2 - {}^3P_1$	481.37	1.15	–
$^3P^0_1 - {}^3P_2$	482.98	1.16	6.18
$^3P^0_2 - {}^3P_2$	482.99	3.66	–
$^3D^0_1 - {}^3P_0$	568.42	1.22	–
$^3D^0_1 - {}^3P_1$	569.76	0.85	–
$^3D^0_2 - {}^3P_1$	569.83	2.75	–
$^3D^0_2 - {}^3P_2$	572.11	0.80	8.76
$^3D^0_3 - {}^3P_2$	572.34	4.84	–
$^5S^0_2 - {}^3P_1$	1137.0	0.28	–
$^5S^0_2 - {}^3P_2$	1146.1	0.70	–

Table 2. Line intensities of Mg V lines (N (Mg) / N (H) = 3.7 × 10^{-5}).

Transition $2s2p^5 \rightarrow 2s^22p^4$	Wavelength (Å)	Intensities (ergs cm^{-2} s^{-1} sr^{-1})	
		Computed (Vernazza & Reeves 1978)	Observed
$^1P^0_1 - {}^1D_2$	276.58	2.87	–
$^3P^0_1 - {}^3P_2$	351.09	1.61	–
$^3P^0_0 - {}^3P_1$	352.20	1.45	–
$^3P^0_2 - {}^3P_2$	353.09	5.52	100
$^3P^0_1 - {}^3P_1$	353.30	0.97	–
$^3P^0_1 - {}^3P_0$	354.22	1.28	–
$^3P^0_2 - {}^3P_1$	355.33	1.73	–
$2s^22p^4 \rightarrow 2s^22p^4$			
$^1S_0 - {}^3P_1$	1324.43	0.19	–
$^1D_2 - {}^3P_2$	2783.34	0.13	–

given by Edlén (1982). For NeV and MgVII the various transition probabilities have been taken from Nussbaumer and Rusca (1979) and Aggarwal (1986). Transition probabilities for the SiVII ion have been taken from Bhatia, Feldman & Doschek

Table 3. Line intensities of Mg VII lines (N (Mg) / N (H) = 3.7×10^{-5}).

Transition	Wavelength	Intensities (ergs cm^{-2} s^{-1} sr^{-1})	
$2s2p^3 \rightarrow 2s^22p^2$	(Å)	Computed	Observed
$^3S_1^0 - {}^3P_0$	276.15	2.46	2.93[a]
$^3S_1^0 - {}^3P_1$	276.99	7.37	23.41[a] (blended)
$^3S_1^0 - {}^3P_2$	278.41	12.38	17.56[a]
$^1P_1^0 - {}^1D_2$	280.74	2.53	
$^1D_2^0 - {}^1D_2$	319.02	5.79	
$^1P_1^0 - {}^1S_0$	320.50	0.59	
$^3P_1^0 - {}^3P_0$	363.75	2.45	
$^3P_0^0 - {}^3P_1$	365.17	2.60	
$^3P_1^0 - {}^3P_1$	365.21	2.88	136.11[b] (blended)
$^3P_2^0 - {}^3P_1$	365.23	2.09	
$^3P_1^0 - {}^3P_2$	367.64	9.77	655[b] (blended by Mg^{+8})
$^3P_2^0 - {}^3P_2$	367.68	2.93	
$^3D_1^0 - {}^3P_0$	429.13	2.89	
$^3D_1^0 - {}^3P_1$	431.17	1.88	38.22[b] (blended by Mg^{+7})
$^3D_2^0 - {}^3P_1$	431.32	6.69	
$^3D_2^0 - {}^3P_2$	434.71	1.70	28.32[b] (blended by Ne^{+5})
$^3D_3^0 - {}^3P_2$	434.92	11.42	
$^5S_2^0 - {}^3P_1$	854.70	0.40	
$^5S_2^0 - {}^3P_2$	868.13	0.95	
$2s^22p^2 \rightarrow 2s^22p^2$			
$^1S_0 - {}^3P_1$	1189.82	0.54	–
$^1D_2 - {}^3P_2$	2629.00	0.63	–

[a] Malinovsky & Heroux (1973).
[b] Vernazza & Reeves (1978).

Table 4. Line intensities of SiVII lines (N (Si) / N (H) = 3.9×10^{-5}).

Transition	Wavelength	Intensities (ergs cm^{-2} s^{-1} sr^{-1})	
$2s2p^5 \rightarrow 2s^22p^4$	(Å)	Computed	Observed
$^1P_1^0 - {}^1D_2$	217.83	1.78	–
$^3P_1^0 - {}^3P_2$	272.64	6.0	5.12[a]
$^3P_0^0 - {}^3P_1$	274.17	3.74	–
$^3P_2^0 - {}^3P_2$	275.35	22.39	14.63[a]
$^3P_1^0 - {}^3P_1$	275.67	3.46	2.19[a]
$^3P_1^0 - {}^3P_0$	276.84	4.52	3.66[a]
$^3P_2^0 - {}^3P_1$	278.45	7.14	–
$2s^22p^4 \rightarrow 2s^22p^4$			
$^1S_0 - {}^3P_1$	1049.15	0.40	–
$^1D_2 - {}^3P_2$	2147.35	1.03	–

[a] Malinovsky & Heroux (1973).

(1979). In the case of MgV ion, for a given transition logarithmic value of transition probabilities of SiVII, SIX and ArXI (Bhatia, Feldman & Doschek 1979) were plotted against the inverse atomic number. A linear fit between log A_{ji} and $1/Z$

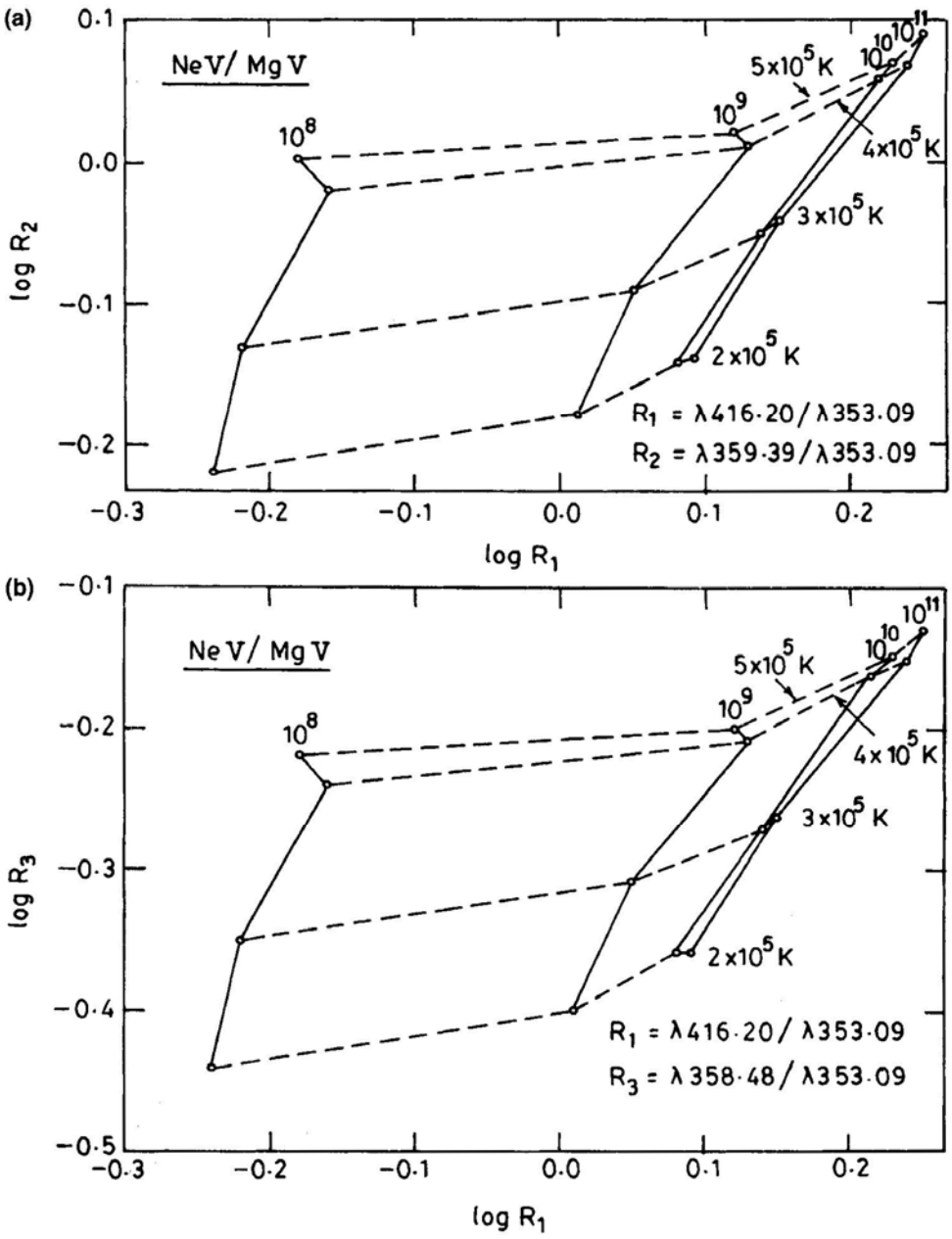

Figure 1. (**a**) Plot of the theoretical NeV/MgV emission line ratio log R_1 = λ416.20/ λ353.09 against log R_2 = λ359.39/λ353.09 for a range of electron temperatures (log T_e = 5.3–5.7; T_e in K) and electron densities (logN_e = 8–11, N_e in cm^{-3}). Points of constant T_e are connected by dashed lines, while those of constant N_e are joined by solid lines; (**b**) Same as Fig. l(a) except for log R_1=λ416.20/λ353.09 against log R_3 =λ 358.48/λ353.09.

(Z being the atomic number) was found to be valid for all the transitions. The respective A_{ji} values for MgV transitions were thus obtained.

The various collision strengths required to solve the steady state equations for the atomic levels have been expressed in terms of effective collision strengths. Collision strengths are in general a function of the incident electron energy. The integral of the collision strength over the incident electron energies gives us the effective collision strength. In simple cases it is possible to get an analytical form for the effective

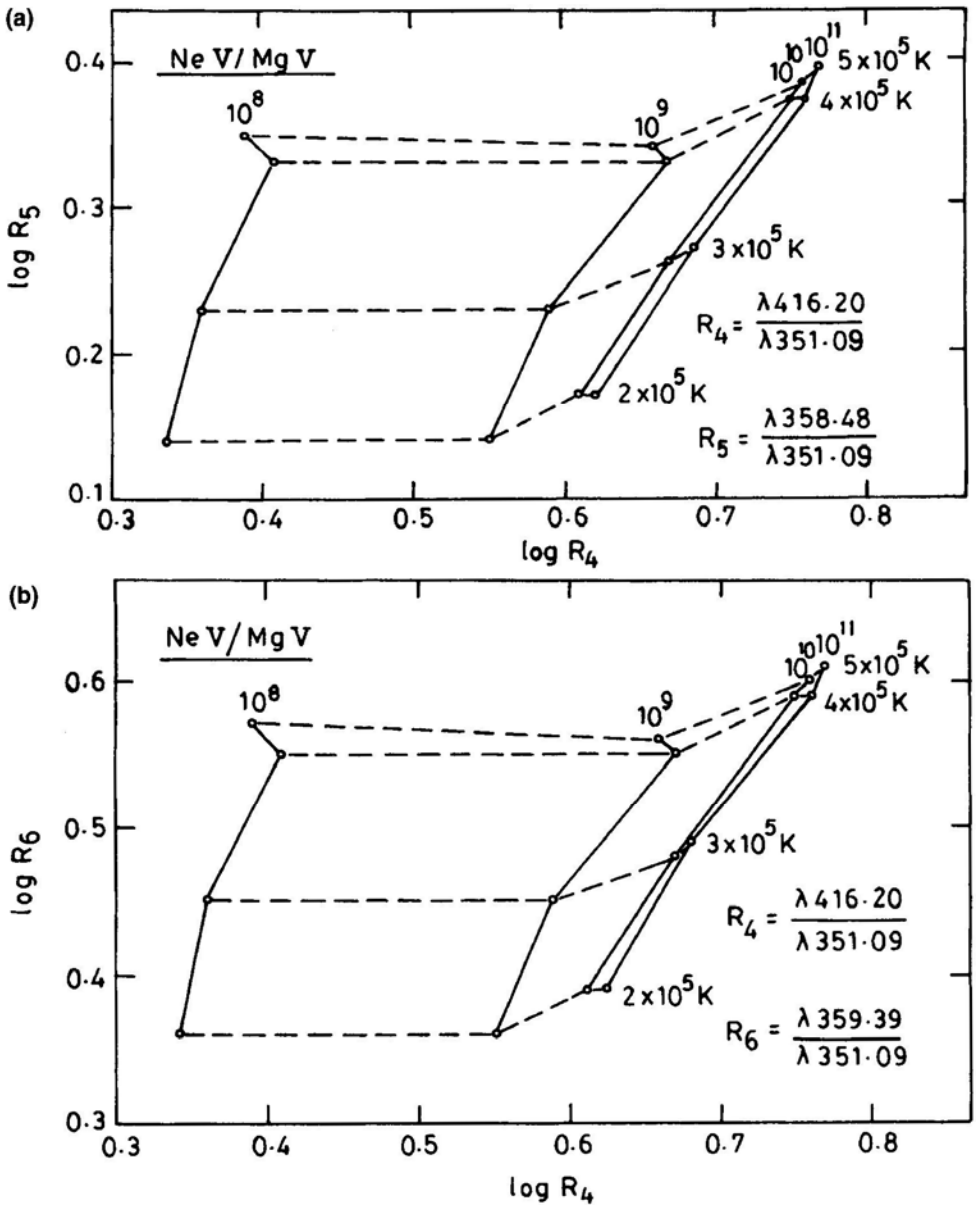

Figure 2 (**a**) same as Fig. 1 except for log R_4 = λ416.20/ 351.09 against log R_5 = λ358.48/λ351.09; (**b**) log R_4 =λ416.20/λ351.09 against log R_6 =λ359.39/λ351.09.

collision strength. In such cases we get the effective collision strengths as a function of the electron temperature for which suitable analytical forms have been derived. Thus we have obtained effective collision strengths for NeV, MgV, SiVII and MgVII ions for the various transitions as a function of electron temperature using the following sources : Aggarwal (1984,1985,1986) for NeV and MgVII ions; and Bhatia, Feldman & Doschek (1979) for MgV and SiVII ions.

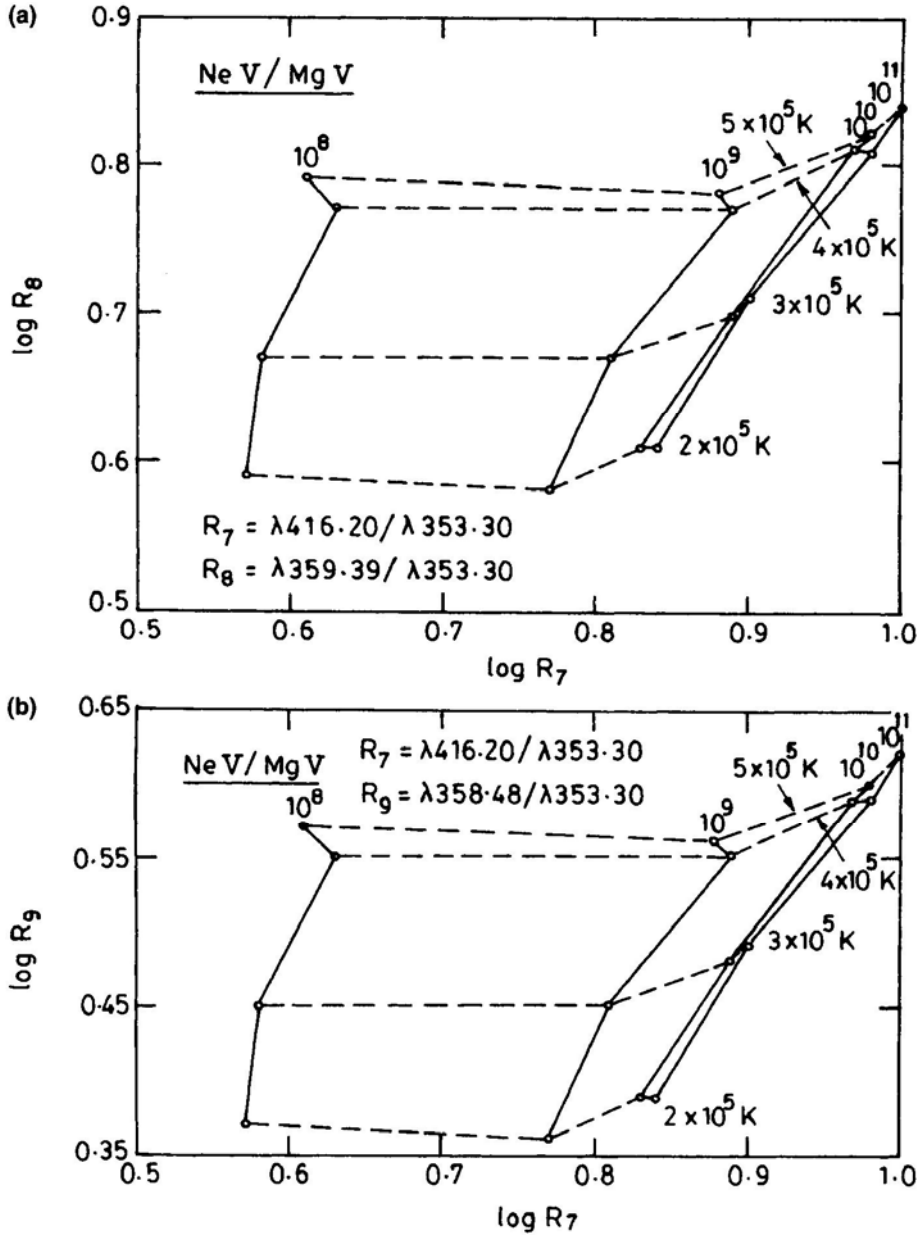

Figure 3. (**a**) Same as Fig. 1 except for log R_7 =λ416.20/λ353.30 against log R_8 = λ359.39/λ353.30; (**b**) logR_7 = λ416.20/λ353.30 against log R_9 = λ358.48/λ353.30.

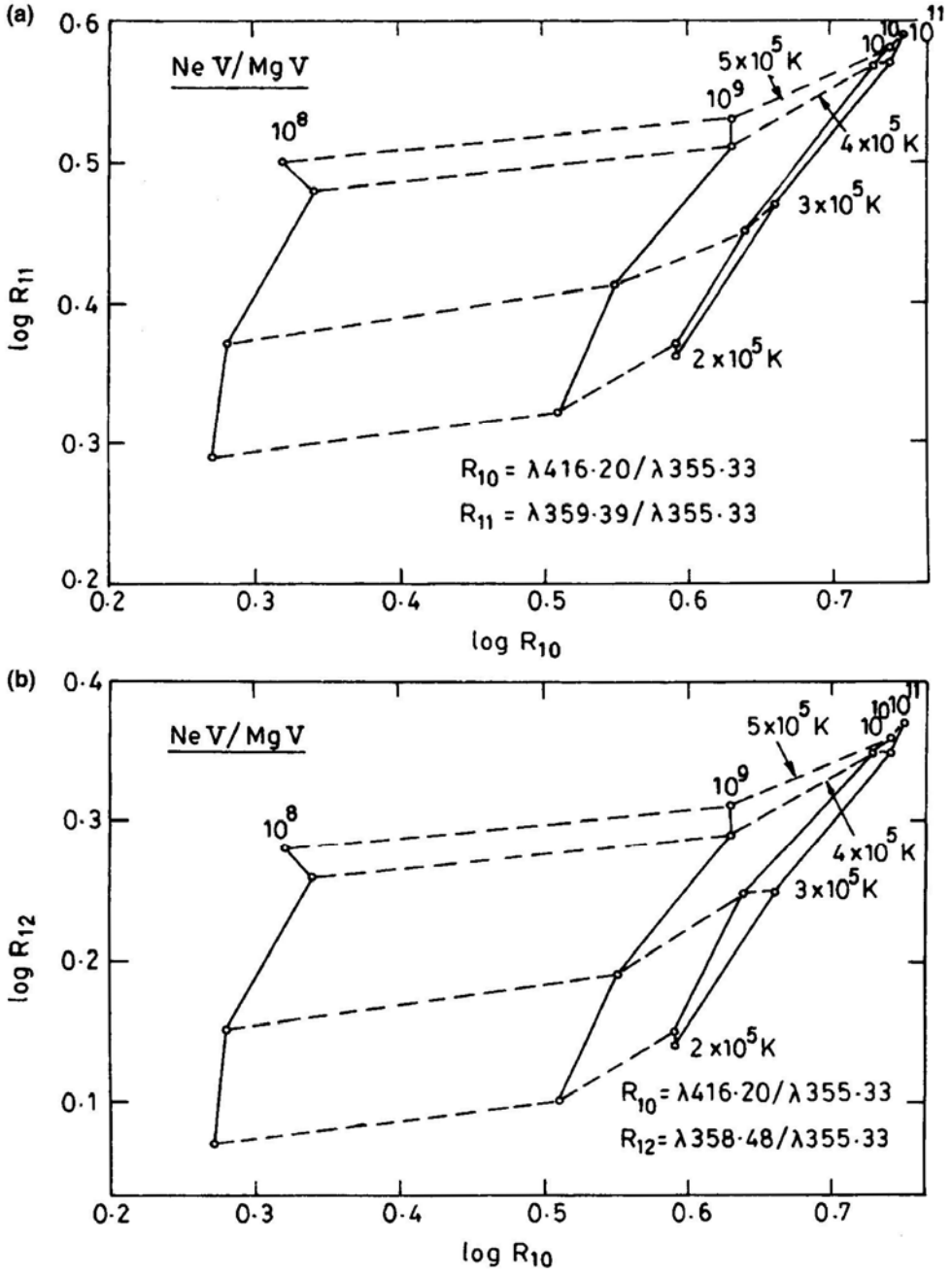

Figure 4 (**a**) Same as Fig 1 except for log R_{10} = λ416.20/λ355.33 against log R_{11} = λ359.39/λ355.33; (**b**) log R_{10} =λ416.20/λ355.33 against log R_{12}= 358.48/ 355.33.

4. Observed and theoretical intensities

Theoretical line intensities have been computed assuming: (**i**) spherically symmetric quiet-Sun model-atmosphere (Elzner 1976), (**ii**) ionic concentrations tabulated by

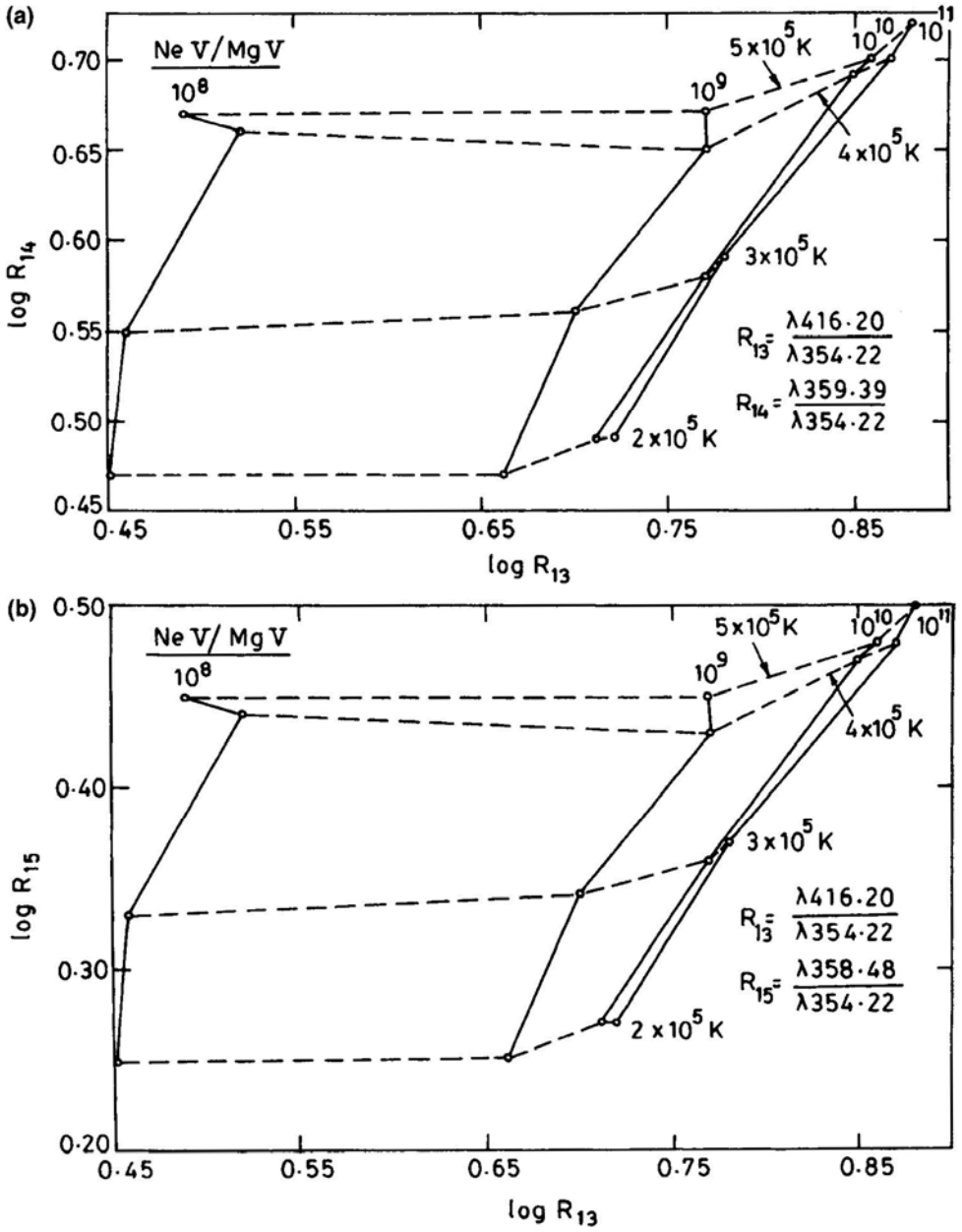

Figure 5. (**a**) Same as fig 1 except for log R_{13} = λ416.20/λ354.22 against log R_{14} = λ359.39/λ354.22; (**b**) log R_{13} =λ416.20/λ354.22 against log R_{15} =λ358.48/λ354.22.

Arnaud & Rothenflug (1985), and (**iii**) values of 3.5×10^{-5}, 3.7×10^{-5} and 3.9×10^{-5} for the element abundance (relative to hydrogen) of Ne, Mg and Si, respectively (Meyer 1985). In Tables 1 to 4 we have listed the theoretical and available observed line intensities for the ions NeV, MgV, SiVII and MgVII. To resolve the discrepancies between the computed and observed intensities, observations at high spectral resolutions with a more sensitive spectral scan are needed. Moreover, the solar atmosphere is completely

Table 5. NeV/MgV electron densities and temperatures (N_e , T_e) derived from ratio-ratio diagrms (cf., Figs. 1 to 5).

Ratio-ratio	N_e (cm^{-3})	T_e (K)
(R_1, R_2)	8.4×10^8	2.4×10^5
(R_1, R_3)	8.5×10^8	2.3×10^5
(R_4, R_5)	8.6×10^8	2.2×10^5
(R_4, R_6)	8.7×10^8	2.2×10^5
(R_7, R_8)	8.6×10^8	2.2×10^5
(R_7, R_9)	8.6×10^8	2.2×10^5
(R_{10}, R_{11})	8.6×10^8	2.4×10^5
(R_{10}, R_{12})	8.6×10^8	2.4×10^5
(R_{13}, R_{14})	8.2×10^8	2.1×10^5
(R_{13}, R_{15})	8.2×10^8	2.1×10^5

Table 6. SiVII/MgVII electron densities and temperatures (N_e, T_e) derived from ratio-ratio diagrams (cf., Figs. 6 to 9).

Ratio-ratio	N_e (cm^{-3})	T_e (K)
(R_1, R_2)	8.1×10^8	6.2×10^5
(R_3, R_2)	8.0×10^8	6.2×10^5
(R_4, R_2)	8.5×10^8	6.2×10^5
(R_5, R_6)	8.3×10^8	6.0×10^5

inhomogeneous and the computed intensity values based on a spherically symmetric model can at best serve as an indicator of the feasibility of observing these lines.

5. Density and temperature diagnostic NeV/MgV line ratios

We show in Figs. 1 to 5 line diagnostics for NeV/MgV ions in the form of ratio-ratio diagrams. The line intensity ratios are dependent, in addition, on the relative element abundances and relative ionic concentrations of NeV and MgV. The ratio-ratio curves in these figures are drawn for equal element abundances of Ne and Mg, which means these line ratio curves are the normalized values given by the expressions

$$\{I(\lambda(\mathrm{NeV}))/I(\lambda(\mathrm{MgV}))\}^{\mathrm{normalized}} = \{I(\lambda(\mathrm{NeV}))/I(\lambda(\mathrm{MgV}))\}^{\mathrm{actual}} \Big/ \left\{ \left(\frac{N(\mathrm{Ne})}{N(\mathrm{H})}\right) \Big/ \left(\frac{N(\mathrm{Mg})}{N(\mathrm{H})}\right) \right\}. \quad (6)$$

Making use of the computed line intensities, we derive N_e and T_e from the ratio-ratio diagrams. These are listed in Tables 5–6. A consistent value of about 8×10^8 cm^{-3} for the density and 2×10^5 K for the temperature are obtained from the NeV/MgV ratio-ratio diagrams (cf., Table 5). It should, however, be noted that abundance anomaly must be taken account of in the analysis and interpretation of observations. These diagrams will be very useful in analysing and interpreting data from the CDS spectrometer when available. Also, we can use these diagrams to deduce N_e and T_e simultaneously from the observed line intensity ratios.

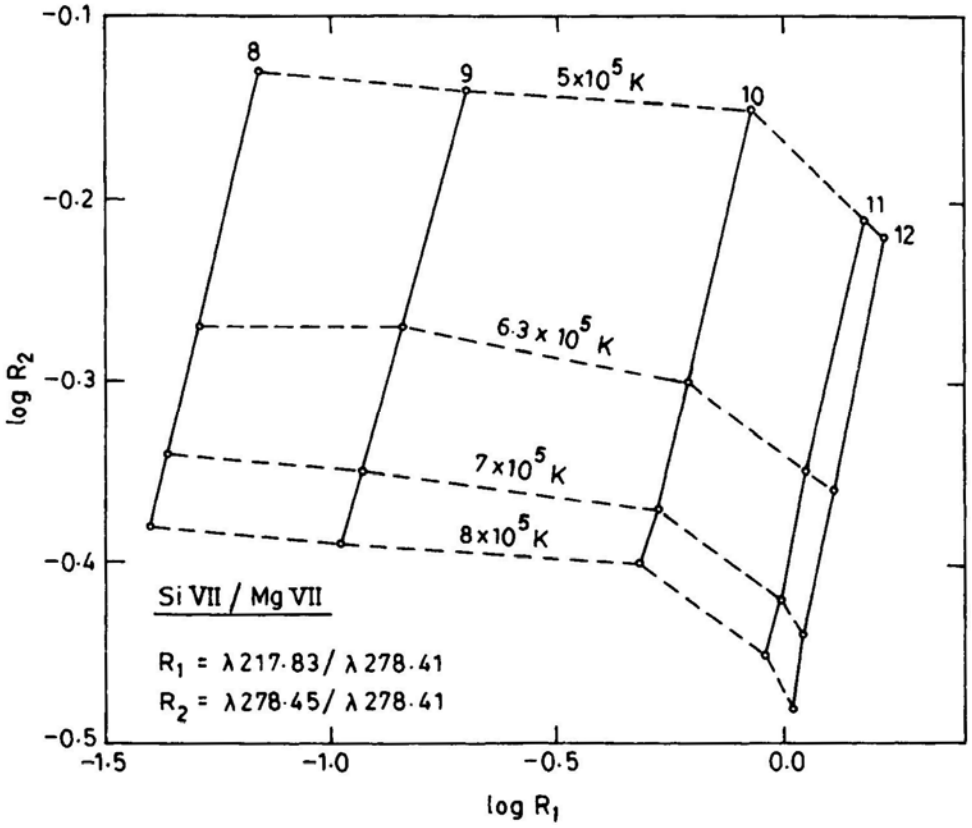

Figure 6. Plot of the theoretical SiVII/MgVII emission line ratio log R_1=λ21.83/ λ278.41 against log R_2 = λ278.45/λ278.41 for a range of logarithmic electron temperatures (log T_e = 5.7–5.9; T_e in K) and logarithmic electron densities (log N_e = 8–11; N_e in cm^{-3}). Points of constant T_e are connected by dashed lines, while those of constant N_e are joined by solid lines.

5.1 *SiVII /MgVII line ratios*

We show in Figs. 6 to 9 line diagnostic ratio-ratio diagrams for SiVII/MgVII The derived values for N_e and T_e are given in Table 6. A consistent value of about 8×10^8 cm^{-3} for density and 6×10^5 K for temperature is obtained. The constant electron pressure of about 5×10^{14}cm^{-3} K is consistent with the transition region solar plasma. The current observations from the CDS/SOHO will be meaningfully used to deduce N_e and T_e from these ratio-ratio diagrams. This will also greatly help in the analysis and interpretation of high-quality EUV data from the CDS spectrometer.

6. Concluding remarks

Simultaneous evaluation of the plasma temperature and the density through the line intensity ratio-ratio diagrams seem to be an excellent diagnostic technique for inhomogeneous solar atmosphere. Although we have used theoretical line intensities based on an unrealistic solar model to deduce N_e and T_e, the ratio-ratio diagrams

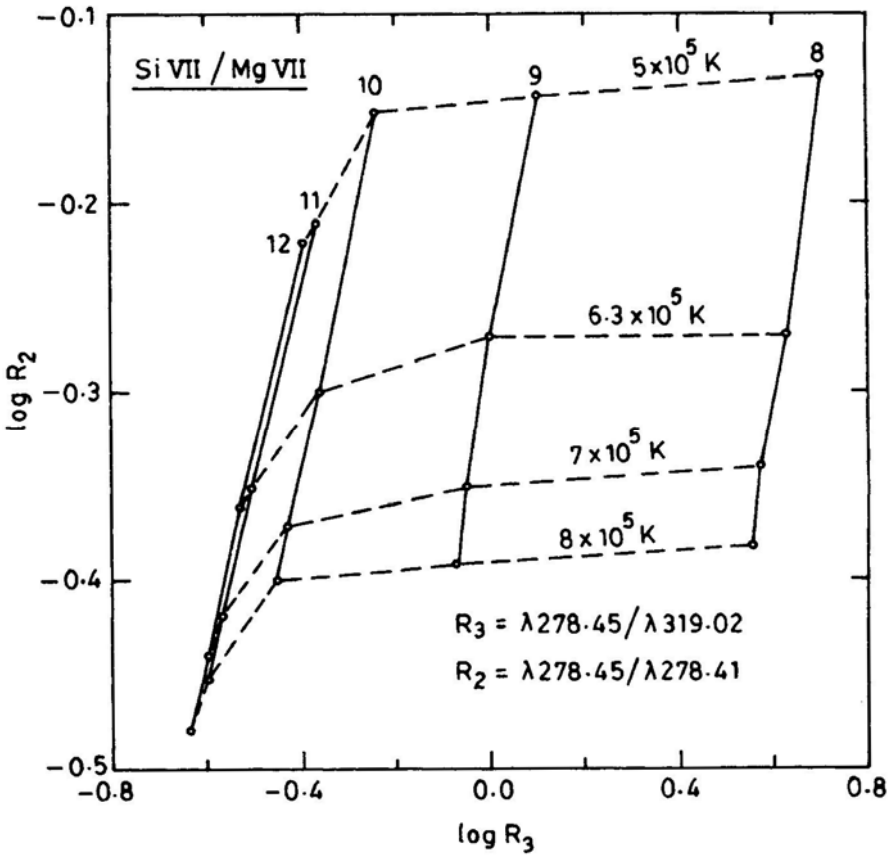

Figure 7. Same as Fig. 6 except for log R_4 =λ278.45/λ319.02 against log R_2 =λ278.45/ λ278.41.

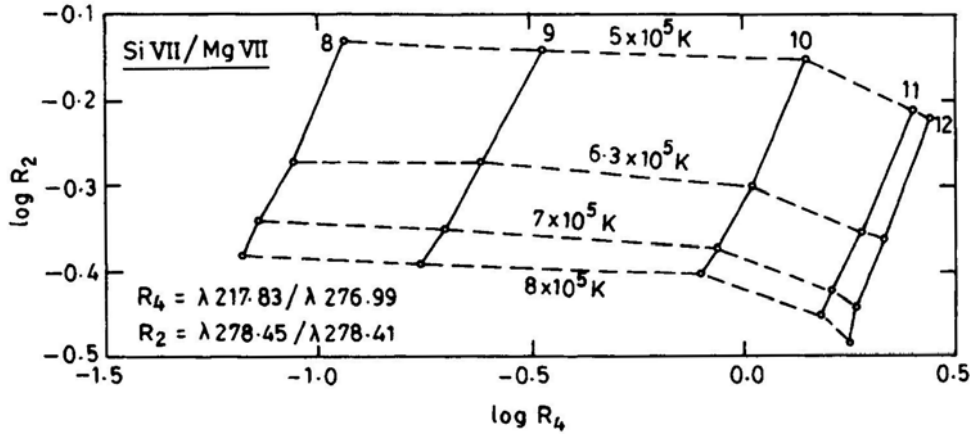

Figure 8. Same as Fig. 6 except for log R_4 = λ217.83/λ276.99 against log R_2 = λ 278.45/ λ278.41.

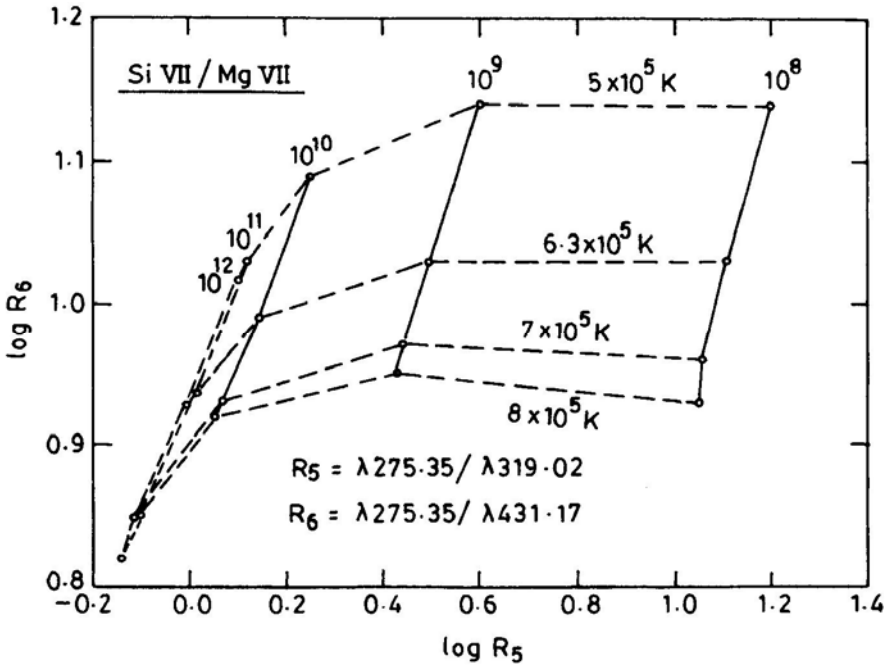

Figure 9. Same as Fig. 6 except for log R_5 =λ275.35/λ319.02 against log R_6 =λ275.35/ λ431.17.

presented in this paper can always be applied, making use of the observed line intensities. Such observations are expected from the high-quality EUV data from CDS/SOHO. While we have shown NeV/MgV and SiVII/MgVII lines to be potentially useful for diagnostics, we expect to estimate Ne/Mg and Si/Mg relative element abundances and their possible variation in different solar structures when such data become available from CDS/SOHO.

Acknowledgements

This work was enabled by the financial support to Anita Mohan from the Department of Science and Technology, New Delhi under the SERC Young Scientist Programme. We wish to thank the Editor for critical reading of the paper.

References

Aggarwal, K. M. 1984, *Astrophys. J. Suppl.*, **56,** 303.
Aggarwal, K. M. 1985, *Astrophys. J. Suppl,* **58,** 289.
Aggarwal, K. M. 1986, *Astrophys. J. Suppl.*, **61,** 699.
Arnaud, M., Rothenflug, R.1985, *Astr. Astrophys. Suppl.*, **60,** 425.
Bhatia, A. K., Feldman, U., Doschek, G.1979, *Astr. Astrophys.*,**80.** 22.
Dwivedi, B. N.1994, *Space Sci. Rev.*, **65,** 289.
Dwivedi, B. N., Mohan, A, Raju, P. K. 1997, *Adv. Space Res.*, **20,** 12, 2271.
Edlén, B. 1982, *Phys. Scripta*, **26**, 71.

Elzner, L. R. 1976, *Astr.Astrophys.*, **47,** 9.
Kelly, R. L., Palumbo, L. J. 1973, Atomic and Ionic Emission Lines below 2000 Å, NRL Report 7599.
Keenan, F. P., Foster, V. J., Reid, R. H. G., Doyle, J. G., Zhang, H. L., Pradhan, A. K. 1995, *Astr. Astrophys.*, **300,** 534.
Malinovsky, M., Heroux, L. 1973, *Astrophys. J.*, **181,** 1009.
Mason, H. E., Monsignori Fossi, B. C. 1994, *Astr. Astrophys. Rev.*, **6,** 123.
Meyer, J. P. 1985, *Astrophys. J. Suppl.*, **57,** 151.
Nussbaumer, H., Rusca, C. 1979, *Astr. Astrophys.* **72,** 129.
Vernazza, J. E., Reeves, E. M. 1978, *Astrophys. J. Suppl*, **37**, 485.

J. Astrophys. Astr. (1999) **20**, 67–77

Cosmological Models with Shear and Rotation

Shwetabh Singh, *Physics Department, Indian Institute of Science, Bangalore 560 012, India.*
Present address: Physics and Astronomy Department, University of Pennsylvania, Philadelphia, PA-19104, U.S.A.

Received 1998 October 13; accepted 1999 August 10

Abstract. Cosmological models involving shear and rotation are considered, first in the general relativistic and then in the Newtonian framework with the aim of investigating singularities in them by using numerical and analytical techniques. The dynamics of these rotating models are studied. It is shown that singularities are unavoidable in such models and that the centrifugal force arising due to rotation can never overcome the gravitational and shearing force over a length of time.

Key words. Cosmological models—singularity—rotation—centrifugal force—shear.

1. Introduction

Cosmological models with rotation have drawn attention in the past for various reasons. Gödel first proposed a rotating universe model, with the aim of demonstrating that general relativity does not incorporate Mach's principle. Another reason for proposing the rotating models has been with the view to avoid a singularity in the universe. This approach was pioneered by Heckmann and Schücking. The basis for this line of thought has been that the centrifugal force arising due to rotation would prevent a collapse of the universe. Finally, since all astrophysical systems have rotation in them so it has been conjectured that there might be rotation in the universe too.

The present work will deal with the investigation of singularities in a special class of metrics namely the Heckmann Schücking metrics in the relativistic and Newtonian framework. Although in general relativity a singularity is unavoidable, as shown by Hawking and Penrose, the study of the dynamics of rotating models does yield an idea of how far rotation may be able to prevent a collapse. The centrifugal force may prevent collapse towards the axis but the collapse may nonetheless take place perpendicular to the plane of rotation. The first part of the work deals with the relativistic Heckmann Schücking model while in the second part the Newtonian analogues of relativistic models will be looked at.

2. The relativistic Heckmann Schücking metric

The most common approach to describe the cosmology of the universe is on the basis of the assumption of homogenity and isotropy of the universe which leads to the

Robertson-Walker metric

$$ds^2 = dt^2 - a^2(t)\,[dr^2/(1-kr^2) + r^2\,(d\theta^2 + \sin\theta^2 d\phi^2)] \tag{1}$$

where $k = 1, 0, -1$.

This line element when coupled with the field equations,

$$R^{\mu\nu} - \tfrac{1}{2}g^{\mu\nu}R = -\kappa T^{\mu\nu} - \Lambda g^{\mu\nu}, \tag{2}$$

can give the value of $a(t)$ where $T^{\mu\nu}$ is given by

$$T^{\mu\nu} = \left(p + \rho + \frac{4}{3}u\right)\frac{dx^\mu}{dt}\frac{dx^\nu}{dt} - \left(p + \frac{u}{3}\right)g^{\mu\nu}. \tag{3}$$

where ρ and u are the matter and radiation densities and ρ is the pressure.

This form of the solution gives rise to a universe with a singular origin. Can the singularity be averted if one considers a modification in the assumptions governing the choice of the line element? The postulate regarding isotropy of the universe is replaced by anisotropy which may be due to rotation and shear. The world lines of the galaxies being still geodesics along which the world time is measured by the t = constant surfaces are no longer orthogonal to the geodesics. This gives rise to cross product terms of the type $g_{\mu 4}dx^\mu\,dt$ ($\mu = 1, 2, 3$) in the metric.

One such type of metric was proposed by Gödel (1949),

$$ds^2 = dt^2 + 2e^{x_1}dtdx^2 - (dx^1)^2 + \tfrac{1}{2}e^{2x_1}(dx^2)^2 - (dx^3)^2. \tag{4}$$

The solution of the field equations with this metric indicates a rotating and stationary universe with the angular velocity of the various components given by

$$\omega^\mu = \frac{\epsilon^{\mu\nu\lambda\kappa}}{6\sqrt{g}}a_{\nu\lambda\kappa}. \tag{5}$$

The non-vanishing of ω^3 shows a rotation about the x^3 = constant axis.

The failure of this model to account for the red shift of galaxies because of its stationary character prompted Heckmann & Schücking (1958) to give a generalisation of the Gsödel's metric to obtain a universe with a non-singular origin and also to account for the red shift by making this model non-static. They hoped to find finite oscillating universes with maximum and minimum radii.

One such model having Gödel's model as a special case, has the line element

$$\begin{aligned} ds^2 = {} & dt^2 + 2e^{x^1}dtdx^2 - c_{11}(t)(dx^1)^2 - 2c_{12}(t)e^{x^1}dx^1dx^2 \\ & + \alpha c_{11}(t)e^{2x_1}(dx^2)^2 - S^2(t)(dx^3)^2. \end{aligned} \tag{6}$$

which when solved with the field equations gives the following set of differential equations:

$$\frac{1}{4R^2}[(\dot{c}_{12}+1)^2 + \alpha\dot{c}_{11}^2 - 4\alpha c_{11}] + \frac{\dot{S}}{S}\left(\frac{2c_{12}+\dot{c}_{11}}{2R^2} - \frac{\dot{R}}{R}\right) = -\Lambda - \frac{\alpha^2}{RS}, \tag{7}$$

$$\dot{c}_{12}c_{11} - \dot{c}_{11}c_{12} - c_{11} = -\frac{\alpha}{RS}, \tag{8}$$

$$\frac{\ddot{R}}{R}+\frac{\ddot{S}}{S}-\frac{1-\dot{c}_{12}^2-\alpha\dot{c}_{11}^2}{2R^2}=\Lambda-\frac{\alpha^2}{2RS}, \tag{9}$$

where

$$R^2=c_{11}-\alpha c_{11}^2-c_{12}^2, \tag{10}$$

$$\rho=\frac{\alpha^2}{RS}, \tag{11}$$

$$\omega=\frac{1}{\sqrt{2}R}. \tag{12}$$

Heckmann and Schücking did not solve these equations which are too complicated for an analytical solution. Narlikar (1960) showed, however, that these models would not fulfill their original purpose. Later the interest in these models waned as the singularity theorems gained wide acceptance. Nevertheless we return to the Heckmann Schücking model to investigate its dynamical behaviour by numerical methods, to see how the singularity actually develops.

The above equations were thus solved numerically with the initial values taken in correspondence with the Gödel (1949) metric except for the introduction of a $\dot{c}_{11}$ term to introduce a non-static character in the metric. Thus,

$$c_{11}(0)=1, \tag{13}$$

$$c_{12}(0)=0, \tag{14}$$

$$\Lambda=-\tfrac{1}{2}, \tag{15}$$

and different values of $\dot{c}_{11}$ and α were taken to highlight different cases. The parameters were varied as follows:

$$\dot{c}_{11}=0.1, 0.01, 0.001, \tag{16}$$

$$\alpha=\tfrac{1}{2}, -\tfrac{1}{2} \tag{17}$$

in correspondence with the two separate cases pointed out by Narlikar (1960).

The results for $\alpha=\frac{1}{2}$ as shown in Fig. 1 show a collapse in R and Fig. 2 reveals the small oscillations R makes before the collapse. Fig. 3 show a continuously shrinking x^3 dimension. The coefficient of $(dx^3)^2$ in the metric goes to zero linearly showing a one-dimensional singularity. R, which shows the extent of the universe in the other two dimensions also goes to zero (Fig. 1) pointing to a singularity in the volume after a period of oscillations since $V \propto R^2S$. The equations also imply a rapidly increasing angular velocity by consequence of $\omega = 1/\sqrt{2}\ R$ as 0pointed out by Gödel (1950) in one of his discussions of the properties of his metric.

The oscillations are sustained for a longer period of time if the initial rate of change of c_{11} with respect to time is large as this delays R from falling to zero as a consequence of

$$R^2=c_{11}-\alpha c_{11}^2-c_{12}^2, \tag{18}$$

becoming smaller and smaller more slowly.

It can also be seen from Figs. 4 and 5 that for the case $\alpha=-\frac{1}{2}$ oscillations are not possible in the Gödel like universe and for the same other parameters the universe

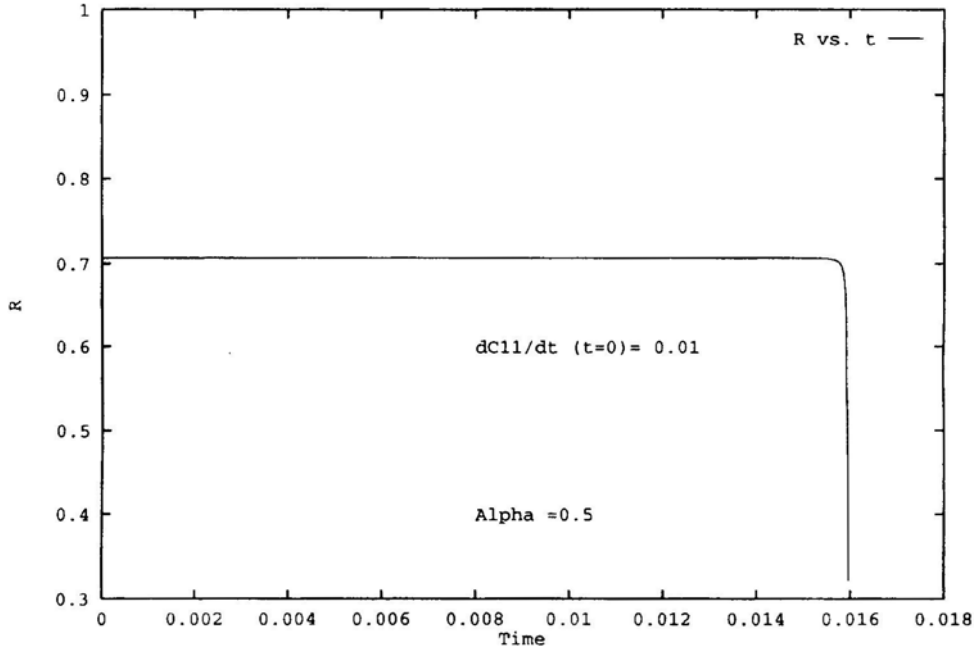

Figure 1. Variation of R with time with $\alpha = 0.5$. The figure shows the behaviour of R with time for a positive α and reveals a behaviour culminating in a collapse.

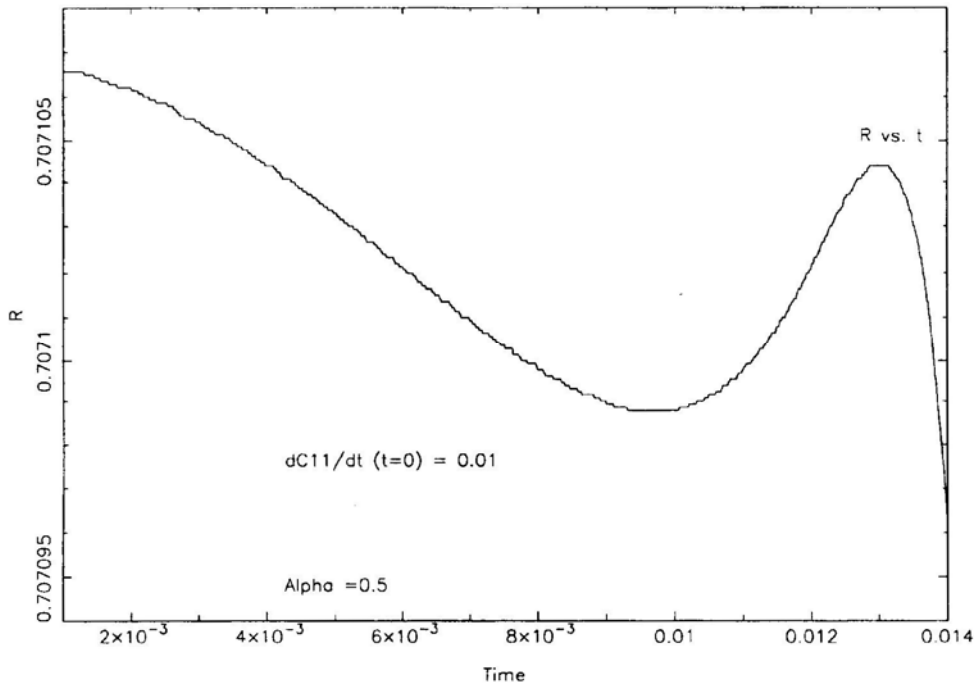

Figure 2. Variation of R with time with $\alpha = 0.5$. The figure shows that during the initial time the behaviour of R for a positive α reveals an oscillating pattern.

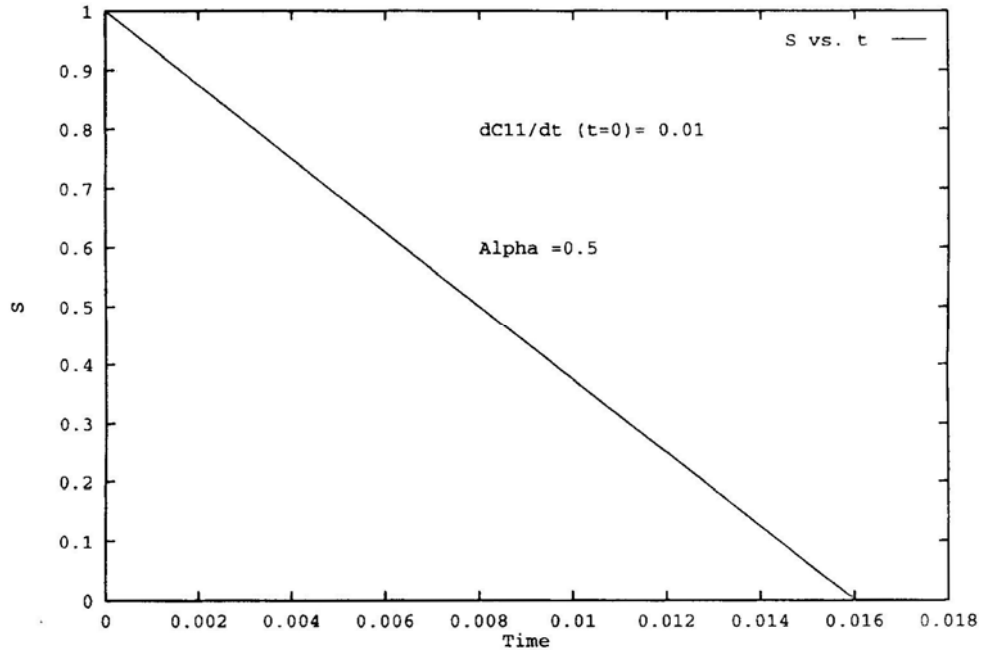

Figure 3. Variation of S with time with $\alpha = 0.5$. The figure shows the behaviour of the coefficient of the x^3 dimension, for a positive α, which is seen to go to zero linearly with time.

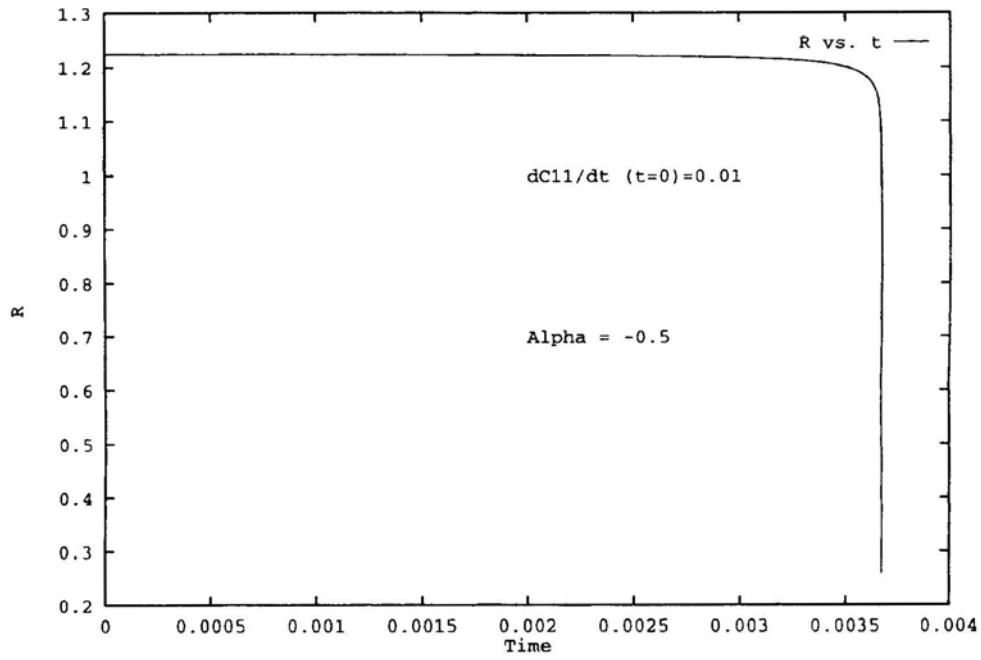

Figure 4. Variation of R with time with $\alpha = 0.5$. The figure shows that for a negative α, R goes to zero monotonically without executing any oscillations about an order of magnitude faster than the case with a positive α.

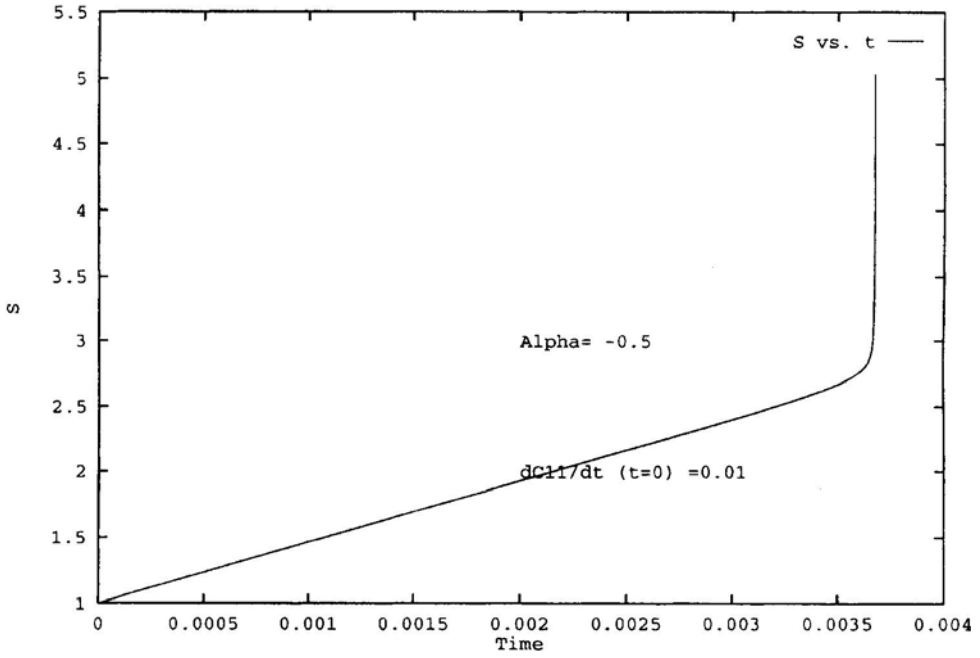

Figure 5. Variation of S with time with $\alpha = -0.5$. The figure shows that for a negative α, the coefficient of the x^3 dimension increases with time first linearly and then very rapidly.

shrinks to a two-dimensional singularity with a rapidly increasing S (Fig. 5), the coefficient of $(dx^3)^2$ and a decreasing R (Fig. 4).

The volume of the universe given by $V \propto R^2 S$ also goes to zero suggesting that a singularity results in this case too. The results for $\alpha = -\frac{1}{2}$ are in agreement with the analytical argument of Narlikar (1960) where he has shown that for $\alpha = -\frac{1}{2}$ the universe does not oscillate between finite limits by considering equations (7), (8), and (9) coupled with the signature conditions. This form of a universe leads to a situation similar to the isotropic case.

Another result which may be directly seen from the numerical simulation is that the universe for $\alpha = -\frac{1}{2}$ is inherently unstable and collapses about an order of magnitude faster than for $\alpha = \frac{1}{2}$, all other parameters being comparable.

3. Newtonian cosmologies with rotation

One of the first works done on developing cosmology on Newtonian terms was by McCrea & Milne (1934). Their work was based on the concept of isotropy and homogenity in the universe and they showed that models analogous to the Friedmann Robertson Walker model could be obtained by using the Newtonian theory too. Heckmann & Schücking (1955) formulated the Newtonian equations of cosmology by solving the following three equations used to describe the general behaviour of a homogeneous universe, but doing away with the assumption of isotropy:

$$\frac{\dot{\rho}}{\rho} + \mathrm{div} \cdot \mathbf{v} = 0. \tag{19}$$

$$\dot{\mathbf{v}} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\phi, \tag{20}$$

$$\nabla^2\phi + \lambda = 4\pi\rho G. \tag{21}$$

The Newtonian equations of cosmology formulated by them include both shear and rotation terms. Thus while rotation prevents the universe from collapsing, shear has the opposite effect. This is analogous to the general relativistic result derived by Raychaudhuri (1955). They tried to avoid singularity by setting the shear equal to zero which is possible in the Newtonian framework. They were able to show that it is possible to avoid singularities because the rotation term dominates as $R \to 0$, hence preventing a collapse.

This freedom to take the shear equal to zero in Newtonian models does not exist in general relativistic models as was shown by Ellis (1967). Hence, it is not possible to avoid singularities in the general relativistic models which give a more accurate description than their Newtonian analogues, in accordance with the Hawking-Penrose theorem (see Hawking & Ellis 1975).

However in 1963, Narlikar formulated the gravitational force approach which was based on the inverse square law rather than the Poisson equation,

$$\nabla^2 \phi = 4\pi G\rho. \tag{22}$$

This puts a greater restriction on ϕ and implies that even though it might be possible to put the shear equal to zero initially, the time dependence of the shear has the effect of it acquiring some finite value in the course of time. Thus it is not obvious that singularity can be prevented.

By the axioms of Newtonian mechanics we have a Euclidean space with rectangular co-ordinates $x_\mu (\mu = 1, 2, 3)$ and an even-flowing uniform time, t. Thus at any time t the universe will present the same large scale view to all observers whose motion is idealized as the streaming of an ideal fluid. The velocity-distance relation can be written in the form,

$$\upsilon_\mu = H_{\mu\nu} x_\nu, \tag{23}$$

where $H_{\mu\upsilon}$ is a function of t only, which appears in analogy with the Hubble's constant in the Robertson-Walker cosmology, but is direction dependent in accordance with the assumption of anisotropy. After writing the solution of (23) in the form,

$$x_\mu = a_{\mu\nu}(t) x_\nu^0 \Delta = det\|a_{\mu\nu}\| \tag{24}$$

and solving equations (19) and (20), Narlikar (1963) arrived at the following equation of motion,

$$\Delta \frac{\ddot{a}_{\mu\nu}}{2} = -\frac{4\pi G\rho_0 a_{\mu\nu}}{3}. \tag{25}$$

After replacing the time t by a dimensionless co-ordinate,

$$\tau = \left(\frac{4\pi G\rho_0}{3}\right)^{\frac{1}{2}} t \tag{26}$$

and continuously differentiating equation (25), Narlikar (1963) was able to reduce the six equations of motion to a single fourth order equation:

$$\Delta^2\Delta'''' + 7\Delta\Delta'' = 4\Delta'^2 + 9\Delta = 0. \tag{27}$$

This fourth order nonlinear differential equation may be reduced by the following transformations,

$$\Delta = F^2, \left(\frac{\mathrm{d}F}{\mathrm{d}\tau}\right)^2 = X(F), \quad F = e^U, \quad \frac{\mathrm{d}X}{\mathrm{d}U} = Y(X), \tag{28}$$

$$X = F'^2, Y = 2FF'', \tag{29}$$

to the following second order equation,

$$XY^2\left(\frac{\mathrm{d}^2Y}{\mathrm{d}X^2}\right) + XY\left(\frac{\mathrm{d}Y}{\mathrm{d}X}\right)^2 + \left(X + \frac{Y}{2}\right)Y\frac{\mathrm{d}Y}{\mathrm{d}X} + Y^2 - 2XY + 7Y - 2X + 9 = 0. \tag{30}$$

The above equation was solved numerically by Narlikar (1963) for various initial conditions and all of them did show a singularity. Davidson & Evans (1973) investigated equation (25) further and all their numerical and analytical results also show a singularity.

But, since numerical analysis cannot be exhaustive of all initial conditions, equation (30) was analysed asymptotically to arrive at a more general result. We take the form of the asymptote as,

$$Y = mX + K. \tag{31}$$

Substituting Y in equation (30) and equating the coefficients of X^2 and X to zero we find the following asymptotes to the curves

$$Y = K_1, \tag{32}$$
$$Y = -2X + K_2 \tag{33}$$
$$Y = (34)\,\tfrac{2}{3}\,X - 3. \tag{34}$$

Here K_1 and K_2 are constants and equation (34) is an exact solution and was also found by Narlikar (1963). The curves with asymptotes $Y = -2X + K_1$ were also investigated by Narlikar (1963), and were found to go asymptotically as $Y = -2X$ as is shown in Fig. 6. However the curves with asymptotes $Y = K_1$ were missed in that work.

Substituting the value of Y from equation (29) in (32), we get,

$$FF'' = k. \tag{35}$$

Solving the above equation by integrating leads to,

$$\frac{\mathrm{d}F}{\sqrt{\ln F}} = c_1\mathrm{d}\tau. \tag{36}$$

Substituting x^2 = In F, we can get the equation in the form of the integral of the error function,

$$\int e^{x^2}\mathrm{d}x = c_2\tau, \tag{37}$$

where c_2 is some constant independent of Δ and τ.

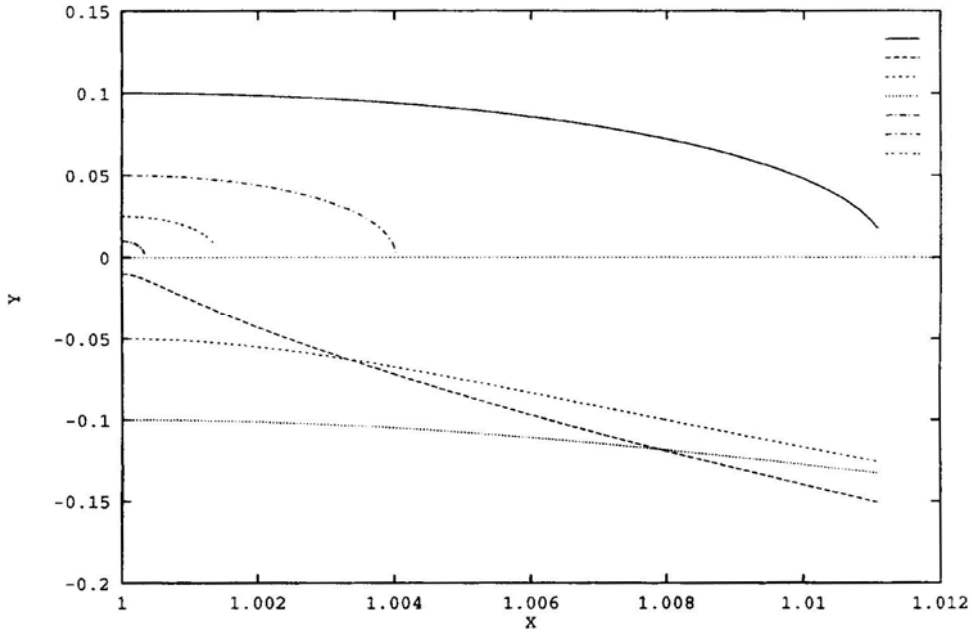

Figure 6. Variation of X with Y for some different initial conditions. The curves show that the curves for $Y > 0$ can never be joined to the curves for $Y < 0$ and that Y eventually decreases with X corresponding to universes where rotation has been ineffective in preventing a singular state. All curves in the figure asymptotically go as $Y = -2X$.

It may be seen from equation (37) that τ has a range of values from $-\infty$ to ∞ being the time co-ordinate. By readjusting the time scale we can always arrange that at some point in the range of τ, Δ must become equal to zero which will imply a singularity.

The two other asymptotes of the curve have already been investigated numerically by Narlikar (1963) and Davidson & Evans (1973) and were shown to be singular. The asymptote Y = constant does indeed correspond to a few numerical solutions which were compiled by Davidson & Evans (1973).

Another way of looking at the singular nature of the solution is by investigating the case when asymptotically $Y = -2X$. Then we can take,

$$Y = -2X + c = -2XZ(X) \tag{38}$$

where the function $Z(X) \to 1$ as $X \to \infty$. thus we get,

$$Z(X) = 1 - \frac{c}{2X} \tag{39}$$

where c is a constant independent of X.

Substituting these values of Y and its first and second derivatives in equation (30) we see that the quadratic terms of (30) exactly cancel to zero. We are then left with $7Y - 2X + 9$. For this to tend to zero asymptotically we arrive at the condition,

$$-16X + 7c + 9 = 0. \tag{40}$$

Substituting the value of X from (29) and putting it in terms of Δ from equation

(28) we have,

$$\frac{4}{\Delta}\frac{\mathrm{d}\Delta}{\mathrm{d}\tau} + 7c + 9 = 0. \tag{41}$$

Thus,

$$\frac{\Delta_2}{\Delta_1} = \exp\left[\left(\frac{-9-7c}{4}\right)(\tau_2 - \tau_1)\right]. \tag{42}$$

It is clearly seen from equation (42) that for $c > -9/7$ $\Delta_2 \to 0$ for $\tau_2 \gg \tau_1$ and for $c < -9/7$, $\Delta_2 \to 0$ for $\tau_2 \ll \tau_1$.

Since the time scale can be arbitrarily adjusted, we will get a singularity in one of the cases mentioned above.

The conclusion arising from the fact that the volume element of the universe, Δ goes to zero in all possible cases taken into consideration implies a singular solution to the generic Newtonian cosmological model with shear and rotation. The results show that in an expanding universe the outward force arising due to rotation can never be enough to overcome the combined force due to shear and gravitational pull. It may also be seen that the shear term cannot be taken equal to zero over the entire time scale if the expanding model of the universe is considered.

4. Conclusion

Though singularities are unavoidable as shown by Hawking & Ellis (1975) in the general singularity theorem, nonetheless it is interesting to observe in these models how we arrive at singularities in the course of time. This result holds equally well for the Newtonian models with shear and rotation as they have been shown to be singular in the history or future of the universe. The centrifugal force arising due to the rotation of the universe can never adequately combat the gravitational and shearing force.

Acknowledgements

I would like to thank Professor, J. V. Narlikar, under whose guidance the work was undertaken, for discussions and constant support. I also thank all the members of the Inter University Centre for Astronomy and Astrophysics, Pune for their help and the Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore for providing me the opportunity to undertake this work under the Summer Research Fellowship Programme, 1998.

References

Davidson, W., Evans, A. B. 1973, Newtonian Universes Expanding or Contracting with Shear and Rotation; *International Journal for Theoretical Physics*, **7**, 5, 353–378.

Ellis, G. F. R. 1967, Dynamics of Pressure-Free Matter in General Relativity, *J. Mathematical Physics* **8**, 1171.

Gödel, K. 1949, *Review of Modern Physics*, **21**, 447.

Gödel, K. 1950, Rotating universes in General Relativity Theory in Proceedings 1950 *International Congress of Math.* **I,** 175–181.

Hawking, S., Ellis, G. I. F. 1975, Large Scale Structure of Space-Time, Cambridge University Press.

Heckmann, O., Schücking E.1995, *Zeitschrift für Astrophysik*, **38** 95.

Heckmann, O., Schücking E.1998, Solvay Conferences (Brussels).

McCrea, W. H., Milne, E. A. 1934, *Quarterly Journal of Mathematics*, **5**, 73.

Narlikar, J.V.1960, Estratto da Rendiconti della Seuola Internazionale di Fisica - XX Corso, 222–227, 2

Narlikar, J. V. 1963, Newtonian Universes with Shear and Rotation, *Mon. Not. R. Astr. Soc.* **126,** 203-208.

Raychaudhuri, A. K. 1955, *Phys. Rev.* **98**, 1123.

J. Astrophys. Astr. (1999) **20**, 79–88

Formation of Giant Molecular Clouds by Aggregation in a Galactic Disc with Rigid Rotation

Tong-jie Zhang & Guo-xuan Song, *Shanghai Observatory, Chinese Academy of Sciences, Shanghai, 200 030, China* and
National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100 012, China
email: tjzhang @ center. shao. ac. cn

Received 1995 September 20; accepted 1999 June 10

Keywords. Aggregation of clouds—rigid rotation.

1. Introduction

Giant Molecular Clouds (GMCs) in disc galaxies are the main regions within which star formation takes place. Due to star formation, subsequent stellar wind, expanding HII regions surrounding the stars, and finally supernova explosions, the parent GMCs get disrupted. Thus with the passage of time the number of giant clouds will decrease. This in turn will result in a decrease in the star formation rate unless, of course, the more massive clouds continue to be formed from presumably the less massive clouds. This important question is the main motivation for this investigation.

In general there are two mechanisms for the formation of the GMCs: (**i**) Coagulation process in which the colliding clouds coalesce to form more massive clouds (Kwan & Valdes 1987), and (**ii**) Aggregation process in which the colliding clouds do not coalesce, but form a clump with several distinct clouds—inelastic collisions dissipate the relative kinetic energy of the original clouds (Song 1991). Such a clumpy aggregation of smaller clouds could also serve as a model for the GMCs. High resolution observations in the lines of Carbon Monoxide support such a scenario (Rivola *et al.,* 1986). Kahn & Song (1994) studied the efficiency of such a mechanism and concluded that it is fairly efficient if the coefficient of restitution is less than or equal to 0.5, or in other words the collisions are sufficiently inelastic.

In reality the formation of the GMCs by such a process of aggregation may depend on other factors as well. For example, Song (1996a) concluded that gravitation between the clouds may be as important as the inelasticity of the collisions.

In this communication we examine the role of the rotation of the disc in the formation of the GMCs. In general disc galaxies exhibit differential rotation. The importance of the corotation resonance, as well as that of the inner and outer Lindblad resonances is well appreciated. In order to highlight the possible role of rotation, we exclude for the time being the perturbations which lead to stable spiral patterns in disc galaxies.

2. The model

The model adopted in this paper is basically the same as the one used by Kahn and Song (1994), except that the Toomre disc is replaced by another axisymmetric disc in which the rigid rotation is established. The individual clouds are randomly distributed in a ring between 3 kpc and 7 kpc, with an initial half width of 100 pc normal to the disc. All individual clouds are assumed to have a mass of $10^4\ M_\odot$, with a radius of 5.85 pc. The total number of clouds is 120,000, corresponding to a total mass of molecular hydrogen in the ring of $1.2 \times 10^9 M_\odot$. Due to the magnetic field of the Galaxy, the *effective radius* of the clouds for the collision process is 11.7 pc (Clifford & Elmegreen 1983).

Khan and song studied to process of aggregation of the GMCs with a Toomer disc with a frequency of rotation $\Omega(r) = B(1 + r^2/a^2)^{-3/4}$, with $B = 0.0576$ Myr^{-1} and the scale length $a = 7$ kpc. For comparison here we study a model with rigid rotation. In this model $\Omega(r) = \text{constant} = 0.05$ Myr^{-1}.

At the beginning of the run every cloud of mass $10^4 M_\odot$ has a circular motion corresponding to the rigid rotation *plus* a velocity with dispersion equal to 10 kms^{-1}, 7.5 kms^{-1} and 5 kms^{-1} in the radial, circumferential, and normal to the disc, respectively. The collision between the clouds was taken to be inelastic (with a coefficient of restitution equal to 0.1) in the direction of the line joining the centres of the two clouds; it was treated as elastic in the perpendicular direction. Gravitation between the clouds was included, but with a cut-off range of 235 pc. The *clumps* were identified by using a percolation scheme with the percolation parameter of 30 pc, i.e., when the separation between the centres of the two clouds is less than 30 pc they were considered to be part of a clump. In order to elucidate the role of rotation in the formation of the GMCs, the disruption of the clouds subsequent to star formation was ignored. The model was run for a total of 300 Myr.

3. Results

In Fig. 1 we have plotted the fraction of molecular matter in clumps $F(M)$ with a mass greater than M; this has been done both for a model with differential rotation, as well as rigid rotation. As we can see from the figure, the efficiency of the formation of aggregates or clumps is greater in the model with rigid rotation if one confined oneself to times less than 150 Myr. The reduced efficiency at early times of the model with differential rotation may be understood along the lines pointed out by Song (1996b). Differential rotation plays both positive, as well as negative roles in the formation of aggregates. Whereas it increases the probability of two clouds colliding to eventually form a clump, it also acts to tear apart already formed clumps. In the model with rigid rotation the clumps are formed from nearby clouds—the probability of collisions leading to aggregation is determined by the *velocity dispersion of the clouds.* Once clumps form out of clouds in proximity they are unlikely to grow further. This is the reason one finds that $F(M)$ remains constant after approximately 150 Myr. In contrast, in the second model, differential rotation also helps in the formation of aggregates. Because of the differential rotation there is a possibility of collision between two distinct clumps formed at different galactocentric distances. This will result in a steady increase in the fraction of mass contained in

clumps (in the model with rigid rotation the collision between two clumps seldom occurs).

To understand the details better, we now refer to Fig. 2. Here a clump 3099 is identified at an age of 150Myr. This consists of 171 individual molecular clouds. We now trace backwards in time; what has been shown in the figure is a sequence of steps in the evolution at an interval of 5 Myr. The final clump 3099 at 150 Myr is indicated by a short arrow. It is interesting to note that the general shape of the distribution of the individual clouds is filamentary which gets progressively thinner. Obviously the

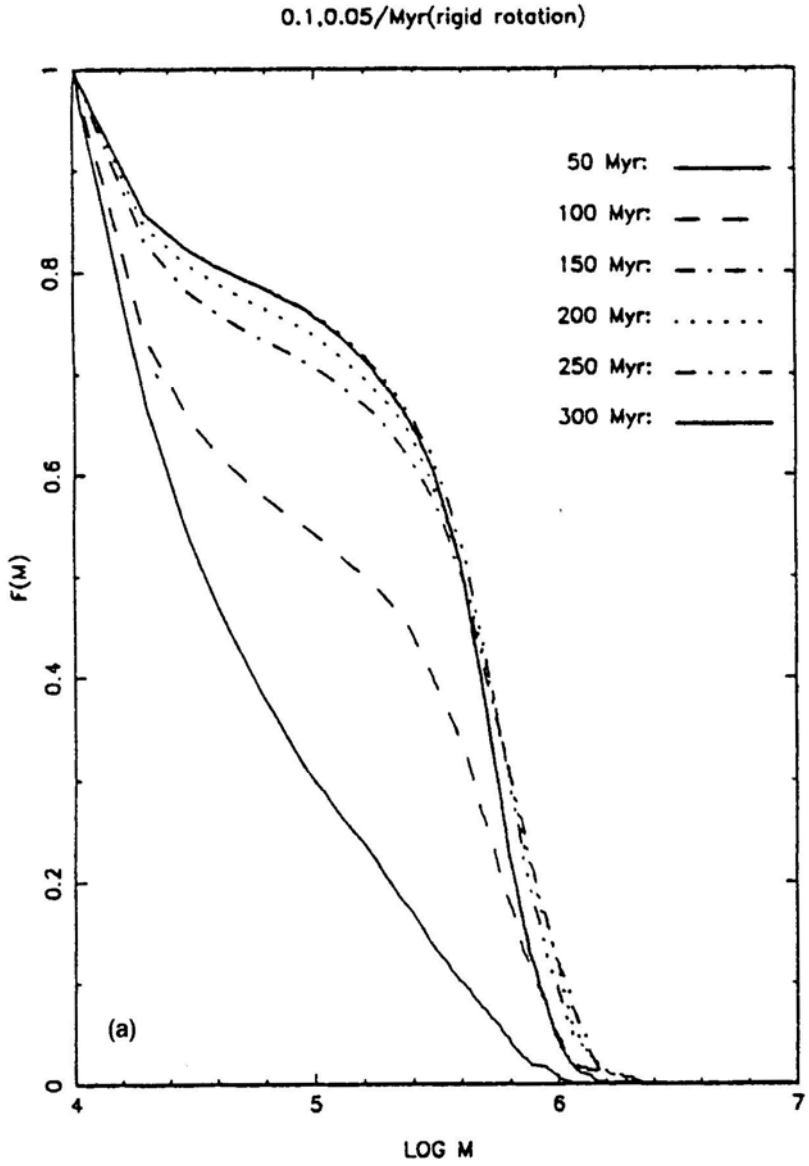

Figure l(a). (*Continued*)

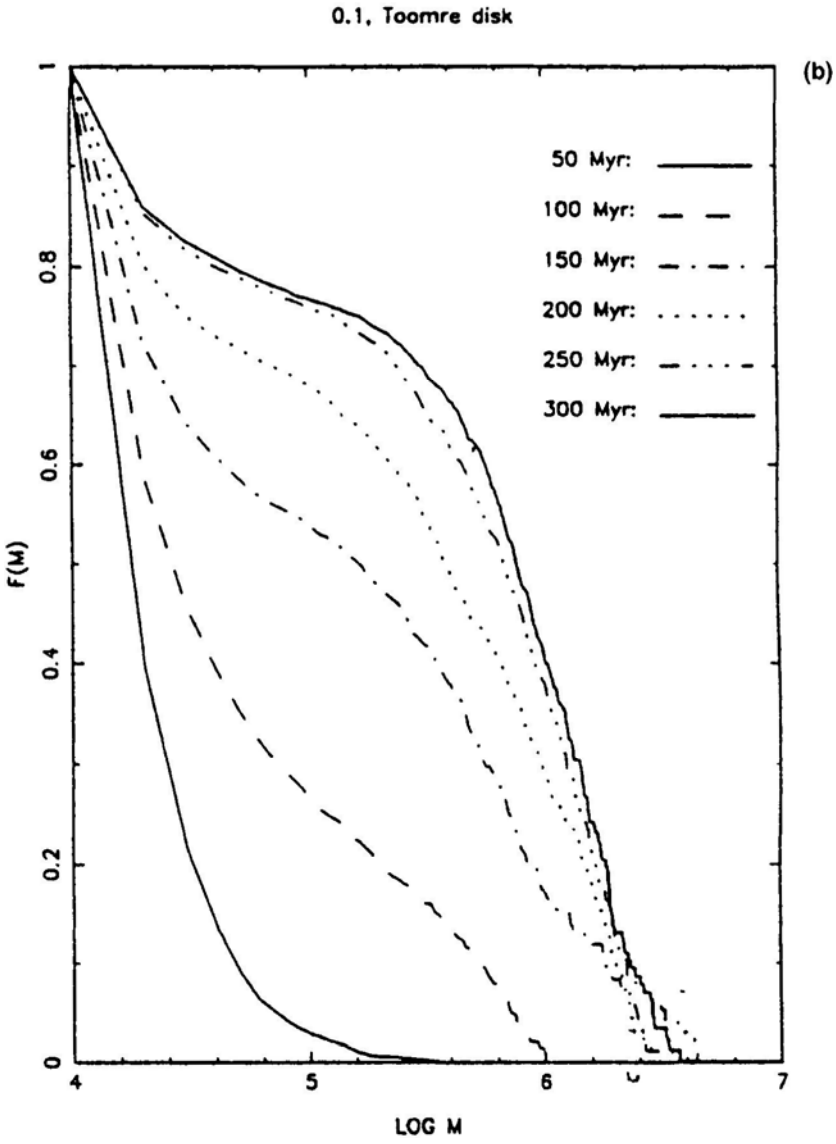

Figure l(a&b). The fraction $F(M)$ of molecular matter in clumps with mass greater than M. The left panel is for the model with rigid rotation in the disc and the right panel is for the model with differential rotation.

orientation of these filaments will be different if one took into account the effects of differential rotation.

In Fig. 3 the evolution of the clump 1100 is shown as time progresses from 50 Myr to 150 Myr. This plot has a higher resolution than in Fig. 2. It may be seen from these plots that some individual clouds which were part of the clump at 50 Myr eventually 'escaped' from the clump. The escape of the clouds will, of course, be more pronounced if there is differential rotation present. This may be better appreciated by referring to Fig. 4. Here a sphere of radius 500 pc is defined with the centre of this

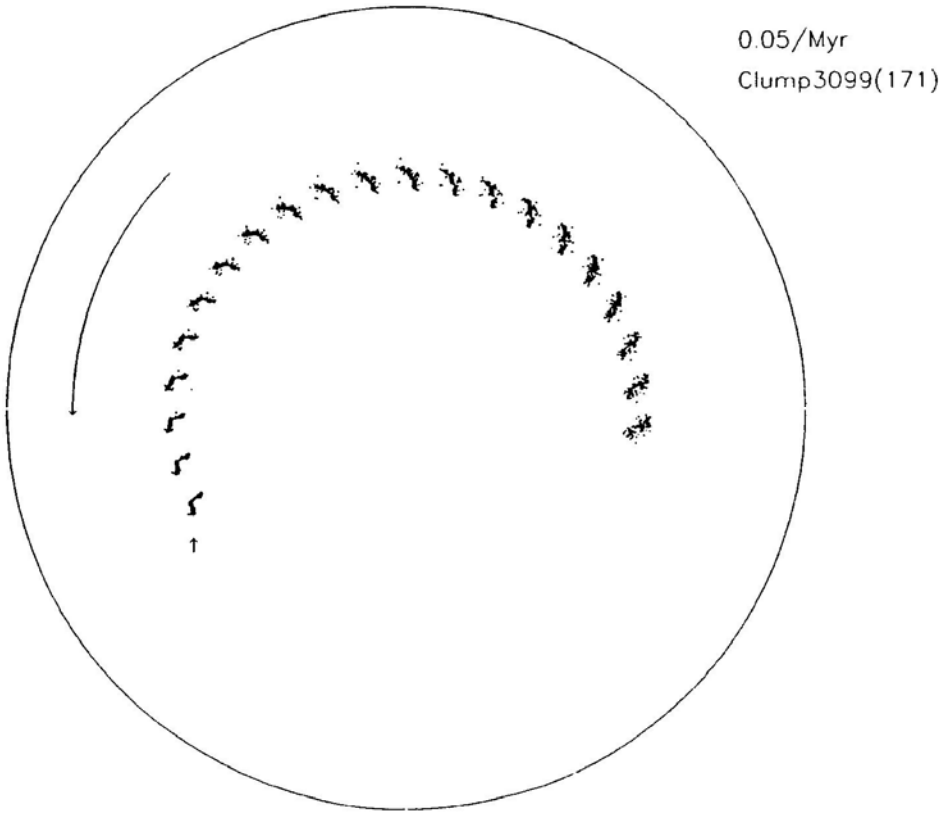

Figure 2. The evolution of the individual clouds in Clump 3099 identified at 150 Myr, which is indicated by a short arrow. The large arrow means the direction of rotation.

sphere coinciding with the centre of mass of the clump 1100 at 50 Myr. The evolution of the individual clouds within this sphere is followed till 150 Myr in steps of 10 Myr. Initially there were 713 clouds within this sphere. Since there is no differential rotation in the model the projected sphere remains unchanged. With elapse of time many clumps seem to form. As already mentioned, in the absence of differential rotation these smaller clumps do not have a chance to form larger clumps.

4. Conclusion

We have explored a model for the formation of giant molecular clouds through accretion of smaller clouds; the star disc is assumed to be rigidly rotating in this model. The main results are:

(**1**) The aggregates consist mainly of the neighbouring clouds whose collision was induced by the velocity dispersion. In this model aggregates take a shorter time to form compared to the model which includes differential rotation, although the aggregates tend to be less massive.

(**2**) By the same token, the mass of these aggregates does not increase further. The inclusion of differential rotation would have allowed for the possibility of distinct aggregates colliding amongst themselves to form larger aggregates.

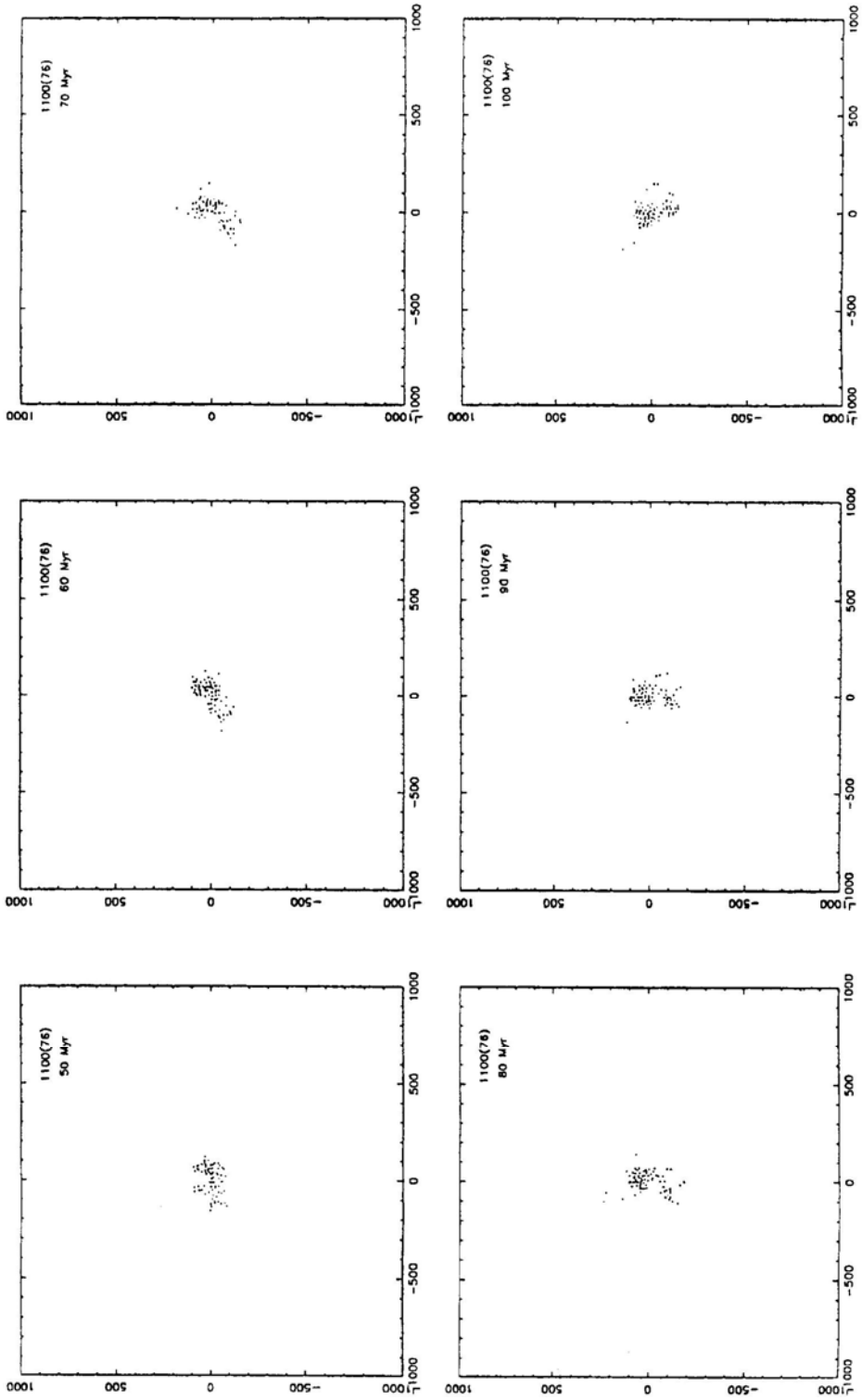

Figure 3(a). *(Continued)*

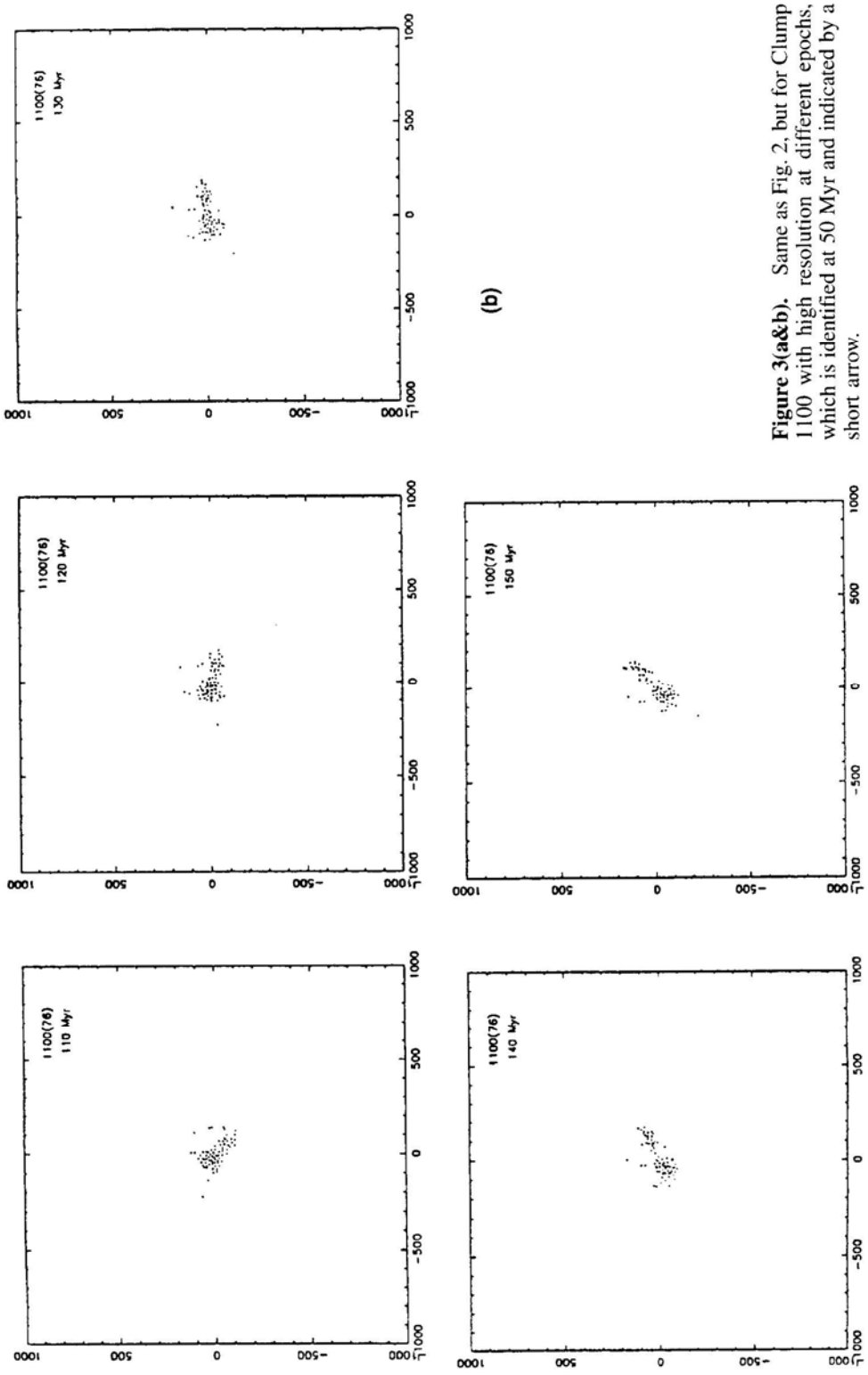

Figure 3(a&b). Same as Fig. 2, but for Clump 1100 with high resolution at different epochs, which is identified at 50 Myr and indicated by a short arrow.

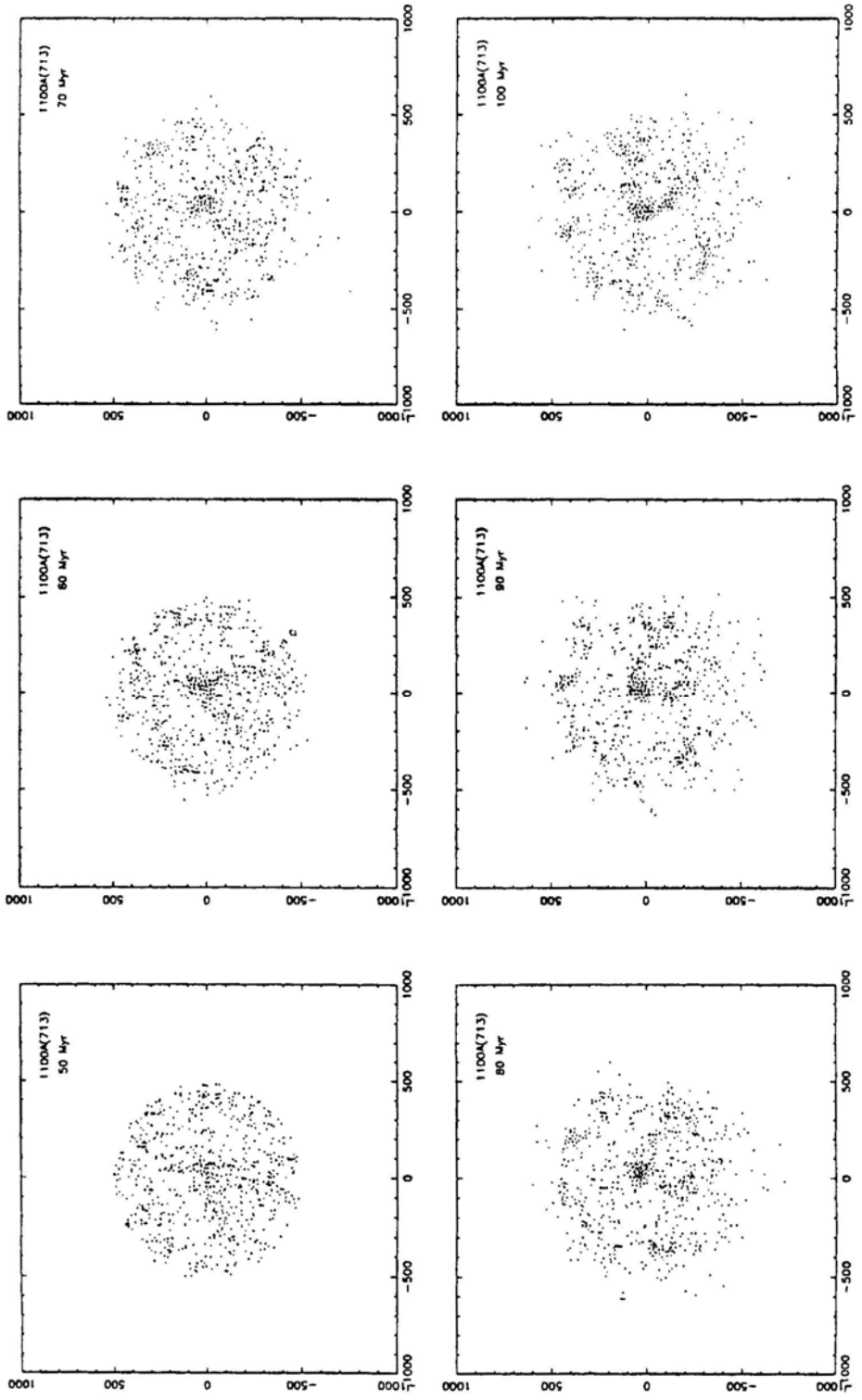

Figure 4(a). *(Continued)*

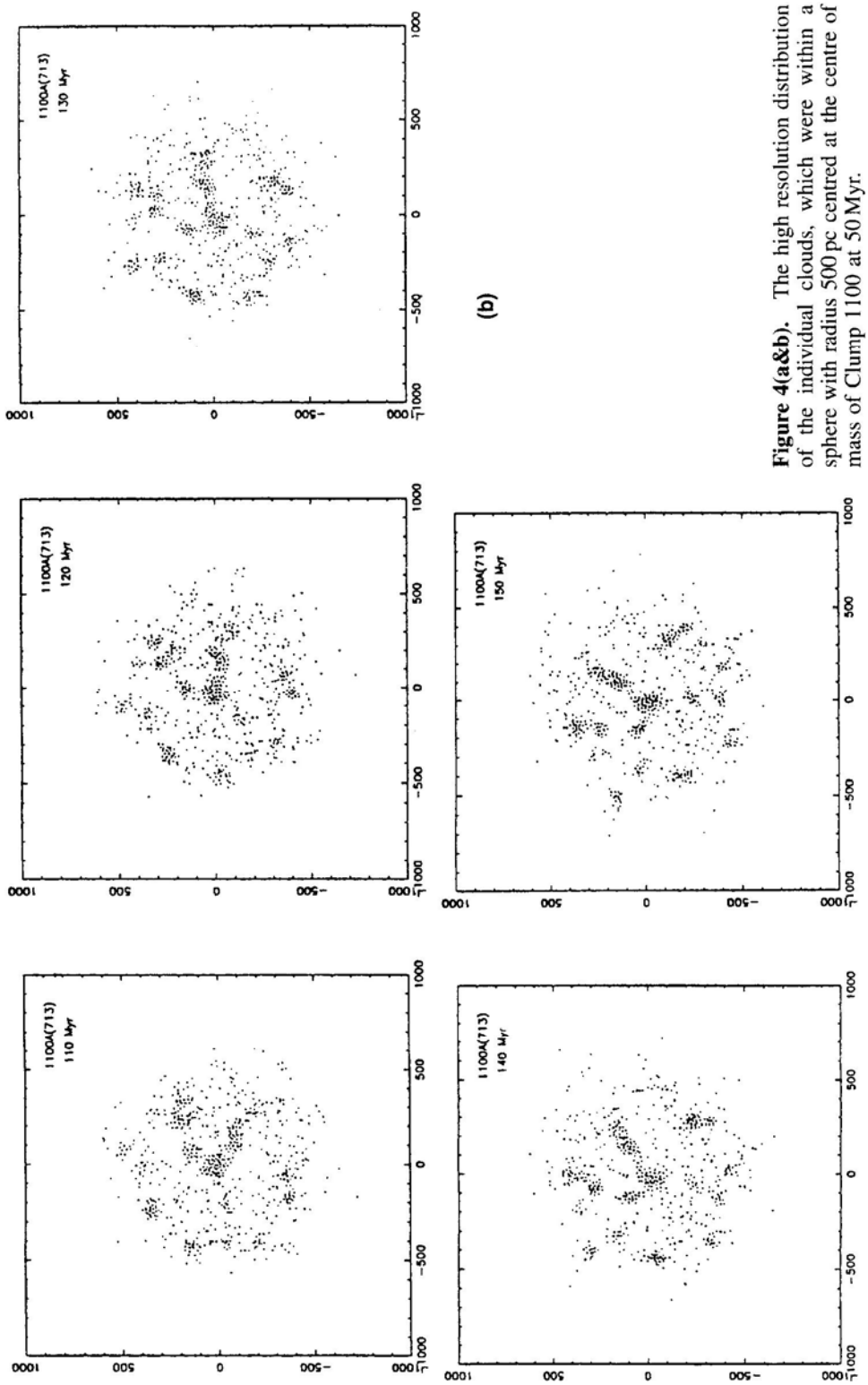

Figure 4(a&b). The high resolution distribution of the individual clouds, which were within a sphere with radius 500 pc centred at the centre of mass of Clump 1100 at 50 Myr.

Acknowledgement

This work was supported by the Young Astronomical Foundation of Shanghai Observatory and the National Nature Science Foundation of China.

References

Clifford, P., Elmegreen, B. G. 1983, *Mon. Not. R. Astr. Soc.*, **202,** 629.
Kahn, F. D., Song, G. X. 1994, *Ap. S. S.*, **211,** 127.
Kwan, J., Valdes, F. 1987, *Ap. J.*, **315,** 92.
Rivola, A. R., Solomon, P. M., Sanders, D. B. 1986, *Ap. J.*, **301,** L19.
Song, G. X. 1991, *Proceedings of Australian Astronomical Society*, **9**, 200.
Song, G. X. 1996a, *Annals of Shanghai Observatory Academia Sinica*, **17**, 234.
Song, G. X. 1996b, *J. Korean Astron. Soc.*, **29**, S165.

J. Astrophys. Astr. (1999) **20**, 89.

Preface

Research in the area of Black Holes has witnessed extremely significant and rapid growth both in experiment and in theory in recent times. This special issue of *Journal of Astrophysics and Astronomy* published by the **Indian Academy of Sciences, Bangalore** comprises a collection of important papers presented at a discussion meeting on '**The Physics of Black Holes'** held at the Indian Institute of Science in Bangalore during 8–10 December, 1997.

This meeting was made possible by the generous support received from the Jawaharlal Nehru Centre for Advanced Scientific Research, Indian Institute of Astrophysics, Raman Research Institute, all at Bangalore, S. N. Bose Institute for Basic Sciences, Calcutta and besides of course, the Indian Institute of Science.

We are grateful to all the contributors to this volume and also to the panel of reviewers for their painstaking reviews. We are grateful to Professor G. Srinivasan, Editor, *Journal of Astrophysics and Astronomy* for agreeing to bring out these proceedings as a special issue of the journal.

We thank the contributors and also the editorial staff of the journal for their co-operation, patience and understanding.

J. Pasupathy (Guest Editor)
Centre for Theoretical Studies,
Indian Institute of Science,
Bangalore.

J. Astrophys. Astr. (1999) **20,** 91–101

New Light on the Einstein-Hilbert Priority Question

John Stachel, *Department of Physics and Center for Einstein Studies, Boston University, 590 Commonwealth Avenue, Boston, Massachusetts 02215 USA.*
email:stachel@buphyc.bu.edu

1. Introduction

This talk is based on the joint work with Leo Corry and Jürgen Renn.[1] The development of general relativity may be compared to a traditional three-act play:

I (1907-1908): The formulation of the equivalence principle;

II (1912-1913): The choice of the metric tensor field as the appropriate relativistic generalization of the scalar Newtonian gravitational potential;

III (1913-1915): The search for the correct field equations for the metric tensor field.

The first two acts were essentially monologues starring Albert Einstein. The third act was a dialogue between Einstein and David Hilbert, the leading Göttingen mathematician. I have told the story of all three acts from Einstein's point of view elsewhere in some detail,[2] so I shall be brief in reviewing the first two acts. But I must say more about the third act, since this is where Hilbert entered the story and the priority question between them arose.

Acts I and II. In 1907, Einstein prepared a review article on the principle of relativity and its consequences for the various branches of physics.[3] In the course of the review, he discussed the question of a relativistic theory of gravitation.[4] It was clear from the outset that the Newtonian theory would have to be modified, since it is based on the concept of a force between pairs of particles that depends on the distance between them at some moment of time. But according to the special theory of relativity, simultaneity is no longer absolute, but depends on the inertial frame of reference being considered; hence the spatial distance between two particles is now frame -dependent. The example of electromagnetism suggests how to proceed. The electrostatic Coulomb force between two charged particles also seems to depend on their simultaneous positions; but we know how to recast the theory into a field form (Maxwell's equations) that is special-relativistically invariant, and in which interactions between charges actually propagate with a finite speed—the speed of light. The problem was to find a similar field theory of gravitation. Indeed, Newton's theory had already been cast into a field form: the gravitational potential obeys Poisson's equation with the density of matter as its source. The problem was to find a relativistic generalization of Poisson's equation. But here Einstein hit upon a fundamental feature of the gravitational field that has no parallel in electromagnetism. All objects fall with the same acceleration in a given gravitational field (Galileo's principle). Newton had explained (or better reformulated) this observation as a consequence of the equality of gravitational and inertial mass. He verified this equality to one part in a thousand, but offered no further explanation of it. It follows

from this equality and Newton's laws of motion that no mechanical experiment can ever distinguish between an inertial frame of reference, in which there is a constant gravitational field + **g,** and an accelerated frame of reference with acceleration –g with respect to the inertial frames, in which there is no gravitational field. Einstein went Newton one better, and proposed what he soon called the principle of equivalence:[5] No experiment, mechanical, optical, electromagnetic, or what have you, can distinguish between these two situations because there is no physical difference between them: inertia and gravitation are essentially just two sides of the same coin ("wesensgleich," as Einstein put it). But this means that, once gravitation is taken into account, the privileged role of the inertial frames and with it the special-relativity principle must be abandoned. Or rather, one must look for a field theory of gravitation that is invariant under some generalization of the special principle, i.e., under a group that includes transformations to (at least some) accelerated frames of reference. Einstein reached this conclusion (which forms the end of Act I) by late 1907, but for a long time he was the only physicist looking for a field theory of gravitation to accept it.[6]

At first Einstein looked for a scalar generalization of Newton's theory, based on the gravitational potential. By the middle of 1912, he had worked out what he regarded as a satisfactory theory for the case of a static gravitational field.[7] He developed a field equation for the gravitational potential, which he identified in this case with a variable speed of light $c(x,y,z)$ instead of the constant speed of the special theory; and worked out the equations of motion for a particle in the static gravitational field. He soon realized that these equations of motion could be derived from a variational principle that implicitly involves a non-flat metric tensor field. Planck had shown that the equation of motion of a particle in Minkowski space-time (i.e., in the absence of a gravitational field) could be derived from a variational principle involving the proper time along a timelike worldline in a flat space-time with Minkowski metric tensor. Einstein found that his equations of motion for a particle in a static gravitational field could be derived from the same variational principle if he just replaced the constant c in the Minkowski metric with $c(x,y,\ z)$.[8] This yields a non-flat metric tensor field $g_{\mu\nu}$ of a special type; Einstein guessed that the way to proceed beyond the static case was to introduce a general non-flat metric tensor field as the generalization of the scalar Newtonian potential.[9] He presented this idea in the opening section of a paper written in early 1913,[10] concluding Act II of the play.

Act III. The rest of his 1913 paper begins the lengthy final Act III. With the help of his friend and colleague Marcel Grossmann, Einstein succeeded in showing that the effects of gravitation on all other physical processes could be taken into account by means of the non-flat metric tensor field, and that the resulting equations could be put into a generally covariant form. That is, the group of transformations under which these equations are invariant includes all invertible differentiable point transformations. (Einstein, in keeping with the mathematical treatments of the time, gave the transformations a passive interpretation as coordinate transformations; but from the modern viewpoint it is better to give them an active interpretation as differentiable point transformations, or diffeomorphisms.) Thus the group automatically includes all possible transformations to accelerated frames of reference, which seemed to Einstein a very satisfactory solution to the problem of generalizing the relativity principle.

However, one major difficulty remained: He could not make the differential equations for the gravitational field itself generally covariant. If they were based on the Ricci tensor, just about the only generally-covariant candidate for second-order field equations constructed from the metric tensor, the equations relating the latter to its source, the stress-energy-momentum tensor describing all non-gravitational matter and fields, did not seem to allow passage to the well-verified static Newtonian limit (Poisson's equation).[11] The root of the problem lay in a misconception of what the static metric tensor field should be, and how the Newtonian limit should be taken. But this did not emerge until well after Einstein had abandoned generally covariance in favor of a set of non-covariant field equations, now usually called the Einstein-Grossmann field equations, first presented in the 1913 paper.

Being Einstein, he did not rest content with the Newtonian-limit argument against field equations based on the Ricci tensor (indeed he seems to have become suspicious of that argument rather soon). He soon came up with what appeared to be a devastateing argument against generally covariant gravitational field equations of any kind, which he dubbed the hole argument:[12] Consider a finite region of space-time (the hole), inside of which there is no matter. Einstein required of any gravitational theory that the specification of all sources of the field outside the hole -together with any appropriate boundary conditions on the hole, at infinity, etc. -should uniquely determine the gravitational field inside the hole. But this seems to be impossible if the gravitational field equations are generally covariant: Suppose we have one solution to the field equations outside and inside the hole. Then let us carry out a diffeomorphism (i.e., a one-one differentiable point transformation)[13] that reduces to the identity outside and on the boundary of the hole, but is different from the identity inside the hole. For generally covariant equations, the diffeomorphic-transform of a solution is also a solution to the field equations. Thus, the same source and boundary conditions outside the hole correspond to two distinct solutions inside the hole - indeed, since the group of diffeomorphisms depends on four arbitrary functions, the number of such distinct solutions is unlimited. Therefore one must look for gravitational field equations that are not generally covariant; yet they must be invariant under a group of transformations that includes at least some non-linear ones (accelerations) –but not so many as to fall foul of the hole argument.

This is how the situation stood, as Einstein saw it, in late 1913, and so it stayed until late 1915. He remained wedded to the Einstein-Grossmann field equations and kept trying to find better and better arguments in their favor; in particular, arguments for their uniqueness and for their invariance under the maximum invariance group compatible with avoiding the hole argument.

It was not until the end of 1915, after he had returned to general covariance, that he found the flaw in his hole argument against it: In modern terms, the hole argument presupposes that some intrinsic physical significance attaches to the points of a four-dimensional manifold *before* the metric tensor field is defined on it. Once this idea is abandoned, and it is accepted that the points of the manifold inherit all their physical properties from the metric tensor field (and any other physical fields that may be present in regions outside the hole), then it is clear that two mathematical solutions inside the hole that differ only by a diffeomorphism transformation represent the *same* gravitational field; so the uniqueness requirement is satisfied. But as noted above, Einstein did not realize this until after he had returned to general covariance for other reasons.

Enter Hilbert. In the summer of 1915, while he was at the stage of the non-covariant theory, Einstein was invited to give a series of lectures on his gravitational theory in Göttingen (he stayed from June 29th to July 7th), then the mathematical capital of Germany if not the world. In addition to Hilbert, it boasted the presence of Felix Klein, the doyen of German mathematics, Emmy Noether, and many other luminaries. Einstein, who had been having his troubles getting the physicists to accept his approach to gravitation, was delighted with the reception of his work by the Göttingers.

> In Göttingen I had the great pleasure of seeing everything understood, down to the details. I am quite enthusiastic about Hilbert. An important man. I am very curious about his opinion.[14]

For his part, Hilbert, who had been working on Mie's nonlinear electrodynamic theory, which aimed at explaining the structure of matter, was delighted with Einstein's approach to gravitation. He decided to combine Mie's four electrodynamic potentials q_μ with Einstein's gravitation potentials $g^{\mu\nu}$ in order to forge a theory of matter that would enable him to explain the structure of the electron, as well as its curious non-radiative behavior in the Bohr atom.[15] In the fall of 1915, Einstein and Hilbert entered into an intense correspondence - the only serious one either had on this topic -just as Einstein was reaching the crisis point in his efforts to shore up the Einstein-Grossmann equations, leading to their abandonment.[16] Einstein's letters to Hilbert report such milestones as: his return to general covariance and re-adoption of field equations based on equating the Ricci tensor to the stress-energy tensor in two papers submitted on 4th and 11th November;[17] his solution of the problem of the anomalous precession of the perihelion of Mercury submitted on 18th November;[18] and his adoption of the final form of the field equations submitted on 25th November.[19] These equations are still based on the Ricci tensor, but now with $-\ 1/2 g\mu_V T$ (where T is the trace of the stress-energy tensor—this term is often referred to as the trace term) subtracted from the stress energy tensor on the right hand side. Meanwhile, Hilbert had not been inactive. He, too had formulated field equations—but for a combined theory of gravitation and Mie's electromagnetism, based on a generally-covariant variational principle.

The accepted account of just what happened during this period has been succinctly formulated in a recent biography of Einstein:

> In the decisive phase [of his work on general relativity] Einstein even had a congenial colleague, though this caused him more annoyance than joy, as it seemed to threaten his primacy. "Only one colleague truly understood it, and he now tries skilfully to 'nostrify' [i.e., appropriate] it," he complained to [Heinrich] Zangger about what he evidently regarded as an attempt at plagiarism. This colleague was none other than David Hilbert[...] What must have irritated Einstein was that Hilbert had published the correct field equations first—a few days before Einstein. [...]
>
> In November, when Einstein was totally absorbed in his theory of gravitation, he essentially corresponded only with Hilbert, sending Hilbert his publications and, on November 18th, thanking him for a draft of his article. Einstein must have received that article immediately before writing this letter. Could Einstein, casting his eye over Hilbert's paper, have discovered the term which was still lacking in his own equations [i.e., the trace term], and thus 'nostrified' Hilbert?[20]

The author argues that this is improbable, and most authorities agree; but the point is that, on the accepted account, which based on comparison of the published version of Hilbert's paper on the foundations of physics[21] with Einstein's papers, the dating of the papers makes such an act of plagiarism by Einstein *possible.*

Recently, Leo Corry found a set of printer's proofs of Hilbert's paper, marked "First proofs of my first note" in his handwriting.[22] As a result of a study of these proofs and their dating, one can conclude that Einstein *could not* have taken anything pertaining to the field equations from Hilbert's paper, or the summary sent him in November. Indeed, the situation is quite the reverse: The question is whether Hilbert might have taken from Einstein's 25th November paper, because it was *quite possible* for him to do so. Let me explain how this reversal of the direction of possible influence came about.

The published article by Hilbert is dated, "submitted to the session [of the Göttingen Academy of Sciences] of 20 November 1915." Einstein's conclusive paper, in which he gave the final form of his generally-covariant field equations was submitted to the session of the Prussian Academy of Sciences in Berlin of 25th November 1915. Bearing in mind that Hilbert sent a summary or draft of his paper to Einstein, to which the latter replied on November 18th, it seems that Einstein had available at least the essence of Hilbert's work a week before he submitted his conclusive paper of 25th November. And so he did—but it was Hilbert's work as presented in the proofs. These proofs are dated exactly the same way as the published version – "submitted to the session of 20 November 1915" – and bear a printer's date stamp –"6th December 1915" – indicating that is when they were typeset. These dates would mean nothing if the proofs were essentially the same as the published version. But they are not: they differ in several important respects, the most crucial being that:

(**1**) The theory that they present is not generally covariant; in addition to generally covariant field equations, there are four equations whose purpose is precisely to limit the coordinate system.

(**2**) The generally covariant gravitational field equations are not written down explicitly but merely as the variational derivative, which is not evaluated, of a Lagrangian. In particular, there is no hint of a trace term.

Since the proofs are dated November 20th, they represent the status of Hilbert's work submitted on that date to the Academy and presumably previewed in his earlier communication to Einstein. Since they bear the printer's stamp dated December 6th, any changes Hilbert made before that date are incorporated in the proofs (indeed, he had probably made no changes since he marked them "first proofs"). Thus, there is no chance that Einstein could have learned about the need for a trace term in his equations from a perusal of Hilbert's work before Einstein submitted his paper of November 25th including the trace term.

On the other hand, Hilbert did not start correcting the proofs until 6th December at the earliest (he finally gave up the revision and completely rewrote the article, as we shall see); and Einstein's 25th November paper was published on 2nd December. Since Hilbert's paper was not actually published until March 1916, there *was* plenty of opportunity for Hilbert to see Einstein's paper before publishing his own. Indeed, there is no doubt that he did: Hilbert's paper, which was actually published in the issue of the *Göttinger Nachrichten* dated 31st March, includes references to all of Einstein's November papers including that of 25th November.

Since we now know that Hilbert drastically rewrote his original (proofs) version sometime between December 1915 and March 1916, the question is no longer what could Einstein have gotten from Hilbert, but: What could Hilbert have gotten from Einstein?

Hilbert's proofs: We shall only discuss the two major points of difference between the proofs and the published paper that are most relevant to the priority question; for more detailed comparisons one can consult the references in note 15. First of all, as noted above, in the proofs Hilbert asserts that the theory he is developing cannot be generally covariant. He bases this assertion on a slighly more sophisticated version of Einstein's hole argument (he cites Einstein's 1914 discussion of that argument), involving the Cauchy problem on an initial hypersurface rather than boundary conditions on a hole:

> Since our mathematical theorem [a version of what later became known as Noether's theorem, which guarantees the existence of four identities between generally covariant field equations] shows that the previous axioms I [the existence of a Lagrangian for the four electromagnetic and ten gravitational equations] and II [the general covariance of that Lagrangian] can only provide ten essentially independent equations for the 14 potentials [of gravitation and electromagnetism]; and further, maintaining general covariance makes quite impossible more than ten essentially independent equations for the 14 potentials $g_{\mu\nu}$, q_s; then, in order to keep the deterministic character of the fundamental equations of physics, in correspondence with Cauchy's theory of differential equations [that is, to have a well-posed Cauchy problem], the requirement of four further non-invariant equations to supplement [the generally covariant gravitational equations] is unavoidable. In order to find these equations I start out by setting up a definition of the concept of energy (pp. 3–4 of the proofs, translation from reference in note 1).

Without going into the details of the resulting energy theorem, I shall only cite Hilbert's:

> Axiom III (axiom of space and time). The space-time coordinates are those specific world parameters [Hilbert's name for arbitrary coordinates] for which the energy theorem . . . is valid.
> The validity of [the energy] equation ... is a consequence of Axiom III; these four differential equations .. supplement the [generally-covariant] gravitational equations . . . to yield a system of 14 equations for the 14 potentials $g^{\mu\nu}$, q_s: *the system of the fundamental equations of physics* (original emphasis; p. 7 of the proofs, translation from reference in note 1).

Thus, Hilbert was adopting Einstein's line of reasoning from 1913–mid 1915 just at the time that Einstein was abandoning it, in favor of a return to general covariance. Of course, Hilbert did include ten generally covariant gravitational equations in his "system of the fundamental equations of physics," but Einstein was not impressed by this move. Immediately after receiving Hilbert's account of his new theory, Einstein wrote him on 18th November:

> The difficulty was not to find generally covariant equations for the $g^{\mu\nu}$; this is easy with the help of the Riemann tensor. What was difficult instead was to recognize that these equations form a generalization, and indeed a simple and natural

> generalization of Newton's law. I only succeeded in doing so in the last few weeks (I sent you my first communication); while I had already considered the only possible generally covariant equations, which have now proven to be the correct ones, three years ago with my friend Grossmann. Only with a heavy heart did we give them up, because the physical discussion appeared to show me their incompatibility with Newton's law (my translation).[23]

That Einstein's claims about his earlier work are quite accurate is proved by an examination of his 1912 Zürich research notebook.[24]

The second major point about Hilbert's proofs concerns the form of his gravitational field equations. In the proofs, these equations do not appear explicitly. Hilbert's Lagrangian contains a gravitational term, which is probably $\sqrt{-g}R$ in modern terms,[25] where R is the Ricci scalar (unfortunately, the few lines where the Lagrangian is defined are missing from the proofs), and he indicates that the gravitational field equations result from taking the variational derivative of this term with respect to the metric tensor. But he does not evaluate this variational derivative at all.

Why did Hilbert decide to publish his work before he had accomplished his original aim: to explain the behavior of the electron? He submitted his work to the Göttingen Academy two days after Einstein's letter of 18th November,[26] which opened with a sentence I have not yet quoted:

> The system given by you agrees - as far as I can see – completely with what I have found in the last few weeks and sent to the Academy;

and closed with the news that:

> I have today handed in a work to the Academy, in which without any supplementary hypothesis I have quantitatively derived the perihelion motion of Mercury, discovered by Leverrier, from general relativity (my translation).[27]

As Tilman Sauer suggests,[28] Hilbert may well have felt some urgency to establish his own claims in the light of Einstein's successes.

It was shortly after receiving Hilbert's summary of his theory, which presumably contained both points discussed above, that Einstein complained about Hilbert's "nostrification" of his theory. The letter, cited earlier, continues:

> In my personal experience, I have hardly ever learned to know better the wretchedness of human beings than on the occasion of this theory and what is connected with it. But I don't give a damn (my translation).[29]

Hilbert was apparently aware of Einstein's unhappiness. When he got the proofs of his paper (presumably on or about 6th December), he added an additional complimentary reference to Einstein among several other handwritten insertions; but then he seems to have realized that he would have to completely rewrite the paper in the light of Einstein's four November papers (as noted above, they are all referenced in the published version of Hilbert's paper).[30]

The published paper: Hilbert removed all references to his argument about the need for four non-generally covariant equations to supplement the ten generally-covariant gravitational field equations. He did not discuss the subject of causality for the field equations in this paper at all. Only in 1917, in his second paper on the foundations of physics, did he return to it.[31] He simply asserted Axioms I and II, and dropped Axiom

III and all other references to preferred coordinate systems. He completely changed his energy discussion to take into account the full general covariance of the present version of his theory.

He also now wrote down the explicit form of the gravitational field equations that follow from the variational derivative of his Lagrangian: the trace term $1/2g_{\mu\nu}R$, where R is the trace of the Ricci tensor, is subtracted from the Ricci tensor, giving what we now call the Einstein tensor as the lefthand side of the field equations. (His ighthand side is the stressenergy tensor for Mie's electrodynamic theory.)

Indeed, he was the first to give the field equations in this form. As noted above, Einstein, in his paper of 25th November had the Ricci tensor on the left hand side, with the trace of the stressenergy tensor subtracted on the righthand side. The two forms are completely equivalent, of course; but whether Hilbert or Einstein immediately recognized this is not clear; nor if Hilbert did recognize it, is it clear whether it influenced him in any way. What is clear is that the argument he offers for the form of the left hand side of the field equations is fallacious. Instead of evaluating the Einstein tensor he simply says:

> ...which follows easily without calculation from the fact that, except for $g_{\mu\nu}$, $K_{\mu\nu}$ [the Ricci tensor] is the only tensor of second order and K [the Ricci scalar] is the only invariant that can be constructed from only the $g^{\mu\nu}$ and its first and second order partial derivatives. . . (pp. 404–405 of ref. 20, translation from ref. in note 1)

The argument is fallacious, of course, because many other tensors of second order in the derivatives of the metric and many other invariants can be constructed from the Riemann tensor. Even if one requires linearity in the second derivatives, the fact that the coefficient of the trace term is exactly 1/2 remains quite undetermined by this argument. Hilbert himself seems to have recognized its untenability: When he reprinted the paper in 1924,[32] he dropped it and simply sketched a correct method for evaluating the variational derivative.

Hilbert added several clear aknowledgements of Einstein's priority in constructing a generally-covariant gravitational theory based on the metric tensor, and continued to acknowledge it on many later occasions. If he had only added a few words to the dateline of the published paper: "revised version submitted on any date after 6th December," the whole later priority question could have been avoided.[33]

At any rate, the priority issue between Einstein and Hilbert was happily resolved quite soon. On 20th December 1915, well before Hilbert's paper appeared in print, Einstein wrote him:

> There has been a certain resentment between us, the cause of which I do not wish to analyze. I have fought against the feeling of bitterness associated with it, and indeed with complete success. I again think of you with unclouded friendliness and I ask you to attempt the same with me. It is objectively a pity if two genuine chaps, who have liberated themselves to some extent from this sorry world, do not give each other mutual pleasure (my translation).[34]

Notes

[1] Leo Corry, Jürgen Renn and John Stachel, "Belated Decision in the Hilbert-Einstein Priority Dispute," *Science 278* 1270–1273 (1997).

[2] See John Stachel, "The Genesis of General Relativity," in H. Nelkowski *et al,* eds., *Einstein Symposion Berlin, Lecture Notes in Physics 100* (Berlin/Heidelberg/ New York: Springer-Verlag, 1980), pp. 428–442; "Einstein and the Rigidly Rotating Disc," in A. Held ed. *General Relativity and Gravitation One Hundred Years After the Birth of Albert Einstein* (New York: Plenum, 1980), pp. 1–15, reprinted in D. Howard and J. Stachel, eds., *Einstein and the History of General Relativity/ Einstein Studies, vol. 1* (Boston/Basel/Stuttgart: Birkhäuser, 1989), pp. 48-62; "How Einstein discovered general relativity: a historical tale with some contemporary morals," in M. A. H. MacCallum, ed., *General Relativity and Gravitation/ Proceedings of the 11th International Conference on General Relativity and Gravitation* (Cambridge: Cambridge University Press, 1987), pp. 200-208; "Einstein's Search for General Covariance, 1912-1915," in D. Howard and J. Stachel, eds *Einstein and the History of General Relativity/Einstein Studies, vol. 1* (Boston/ Basel/Stuttgart: Birkhäuser, 1989), pp. 63–100.

[3] Albert Einstein, "Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen," *Jahrbuch der Radioaktivität und Elektronik 4,* 411–462 (1907); reprinted in John Stachel et al., eds. *The Collected Papers of Albert Einstein,* vol. 2, *The Swiss Years: Writings 1900–1909,* (Princeton University Press, 1989), pp. 433-488. At that time he still referred to the principle, rather than the theory, of relativity, when referring to what was later called special relativity.

[4] See ref. in note 3, pp. 476–484.

[5] See Albert Einstein, "Lichtgeschwindigkeit und Statik des Gravitationsfeldes," *Annalen der Physik 38,* 355–369 (1912), p. 365.

[6] Some physicists looking for a relativistic theory of gravitation continued to adhere to the special-relativistic principle (Gunnar Nordström and Gustav Mie); one physicist abandoned the relativity principle entirely in his search for a field theory of gravitation (Max Abraham).

[7] See ref. in note 5, and Albert Einstein, "Zur Theory des statischen Gravitationsfeldes," *Annalen der Physik 38,* 443–458 (1912); both papers are reprinted in Martin J. Klein *et al,* eds, *The Collected Papers of Albert Einstein,* vol. 4, *The Swiss Years: Writings 1912–1914,* (Princeton University Press, 1995), pp. 130–144,147–162.

[8] See second reference in note 7, p. 458.

[9] The consideration of stationary gravitational fields, and in particular the gravita tional field on a rotating disc, also played a role in convincing Einstein of the need to go beyond flat space to solve the gravitational problem. See John Stachel, "Einstein and the Rigidly Rotating Disc," reference in note 2.

[10] See Albert Einstein and Marcel Grossmann, *Entwurf einer verallgemeinerten Relativitätstheorie und einer Theorie der Gravitation,* (Leipzig/Berlin, B.G. Teubner, 1913). Part I, "Physikalische Teil," was written by Einstein, Part II, "Mathematischer Teil," by Grossmann. The paper is reproduced in Martin J. Klein *et al,* eds , *The Collected Papers of Albert Einstein,* vol. 4, *The Swiss Years: Writings 1912 1914,* (Princeton University Press, 1995), pp. 303–339.

[11] See John Stachel, "Einstein's Search for General Covariance, 1912–1915," reference in note 2.

[12] For more detailed discussions, see John Stachel, "The Meaning of General Covariance: The Hole Story,"in John Earman et al., eds., *Philosophical Problems of the Internal and External World/Essays on the Philosophy of Adolf Grünbaum*

(Konstanz: Universitätsverlag/Pittsburgh: University of Pittsburgh Press, 1993), pp 129–160; and "Einstein's Search for General Covariance, 1912–1915," reference in note 2.

[13] By speaking of diffeomorphisms instead of coordinate transformations, I am modernizing the argument, but am not being unfaithful to the essence of Einstein's argument.

[14] Einstein to Arnold Sommerfeld, 15 July 1915, in Robert Schulmann *et al.,* eds. *The Collected Papers of Albert Einstein,* vol. 8, *The Berlin Years: Correspondence, 1914–1918 Part A: 1914–1917,* (Princeton University Press, 1998), Doc. 96, p.147.

[15] For a discussion of Hilbert's work during this period, see Leo Corry, "From Mie's Electromagnetic Theory of Matter to Hilbert's Unified Foundation of Physics," *Studies in History and Philosophy of Modern Physics* **30B,** 159-183 (1999); Tilman Sauer, "The Relativity of Discovery: Hilbert's First Note on the Foundations of Physics," *Archive for History of Exact Sciences* 53, 529–575 (1999); and Jürgen Renn and John Stachel, "Hilbert's Foundations of Physics: From a Theory of Everything to a Constituent of General Relativity," (Berlin, Max-Planck-Institut für Wissenschaftsgeschichte Preprint 118, 1999).

[16] For discussions of the reasons for his abandonment of the Einstein-Grossman equations and his return to general covariance, see John Stachel, "Einstein's Search for General Covariance," reference in note 2; John Norton, "How Einstein Found His Field Equations , 1912–1915," *Historical Studies in the Physical Sciences 14,* 253–316 (1984); Michel Janssen, "Rotation as the Nemesis of Einstein's Entwurf Theory," in Hubert Goenner *et al.* (eds), *The Expanding Worlds of General Relativity, Einstein Studies,* vol. 7, (Boston/Basel/Berlin, Birkäuser 1999), pp. 127–157.

[17] Albert Einstein, "Zur allgemeinen Relativitätstheorie," *Königlich Preußischen Akademie der Wissenschaften* (Berlin), *Sitzungberichte,* 778–786 (1915), "Zur allgemeinen Relativitätstheorie (Nachtrag), *ibid.,* 799–801 (1915).

[18] Albert Einstein, "Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie," *Königlich Preußischen Akademie der Wissenschaften* (Berlin), *Sitzungberichte,* 831–839 (1915).

[19] Albert Einstein, "Die Fieldgleichungen der Gravitation," *Königlich Preußischen Akademie der Wissenschaften* (Berlin) *Sitzungsberichte,* 844–847 (1915). The papers cited in notes 17, 18 and this note are reprinted in Martin J. Klein et al, eds. *The Collected Papers of Albert Einstein,* vol. 6, *The Berlin Years: Writings 1914-1917* (Princeton University Press, 1996), pp. 215–223, 226–228, 234–242, and 245–248.

[20] Albrecht Fölsing, *Albert Einstein: a biography* (Viking, New York, 1997), pp. 375–376. The translation has been slightly modified and is taken from ref. 1.

[21] David Hilbert, "Die Grundagen der Physik (Erste Mitteilung)," *Nachrichten von der Königl. Gesellschaft der Wissenschaften zu Göttingen. Mathematisch-physikalische Klasse,* 395–407 (1915).

[22] The proofs are in the Handschriftenabteilung of the Staats- und Universitätsbibliothek Göttingen, Cod. Ms. D. Hilbert 634.

[23] Albert Einstein to David Hilbert, 18 November 1915, in Robert Schulmann *et al,* eds. *The Collected Papers of Albert Einstein,* vol. 8, *The Berlin Years: Correspondence, 1914–1918 Part A: 1914–1917,* (Princeton University Press, 1998), Doc. 148, pp. 201–202.

[24] Albert Einstein, "Research Notes on a Generalized Theory of Relativity," in Martin J. Klein *et al.,* eds. *The Collected Papers of Albert Einstein,* vol. 4, *The Swiss Years: Writings 1912–1914,* (Princeton University Press, 1995), pp. 201–269 for a transcription, pp. 630–682 for a facsimile. For a brief review of its significance, see Jürgen Renn and Tilman Sauer," Einsteins Züricher Notizbuch," *Physikalische Blätter 52,* 865–872 (1996).

[25] Hilbert does not introduce a minus sign since he is working with one imaginary coordinate, and he writes (K) for what we now usually designate by (R).

[26] It is a sad reflection on the state of the mails today that in wartime Germany, mail between Berlin and Göttingen seems to have taken only a day or so.

[27] Reference in note 23.

[28] See the reference in note 15.

[29] Albert Einstein to Heinrich Zangger, 26 November 1915, in Robert Schulmann *et al.,* eds. *The Collected Papers of Albert Einstein,* vol. 8, *The Berlin Years: Correspondence, 1914–1918 Part A: 1914–1917,* (Princeton University Press, 1998), Doc. 152, pp. 204–205; citation from p. 205.

[30] References in notes 17, 18 and 19.

[31] David Hilbert, "Die Grundlagen der Physik (Zweite Mitteilung)," *Nachrichten von der Königl. Gesellschaft der Wissenschaften zu Göttingen. Mathematisch-physikalische Klasse,* 53–76 (1917).

[32] It was reprinted, with extensive changes, in *Mathematische Annalen 92,* 1 (1924).

[33] It was not impossible to do this in the *Göttinger Nachrichten.* Related papers by Klein and Noether, for example, carry such revised datelines (see the paper by Renn and Stachel cited in note 15).

[34] Albert Einstein to David Hilbert, 20 December 1915 in Robert Schulmann *et al.,* eds. *The Collected Papers of Albert Einstein,* vol. 8, *The Berlin Years: Correspondence, 1914–1918 Part A: 1914–1917,* (Princeton University Press, 1998), Doc. 167, pp. 222.

J. Astrophys. Astr. (1999) **20**, 103–120

Black Holes and Rotation

C. V. Vishveshwara, *Indian Institute of Astrophysics*, *Koramangala, Bangalore 560034*, *India.*

Abstract. In this article, we first consider briefly the basic properties of the non-rotating Schwarzschild black hole and the rotating Kerr black hole Rotational effects are then described in static and stationary spacetimes with arial symmetry by studying inertial forces, gyroscopic precession and gravi-electromagnetism. The results are applied to the black hole spacetimes.

Key words. Black holes—rotation—inertial forces—gyroscopic precession — gravi-electromagnetism.

1. Introduction

In the last three decades, there has emerged a phenomenal amount of research on black holes. This includes studies on their geometrical structure, their physical aspects and phenomena occurring in their strong gravitational fields. These studies were initially confined to the simpler case of the nonrotating Schwarzschild black hole and later on extended to the more complex case of the rotating Kerr black hole. The effect of rotation inherent to the Kerr spacetime manifests itself in almost all physical phenomena, sometimes in a profound manner. For instance, it is rotation that is responsible for the existence of the ergosphere in the Kerr geometry and the consequent possibility of energy extraction via the Penrose process. The subject of black holes and rotation is quite vast. Here we shall review only a few ideas with emphasis on the work my coworkers and I have done over the years. First we shall very briefly compare and contrast some of the basic attributes of the Schwarzschild and Kerr spacetimes. We shall then discuss the notion of ‘rest frames’ in the two geometrics which is quite important in studying physical phenomena. In recent years, there has been considerable interest in the general relativistic analogues of inertial forces and their possible reversal in the strong gravitational fields of black holes and ultra compact objects. At the same time there are two other phenomena apparently related to the inertial forces, namely gyroscopic precession and gravito-electromagnetism. We shall demonstrate and discuss how these three aspects of black hole spacetimes can be related to one another in a covariant and elegant manner utilizing the Killing vector fields. The formalism will be presented at a very general level in the context of arbitrary static and stationary spacetimes with obvious application to the Schwarzschild and Kerr metrics as specific examples.

As has been mentioned already, the above topics are but a small part of an extensive field. They suffice, however, to illustrate the important role played by rotation in black hole physics.

2. Basic properties

In this section we give a very brief account of some of the basic properties of the Schwarzschild and Kerr black holes. These aspects are well known and are given here for the sake of completeness.

2.1 *Line elements*

Schwarzschild:

$$ds^2 = \left(1 - \frac{2m}{r}\right)dt^2 - \left(1 - \frac{2m}{r}\right)^{-1} dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2) \tag{1}$$

where $m = MG/c^2$; M = Mass; $c = G = 1$.

Kerr:

$$ds^2 = \left(1 - \frac{2m}{\Sigma}\right)dt^2 + 2\frac{2mar}{\Sigma}\sin^2\theta dt\, d\phi - \left(a^2 + r^2\frac{2ma^2r}{\Sigma}\sin^2\theta\right)\sin^2\theta d\phi^2 - \frac{\Sigma}{\Delta}dr^2 - \Sigma d\theta^2 \tag{2}$$

where m = Mass, $J = ma$ = *Angular Momentum,* $\Sigma = r^2 + a^2\cos^2\theta, \Delta = m^2 - 2mr + a^2$

2.2 *Spacetime symmetries*

Both the Schwarzschild and the Kerr spacetimes admit a globally timelike Killing vector field ξ. The Schwarzschild metric is spherically symmetric with three rotational Killing vector fields.

$$L_x = y\frac{\partial}{\partial z} - z\frac{\partial}{\partial y}, \quad L_y = z\frac{\partial}{\partial x} - x\frac{\partial}{\partial z}, \quad L_z \equiv \eta = x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x} \tag{3}$$

satisfying the usual commutation relations $[L_x, L_y] = -L_z$ etc. while ξ and η commute,

$$[\xi, \eta] = 0$$

Here the coordinates (x, y, z) are related to (r, θ, ϕ) by the flat space formulae connecting the Cartesian coordinates to polar coordinates. Furthermore $\xi \eta = 0$, i.e $g_{03} = 0$, signifying the absence of rotation. In the case of the stationary, axially symmetric Kerr spacetime only $L_Z \equiv \eta$ exists in addition to the timelike Killing vector ξ^α. They satisfy the commutation relation $[\xi,\eta] = 0$, but $\xi . \eta = g_{03} \neq 0$, because of the inherent rotation.

2.3 *Source*

The Schwarzschild spacetime can represent the gravitational field exterior to different spherically symmetric sources. These include the static, collapsing or expanding, and

oscillating spherical sources. And of course, the Schwarzschild spacetime without a material source of finite extension corresponds to the nonrotating, spherical black hole. In contrast, no realistic source has been matched on to the Kerr spacetime. Because of the unusual multipole structure of the Kerr spacetime involving infinite sequence of multipole moments m, ma^2 , ma^4,... it has been conjectured that no material source exists as the interior for the Kerr metric. This fact has not been proved but seems to be true. Neverthless, the Kerr spacetime corresponds to the rotating, stationary black hole with axial symmetry.

2.4 *Black hole structure*

As is well known, the Schwarzschild black hole is located at $r = 2m$. The global timelike Killing vector becomes null on this surface defining the static limit. Furthermore, the normal to this surace is also null. Consequently, the light cone is tangential to this surface which therefore acts as a one-way membrane. Or equivalently it is the event horizon. In other words, the surface $r =2m$ is both the static limit and the event-horizon.

This is no longer true in the case of the Kerr spacetime, an important effect of rotation. The stationary limit, namely the surface on which the global timelike Killing vector becomes null, is given by $r = m + (m^2 - a^2 \cos^2 \theta)^{1/2}$ provided $\alpha \leq m$. On the other hand the null event horizon happens to be the surface $r = m + (m^2 - a^2)^{1/2}$. The region between the two surfaces is the ergosphere which makes unusual phenomena like the Penrose process and super-radiance possible at the cost of the rotational energy of the Kerr black hole.

2.5 *Uniqueness and stability*

The Schwarzschild and the Kerr black holes represent uniquely the time independent, asymptotically flat, uncharged black holes without and with rotation respectively. Stability of the Schwarzschild black hole has been established completely. The Kerr black hole has been shown to be stable against all normal modes. However, the completeness of the radial modes has not been proved though it may be reasonable to assume that this is true.

3. The global rest frame

Within the framework of special theory of relativity, or equivalently in the flat space-time, the global rest frame of an inertial observer plays a fundamentally important role. The general relativistic analogue of such a frame of reference in stationary, axisyrnmetric spacetimes is likewise important in studying physical phenomena in a meaningful way. In the case of black holes, the difference between the spacetimes of the nonrotating Schwarzschild black hole and the rotating Kerr black hole shows up clearly in defining these frames.

The rest frame in flat spacetime is adapted to the inertial observer following a worldline along time t. This is the direction of the timelike Killing vector ξ^α. The four velocity of the observers at different spatial points are orthogonal to the hyperspace t = constant. The four velocity is given by

$$u^a = e^{\psi}\xi^a; \quad u^a u_a = 1; \quad e^{\psi} = (\xi^a \xi_a)^{-\frac{1}{2}} \tag{4}$$

and

$$\xi^a = \delta^a_0 \quad \text{and} \quad \xi_a = t_{,a} \tag{5}$$

in Cartesian coordinates. Therefore t is the synchronous time for the rest observers. The worldlines or the four velocities of the rest observers constitute an irrotational congruence. If we define the vorticity of this congruence or that of the vector field ξ^a by

$$\omega^a_\xi \equiv \frac{1}{\sqrt{-g}} \varepsilon^{abcd} \xi_b \xi_{c;d}; \quad \omega^a_u = e^{2\psi}\omega^a_\xi, \tag{6}$$

then

$$\omega^a_\xi = 0. \tag{7}$$

Also u^a is a geodesic.

The concept of the global rest frame is directly extended to a static spacetime like the Schwarzschild. The four velocity u^a defined as in equation (4), is hypersurface orthogonal since,

$$\xi_a = g_{00} t_{,a} = (\xi_b \xi^b) t_{,a}. \tag{8}$$

Once again t is the common synchronous time for the rest observers. Obviously, the vorticity

$$\omega_\xi = 0 = e^{-2\psi}\omega_u \tag{9}$$

and the four velocities form an irrotational congruence. The rest observer's four velocity, however, is no longer geodetic.

Let us now consider the Kerr spacetime. The timelike Killing vector field is no longer irrotational and hence the Killing observers following ξ no longer define the global rest frame. Nevertheless, consider the vector field

$$\chi^a = \xi^a - \frac{(\xi^b \eta_b)}{(\eta^c \eta_c)} \eta^a. \tag{10}$$

We notice,

$$\chi^a \eta_a = 0 \tag{11}$$

so that χ^a is the projection of ξ^a orthogonal to η^a. Furthermore, it is easy to show that the vorticity of the χ^α - congruence

$$\omega^a_\chi = 0. \tag{12}$$

This was first noticed by Bardeen (1970), who called the frames adopted to χ^α as locally nonrotating frames (LNRF). It was recognized that the physical phenomena in the Kerr spacetime could be studied in a significant manner when referred to LNRF. The observers with four velocity

$$u^a = (\chi^b \chi_b)^{-\frac{1}{2}} \chi^a \tag{13}$$

are in fact the 'rest' observers and the frames adapted to them form the global rest frame since χ_a is in fact hypersurface orthogonal:

$$\chi_a = \left[\xi^b \xi_b - \frac{(\xi^b \eta_b)}{(\eta^c \eta_c)}\right] t_{,a}. \tag{14}$$

As before t is the synchronous time for these observers. The apparently paradoxical situation is that in order to be 'rest' observers, those following χ^α will have to be revolving round the black hole! Properties of the global rest frames were studied in detail and generalized to arbitrary stationary, axisymmetric spacetimes by Greene, Schucking & Vishveshwara (1975).

They showed that if the Killing fields ξ^α and η^α satisfied orthogonal transitivity, as in the Kerr spacetime, χ^α became null on the event horizon similar to ξ^α in the Schwarzschild spacetime. Furthermore, t = constant can be shown to be maximal surfaces.

Physical phenomena can be studied meaningfully in the global rest frames, especially since extended systems can be defined only on spatial surfaces of simultaneity like t = constant.

4. Rotational effects

There are three important approaches related to the study of rotational effects in time-independent, axially symmetric spactimes such as those of black holes. These are gravi-electromagnetism, gyroscopic precession and the general relativistic analogues of inertial forces. They manifest themselves especially when considering particle trajectories. They can be studied elegantly when the trajectories follow Killing vector fields as in the case of stationary worldlines or circular orbits. Moreover, the three approaches can be synthesised in a very nice way for such trajectories. We shall present below the formalism in the general case of the stationary, axisymmetric spacetimes and specialize to black holes. These considerations are based on the paper by Nayak & Vishveshwara (1998).

Recently, the two general relativistic phenomena, namely gyroscopic precession and inertial forces have been studied in detail. Iyer & Vishveshwara (1993) have given a comprehensive treatment of gyroscopic precession in axially symmetric stationary spacetimes making use of the elegant Frenet-Serret (FS) formalism. This forms the basis for a covariant description of gyroscopic precession. At the same time, a general formalism defining inertial forces in general relativity has been presented by Abramowicz, Nurowski & Wex (1993). The motivation for this work stemmed from the earlier interest in centrifugal force and its reversal. Such reversal in the Schwarzschild spacetime at the circular photon orbit was first discussed by Abramowicz and Prasanna (1990) and later in the case of the Ernst spacetime by Prasanna (1991). Abramowicz (1990) showed that centrifugal force reversed at the photon orbit in all static spacetimes. He argued, on qualitative grounds, that gyroscopic precession should also reverse at the photon orbit. Taking the Ernst spacetime as a specific example of static spacetimes Nayak & Vishveshwara (1997) have shown that, in fact, both centrifugal force and gyroscopic precession reverse at the photon orbits. A similar study by Nayak & Vishveshwara (1996) in the Kerr-Newman spacetime indicates that the situation in the case of stationary spacetimes is much more complicated than in

the case of static spacetimes. Neither centrifugal force nor gyroscopic precession reverses at the photon orbit.

The above studies raise some interesting questions. Is gyroscopic precession directly related to centrifugal force in all static spacetimes? If so, do they both necessarily reverse at the photon orbit? In the case of stationary spacetimes is it possible to make a covariant connection between the gyroscopic precession on the one hand and the inertial forces on the other, not necessarily just the centrifugal force? Does such a connecting formula reveal the individual non-reversal of gyroscopic precession and centrifugal force at the photon orbit? We shall consider these and related questions. We shall then take up gravi-electromagnetism and show how this is related to gyroscopic precession and inertial forces. The case of black holes becomes a specific example of this broad-based formalism.

4.1 *Gyroscopic precession*

4.1.1 *Frenet-Serret description of gyroscopic precession*

The Frenet-Serret (FS) formalism offers a covariant method of treating gyroscopic precession. It turns out to be quite a convenient and elegant description of the phenomenon when the worldlines along which the gyroscopes are transported follow spacetime symmetry directions or Killing vector fields. In fact, in most cases of interest orbits corresponding to such worldlines are considered for simplicity. In the FS formalism the worldlines are characterized in an invariant geometric manner by defining along the curve three parameters κ the curvature and the two torsions τ_1 and τ_2 and an orthonormal tetrad. As we shall see, the torsions τ_1 and τ_2 are directly related to gyroscopic precession. All the above quantities can be expressed in terms of the Killing vectors and their derivatives. These considerations apply to a single trajectory in any specific example. However, additional geometric insight may be gained by identifying the trajectory as a member of one or more congruences generated by combining different Killing vectors. For this purpose the FS formalism is generalized to what may be termed as quasi-Killing trajectories. For the sake of completeness we summarize below relevant formulae taken from Iyer & Vishveshwara (1993).

Let us consider a spacetime that admits a timelike Killing vector ξ^a and a set of spacelike Killing vector $\eta_{(A)}$ (A = 1,2,...m).Then a quasi-Killing vector may be defined as

$$\chi^a \equiv \xi^a + \omega_{(A)}\eta^a_{(A)}, \tag{15}$$

where (A) is summed over. The Lie derivative of the functions $\omega_{(A)}$ with respect to χ^a is assumed to vanish,

$$\mathcal{L}_\chi \omega_{(A)} = 0. \tag{16}$$

We adopt the convention that Latin indices $a,b,...$ = 0 – 3 and Greek indices α $\beta, \ldots$ = 1–3 and the metric signature is (+, –, –, –). Geometrized units with $c = G = 1$ are chosen. A congruence of quasi-Killing trajectories is generated by the integral curves of χ^a. As a special case we obtain a Killing congruence when $\omega_{(A)}$ are constants.

Assuming χ^a to be timelike, we may define the four velocity of a particle following χ^a by

$$e^a_{(0)} \equiv u^a \equiv e^{\psi}\chi^a, \tag{17}$$

so that

$$e^{-2\psi} = \chi^a\chi_a, \quad \psi_{,a}\chi^a = 0 \tag{18}$$

and

$$\dot{e}^a_{(0)} \equiv e^a_{(0);b}e^b_{(0)} = F^a_b e^b_{(0)}, \tag{19}$$

where

$$F_{ab} \equiv e^{\psi}(\xi_{a;b} + \omega_{(A)}\eta_{(A)a;b}). \tag{20}$$

The derivative of $\omega_{(A)}$ drops out of the equation. The Killing equation and the equation $\xi_{a;b;c} \equiv R_{abcd}\xi^d$ satisfied by any Killing vector lead to

$$F_{ab} = -F_{ba} \quad \text{and} \quad \dot{F}_{ab} = 0. \tag{21}$$

Now, the FS equations in general are given by

$$\begin{aligned}
\dot{e}^a_{(0)} &= \kappa e^a_{(1)},\\
\dot{e}^a_{(1)} &= \kappa e^a_{(0)} + \tau_1 e^a_{(2)},\\
\dot{e}^a_{(2)} &= -\tau_1 e^a_{(1)} + \tau_2 e^a_{(3)},\\
\dot{e}^a_{(3)} &= -\tau_2 e^a_{(2)}.
\end{aligned} \tag{22}$$

As mentioned earlier k,τ_1 and τ_2 are respectively the curvature, and the first and second torsions while $e^a_{(i)}$ form an orthonormal tetrad. The six quantities describe the worldline completely. In the case of the quasi-Killing trajectories one can show that κ, τ_1 and τ_2 are constants and that each of $e^a_{(i)}$ satisfies a Lorentz like equation:

$$\dot{\kappa} = \dot{\tau}_1 = \dot{\tau}_2 = 0, \tag{23}$$

$$\dot{e}^a_{(i)} = F^a_b e^b_{(i)}. \tag{24}$$

Further, k, τ_1, τ_2 and $e^a{}_{(\alpha)}$ can be expressed in terms of $e^a_{(0)}$ and $F^n_{ab} \equiv F^a{}_{a_1}F^{a_1}_{a_2} \ldots F_{a_{n-1}b}$.

$$\begin{aligned}
\kappa^2 &= F^2_{ab}e^a_{(0)}e^b_{(0)},\\
\tau_1^2 &= \kappa^2 - \frac{F^4_{ab}e^a_{(0)}e^b_{(0)}}{\kappa^2},\\
\tau_2^2 &= \frac{F^6_{ab}e^a_{(0)}e^b_{(0)}}{\kappa^2\tau_1^2} - \frac{\left(\kappa^2 - \tau_1^2\right)^2}{\tau_1^2},
\end{aligned} \tag{25}$$

$$\begin{aligned}
e^a_{(1)} &= \frac{1}{\kappa}F^a_b e^b_{(0)},\\
e^a_{(2)} &= \frac{1}{\kappa\tau_1}[F^{2a}_b - \kappa^2\delta^a_b]e^b_{(0)},\\
e^a_{(3)} &= \frac{1}{\kappa\tau_1\tau_2}[F^{3a}_b + (\tau_1^2 - \kappa^2)F^a_b]e^b_{(0)}.
\end{aligned} \tag{26}$$

The above equations were first derived by Honig, Schiicking & Vishveshwara (1974) to describe charged particle motion in a homogeneous electromagnetic field. Interestingly, they are identical to those that arise in the case of quasi-Killing trajectories. Next let us consider an inertial frame of tetrad ($e^a_{(0)}, f^a_{(\alpha)}$) which undergoes Fermi-Walker (FW) transport along the worldline. The triad $f_{(\alpha)}$ may be physically realized by a set of three mutually orthogonal gyroscopes. Then, the angular velocity of the FS triad $e^a_{(\alpha)}$ with respect to the FW triad $f^a_{(\alpha)}$ is given by Iyer & Vishveshwara (1993)

$$\omega^a_{\mathrm{FS}} = \tau_2 e^a_{(1)} + \tau_1 e^a_{(3)}. \tag{27}$$

Or the gyroscopes precess with respect to the FS frame at a rate given by $\Omega(g) = -\omega_{\mathrm{FS}}$ In case of the Killing congruence ω_{FS} is identical to the vorticity of the congruence.

4.1.2 *Axially symmetric stationary spacetimes*

An axially symmetric stationary metric admits a timelike Killing vector ξ^α and a spacelike Killing vector η^a with closed circular orbits around the axis of symmetry. Assuming orthogonal transitivity, in coordinates ($x^0 \equiv t$, $x^3 \equiv \phi$) adapted to ξ^α and η^a respectively the metric takes on its canonical form

$$\mathrm{d}s^2 = g_{00}\,\mathrm{d}t^2 + 2g_{03}\,\mathrm{d}t\,\mathrm{d}\phi + g_{33}\,\mathrm{d}\phi^2 + g_{11}\,\mathrm{d}r^2 + g_{22}\,\mathrm{d}\theta^2 \tag{28}$$

with g_{ab} functions of $x^1 \equiv r$ and $x^2 \equiv \theta$ only. The quasi-Killing vector field

$$\chi^a = \xi^a + \omega\eta^a \tag{29}$$

generates closed circular orbits around the symmetry axis with constant angular speed w along each orbit. The FS parameters and the tetrad can be determined either by the direct substitution of χ^a or by transforming to a rotating coordinate frame as discussed by Iyer & Vishveshwara (1993). They can be written in terms of the Killing vectors and their derivatives as follows.

$$\kappa^2 = -g^{ab}a_a a_b, \tag{30}$$

$$\tau_1^2 = [g^{ab}a_a d_b]^2, \tag{31}$$

$$\tau_2^2 = \left[\frac{\varepsilon^{abcd}}{\sqrt{-g}}\, n_a \tau_b a_c d_d\right]^2, \tag{32}$$

$$e^a_{(0)} = \frac{1}{\sqrt{\mathcal{A}}}(1,0,0,\omega),$$

$$e^a_{(1)} = -\frac{1}{\kappa}(0, g^{11}a_1, g^{22}a_2, 0),$$

$$e^a_{(2)} = \frac{1}{\sqrt{\mathcal{A}}\sqrt{-\Delta_3}}(\mathcal{B},0,0,-\mathcal{C}),$$

$$e^a_{(3)} = \frac{\sqrt{g^{11}g^{22}}}{\kappa}(0,-a_2,a_1,0). \tag{33}$$

In the above,

$$d_a = \left(\frac{\mathcal{B}}{2\sqrt{-\Delta_3}\kappa}\right)\left[\frac{\mathcal{B}_a}{\mathcal{B}} - \frac{\mathcal{A}_a}{\mathcal{A}}\right] = \left(\frac{\mathcal{B}}{\sqrt{-\Delta_3}\kappa}\right)[b_a - a_a],$$

$$
\begin{aligned}
a_a &= \frac{\mathcal{A}_a}{2\mathcal{A}}, \\
b_a &= \frac{\mathcal{B}_a}{2\mathcal{B}}, \\
\mathcal{A} &= (\xi^a\xi_a) + 2\omega(\eta^a\xi_a) + \omega^2(\eta^a\eta_a), \\
\mathcal{B} &= (\eta^a\xi_a) + \omega(\eta^a\eta_a), \\
\mathcal{C} &= (\xi^a\xi_a) + \omega(\eta^a\xi_a), \\
\mathcal{A}_a &= (\xi^b\xi_b)_{,a} + 2\omega(\eta^b\xi_b)_{,a} + \omega^2(\eta^b\eta_b)_{,a}; \quad a = 1,2, \\
\mathcal{B}_b &= (\eta^a\xi_a)_{,b} + \omega(\eta^a\eta_a)_{,b}; \quad b = 1,2, \\
\Delta_3 &= (\xi^a\xi_a)(\eta^b\eta_b) - (\eta^a\xi_a)^2
\end{aligned}
\tag{34}
$$

where n^a is the unit vector along $\zeta_a = \xi_a - (\xi^b\eta_b)/(\eta^c\eta_c)\,\eta_a$ and τ^i is the unit vector along the rotational killing vector η^a. We may note that all the above equations can be specialized to a static spacetime by setting $\xi^a\eta_a = 0$ and $\zeta^a \equiv \xi^a$

The above expressions when specialized to the equatiorial planes of black hole spacetimes are as follows. We have $\tau_2 = 0$ so that gyroscopic precession is given by τ_1 alone.

Kerr:

$$
\tau_1^2 = \frac{1}{r^2}\,\frac{\left[\frac{Ma}{r^2} - \left(\frac{(r^2+2a^2)M}{r^2} - r\left(1-\frac{2M}{r}\right)\right)\omega + \frac{Ma(3r^2+a^2)\omega^2}{r^2}\right]^2}{\left[1 - (r^2+a^2)\omega^2 - \frac{2M(a\omega)^2}{r}\right]^2}. \tag{35}
$$

Schwarzschild:

$$
\tau_1^2 = \omega^2\,\frac{\left(1-\frac{3M}{r}\right)^2}{\left(1-\frac{2M}{r} - r^2\omega^2\right)^2}. \tag{36}
$$

4.2 *Inertial forces*

4.2.1 *General formalism*

As has been mentioned earlier, in a recent paper Abramowicz *et al.* (1993) have formulated the general relativistic analogues of inertial forces in an arbitrary spacetime. The particle four velocity u^a is decomposed as

$$
u^a = \gamma(n^a + v\tau^a). \tag{37}
$$

In the above, n^a is a globally hypersurface orthogonal timelike unit vector, τ^a is the unit vector orthogonal to it along which the spatial three velocity v of the particle is aligned and γ is the normalization factor that makes $u^a u_a = 1$.

Then the forces acting on the particle are written down as:

Gravitational force $G_k = \phi_{,k}$,

Centrifugal force $Z_k = -(\gamma v)^2 \tilde{\tau}^i \tilde{\nabla}_i \tilde{\tau}_k$,

$$\text{Euler force} \quad E_k = -\dot{V}\tilde{\tau}_k,$$
$$\text{Coriolis-Lense-Thirring force} \quad C_k = \gamma^2 v X_k, \tag{38}$$

where

$$\dot{V} = (ve^{\phi}\gamma)_{,i}u^i,$$
$$X_k = n^i(\tau_{k;i} - \tau_{i;k}),$$
$$\phi_{,k} = -n^i n_{k,i}. \tag{39}$$

Here $\tilde{\tau}^i$ is the unit vector along τ^i in the conformal space orthogonal to n^i with the metric

$$\tilde{h}_{ik} = e^{-2\phi}(g_{ik} - n_i n_k). \tag{40}$$

One can show that the covariant derivatives in the two spaces are related by

$$\tilde{\tau}^i \tilde{\nabla}_i \tilde{\tau}_k = \tau^i \nabla_i \tau_k - \tau^i \tau_k \nabla_i \phi - \nabla_k \phi. \tag{41}$$

We shall now apply this formalism to axially symmetric stationary Spacetimes.

4.2.2 *Inertial forces in axially symmetric stationary spacetimes*

As has been shown by Greene, Schiicking & Vishveshwara (1975), axially symmetric stationary spacetimes with orthogonal transitivity admit a globally hypersurface orthogonal timelike vector field

$$\zeta^a = \xi^a + \omega_0 \eta^a, \tag{42}$$

where the fundamental angular speed of the irrotational congruence is

$$\omega_0 = -(\xi^a \eta_a)/(\eta^b \eta_b). \tag{43}$$

The unit vector along ζ^a is identified with n. Further, if u^a follows a quasi Killing circular trajectory, then τ^i is along the rotational Killing vector η^a. In this case it is easy to show that $\dot{V} = 0$ and hence the Euler force does not exist.

More specifically,

$$u^a = e^{\psi}(\xi^a + \omega\eta^a) = e^{\psi}\chi^a = \gamma(n^a + v\tau^a). \tag{44}$$

Then we have

$$n^a = e^{-\phi}\zeta^a,$$
$$\tau^a = e^{-\alpha}\eta^a,$$
$$\gamma = e^{\psi+\phi},$$
$$v = e^{-\phi+\alpha}(\omega - \omega_0), \tag{45}$$

where

$$\phi = \tfrac{1}{2}\ln(\zeta^a\zeta_a), \alpha = \tfrac{1}{2}\ln(-\eta^a\eta_a), \psi = \tfrac{1}{2}\ln(\chi^a\chi_a). \tag{46}$$

From the above relations, we can write down the inertial forces from their definitions as follows.

Gravitational force

$$G_k = \phi_{,k}, \tag{47}$$

Centrifugal force

$$Z_k = \frac{1}{2} e^{2(\psi+\phi)} \tilde{\omega}^2 \left(\frac{\eta^a \eta_a}{\zeta^b \zeta_b} \right)_{,k}, \tag{48}$$

Coriolis-Lense-Thirring force

$$C_k = e^{2(\psi+\alpha)} \tilde{\omega} \left(\frac{\xi^a \eta_a}{\eta^b \eta_b} \right)_{,k}, \tag{49}$$

Where $\tilde{\omega} = (\omega - \omega_0)$.
On the equatorial plane of the Kerr spacetime they reduce to

$$G_k = \frac{(r-M)\{(r^2+a^2)r + 2Ma^2\} - \frac{\Delta}{r}\{r^3 - Ma^2\}}{\Delta\{(r^2+a^2)r + 2Ma^2\}} (0,1,0,0), \tag{50}$$

$$C_k = \frac{2a\mathcal{W}}{\mathcal{A}\mathcal{G}_3} \left\{ \frac{M}{r^2} a^2 + 3M \right\} (0,1,0,0), \tag{51}$$

$$Z_k = \frac{\mathcal{W}^2}{\mathcal{A}} (0, z_1, 0, 0), \tag{52}$$

where

$$\begin{aligned}
\Delta &\equiv r^2 + a^2 - 2Mr, \\
\mathcal{A} &= 1 - \omega^2(r^2+a^2) - \frac{2M}{r}(1-\omega a)^2, \\
\mathcal{G}_3 &= (r^2+a^2) + \frac{2M}{r} a^2, \\
\mathcal{W} &= \omega - \frac{2Ma}{(r^2+a^2)r + 2Ma^2}, \\
z_1 &= \frac{1}{\Delta r^2}[(r-M)\{(r^2+a^2)r^2 + 2Mra^2\} - 2\Delta\{r^3 - Ma^2\}].
\end{aligned} \tag{53}$$

4.2.3 *Specialization to static spacetimes*

In a static spacetime the global timelike Killing vector ξ^α itself is hypersurface orthogonal. The unit vector n^a is now aligned along ξ^α,

$$n^a = e^{-\phi} \xi^a. \tag{54}$$

Then we have the inertial forces as follows:

Gravitational force

$$G_k = \phi_{,k}, \tag{55}$$

where $\phi = \frac{1}{2} \ln (\xi^a \xi^a)$,

Centrifugal force

$$Z_k = -\frac{\omega^2}{2} e^{2(\psi+\alpha)} \left[\ln\left(\frac{\eta^i \eta_i}{\xi^j \xi_j}\right) \right]_{,k}, \tag{56}$$

Coriolis-Lense-Thirring force is identically zero,

$$C_k = 0. \tag{57}$$

In the specific example of the Schwarschild spacetime we have:

$$G_k = \left(1 - \frac{2M}{r}\right)^{-1} \frac{M}{r^2} (0, 1, 0, 0), \tag{58}$$

$$Z_k = \frac{(r - 3M)}{\left(1 - \frac{2M}{r} - \omega^2 r^2\right)\left(1 - \frac{2M}{r}\right)} (0, 1, 0, 0). \tag{59}$$

4.3 *Covariant connections*

In the preceding section we have derived expressions for τ_1 and τ_2 which give gyroscopic precession rate in terms of the Killing vectors. Similarly, inertial forces in an arbitrary axisymmetric stationary spacetime have also been written down in terms of the Killing vectors. All these quantities have been defined in a completely covariant manner. We shall now proceed to establish covariant connections between gyroscopic precession, i.e. the FS torsions τ_1 and τ_2, on the one hand and the inertial forces on the other. First, we shall consider the simpler case of static Spacetimes.

4.3.1 *Static spacetimes*

We have derived in equation (31) and (32), the FS torsions τ_1 and τ_2 for a stationary spacetime. As has been mentioned earlier, for a static spacetime $\xi^a \eta_a = 0$ and $\zeta^\alpha = \xi^\alpha$ in the above equations as well as in the expressions for inertial forces. With this specialization; centrifugal force can be written from equation (56) as

$$Z_b = e^{-(\phi-\alpha)} \omega \kappa d_b. \tag{60}$$

Substituting equation (60) in equations (31) and (32) we arrive at the relations

$$\tau_1^2 = \frac{\beta^2}{\omega^2} [a^b Z_b]^2, \tag{61}$$

and

$$\tau_2^2 = \frac{\beta^2}{\omega^2} \left[\frac{\varepsilon^{abcd}}{\sqrt{-g}} n_a \tau_b a_c Z_d \right]^2, \tag{62}$$

where

$$\beta = \frac{e^{(\phi-\alpha)}}{\kappa}. \tag{63}$$

The equations above relate gyroscopic precession directly to the centrifugal force The two torsions τ_1 and τ_2, equivalent to the two components of precession, are

respectively proportional to the scalar and cross products of acceleration and the centrifugal force. We shall discuss the consequences of these relations later on.

4.3.2 *Stationary spacetimes*

From equation (34) we have

$$\mathcal{A} = (\xi^a \xi_a) + 2\omega(\eta^a \xi_a) + \omega^2\ (\eta^a \eta_a),$$
$$\mathcal{B} = (\eta^a \xi_a) + \omega(\eta^a \eta_a).$$

We decompose the angular speed ω with reference to the fundamental angular speed of the irrortational congruence $\omega_0 = -(\xi^a\ \eta_a)/(\eta^a \eta_a)$,

$$\omega = \tilde{\omega} + \omega_0. \tag{64}$$

Then we have

$$\mathcal{A} = \zeta^a \zeta_a + \tilde{\omega}^2 \eta^a \eta_a,$$
$$\mathcal{B} = \tilde{\omega} \eta^a \eta_a. \tag{65}$$

Similarly, we get

$$\mathcal{A}_a = (\zeta^b \zeta_b)_{,a} + 2\tilde{\omega}\mathcal{C}_a + \tilde{\omega}^2 (\eta^b \eta_b)_{,a},$$
$$\mathcal{B}_a = \mathcal{C}_a + \tilde{\omega}(\eta^b \eta_b)_{,a}, \tag{66}$$

where

$$\mathcal{C}_a \equiv (\xi^b \eta_b)_{,a} + \omega_0 (\eta^b \eta_b)_{,a} \tag{67}$$

or equivalently

$$\mathcal{C}_a = -(\xi^b \eta_b)\omega_{0,a}. \tag{68}$$

From equations (34), (65) and (66) we can show

$$\mathrm{d}_a = -e^{2\psi} \frac{e^{-(\phi+\alpha)}\tilde{\omega}}{2\kappa} \{(\zeta^p \zeta_p)\mathcal{C}_a + \tilde{\omega}[(\zeta^p \zeta_p)(\eta^q \eta_q)_{,a} - (\eta^p \eta_p)(\zeta^q \zeta_q)_{,a}] - \tilde{\omega}^2 (\eta^p \eta_p)\mathcal{C}_a\}. \tag{69}$$

Further, it is easy to see that C_a is directly proportional to $\mathcal{C}_a$,

$$\mathcal{C}_a = -e^{-2\psi}\tilde{\omega}^{-1} C_a \tag{70}$$

where C_a is the Coriolis-Lense-Thirring force. Then equation (69) takes on the form where Z_a is the centrifugal force.

$$\mathrm{d}_a = \frac{e^{(\phi-\alpha)}}{\tilde{\omega}\kappa} \left\{ Z_a - \frac{1}{2}[1 + \tilde{\omega}^2 e^{2(\alpha-\phi)}] C_a \right\} \tag{71}$$

where Z_a is the centrifugal force.

Substituting this in equation (31) for τ_1^2 we get the relation,

$$\tau_1^2 = \frac{\beta^2}{\tilde{\omega}^2} [g^{ab} a_a (Z_b + \beta_1 C_a)]^2, \tag{72}$$

where

$$\beta = \frac{e^{(\phi-\alpha)}}{\kappa},$$
$$\beta_1 = -\tfrac{1}{2}[1 + \tilde{\omega}^2 e^{2(\alpha-\phi)}]. \tag{73}$$

Again, from equation (32), we obtain the expression

$$\tau_2^2 = \frac{\beta^2}{\tilde{\omega}^2}\left[\frac{\varepsilon^{abcd}}{\sqrt{-g}} n_a \tau_b a_c (Z_d + \beta_1 C_d)\right]^2. \tag{74}$$

These relations are more complicated than those we have derived in the static case Nevertheless, they closely resemble the latter with the centrifugal force replaced by the combination of centrifugal and Coriolis forces ($Z_a + \beta_1 C_a$). The static case formulae are obtained from those of stationary case by setting the Coriolis force to zero.

A formula for gyroscopic precession in the case of circular orbits in axially symmetric stationary spacetimes was derived by Abramowicz, Nurowski & Wex (1995) within a different framework. We note that gyroscopics precession does not involve the gravitational force. In case of geodetic orbits, total force is zero but not the centrifugal and Coriolis force individually. Therefore gyroscopic precession is also nonzero even for geodetic orbits.

4.4 *Reversal of gyroscopic precession and inertial forces*

The condition for the reversal of gyroscopic precession is given by

$$\omega^a_{\mathrm{FS}} = \tau_1 e^a_{(3)} + \tau_2 e^a_{(1)} = 0. \tag{75}$$

Since $e^a_{(1)}$ and $e^a_{(3)}$ are linearly independent vector fields at each point, this condition is the same as requiring

$$\tau_1 = \tau_2 = 0. \tag{76}$$

In the case of static spacetimes, τ_1 ans τ_2 are directly related to the centrifugal force Z_k. Therefore gyroscopic precession and centrifugal force reverse simultaneously. It can be shown that this happens at a photon orbit as borne out by the Schwarschild spacetime. In the case of stationary spacetimes there is no such correlations. This is true in the case of the Kerr spacetime.

5. Gravi-electric and Gravi-magnetic fields

Gravi-electric and gravi-magnetic fields are closely related to the idea of inertial forces. These fields with respect to observers following the integral curves of n^a can be defined as follows.

Gravi-electric field:

$$E^a = F^{ab} n_b, \tag{77}$$

Gravi-magnetic field:

$$H^a = \tilde{F}^{ab} n_b \tag{78}$$

where $\tilde{F}^{ab}$ is the dual of F^{ab},

$$\tilde{F}^{ab} = \tfrac{1}{2}(\sqrt{-g})^{-1} \varepsilon^{abcd} F_{cd}. \tag{79}$$

In the above, as before, $F^{ab} = e^{\psi}(\xi_{a;b} + \omega\eta_{a;b})$The equation of motion is

$$\dot{u}^a = F^{ab} u_b. \tag{80}$$

Projecting onto the space orthogonal to n^a with $h_{ab} = g_{ab} - \eta_a \eta_b$ and decomposing u_a as given in (44), we get

$$\dot{u}_{\perp a} = \gamma[F_{ac} n^c + (v(F_{ac}\tau^c - n_a F_{bc} n^b \tau^c)] \tag{81}$$

where 7 is the normalization factor. This equation can be written in the form

$$\dot{u}_{\perp a} = \gamma[F_{ac} n^c + v\sqrt{-g}\varepsilon_{abcd} n^b \tau^c H^d] \tag{82}$$

or

$$\dot{u}_{\perp a} = \gamma[E + v \times H]. \tag{83}$$

We can therefore define

Gravi-electric force:

$$f_{GEa} = \gamma F_{ac} n^c, \tag{84}$$

Gravi-magnetic force:

$$f_{GHa} = \gamma v\sqrt{-g}\varepsilon_{abcd} n^b \tau^c H^d = \gamma v(F_{bc}\tau^c - n_a F_{bc} n^b \tau^c). \tag{85}$$

5.1 *Relations among gravi-electric, gravi-magnetic and inertial forces*

5.1.1 *Static case*

We have defined the gravi-electric field E_a by

$$yE_a = yF_{ac}n^c$$

If we substitute for $F_{ab} = e^{(\psi}(\xi_{a;b} + \omega\eta_{a;b})$, we get

$$f_{GEa} = \gamma E_a = \gamma F_{ac} n^c = -e^{2(\psi+\phi)} G_a. \tag{86}$$

So,

$$E_a = -e^{(\psi+\phi)} G_a. \tag{87}$$

Here G_a is the gravitational force. Similarly we have for the gravi-magnetic field

$$f_{GHa} = \gamma v(F_{ac}\tau^c - n_a n^b F_{bc}\tau^c).$$

The second term in this equation is identically zero because the Killing vector field ξ^a and η^a commute and we get

$$\begin{aligned} f_{GHa} &= \gamma v \sqrt{-g}\varepsilon_{abcd} n^b \tau^c H^d, \\ &= \gamma v F_{ac} \tau^c, \\ &= [e^{2(\psi+\alpha)}\omega^2 G_a - Z_a]. \end{aligned} \tag{88}$$

The above relation clearly shows the connection between the gravi-magnetic force on the one hand and the gravitational and centrifugal forces on the other.

5.1.2 *Stationary case*

In the stationary case, n^a is given by equation (45). As before we decompose $\omega = \varpi + \omega_0$ where ω_0 is given by (43). Then a straightforward computation gives the expression for the gravi-electric field.

$$E_a = -e^{(\psi+\phi)} G_a + e^{-(\psi+\phi)} C_a \tag{89}$$

and the gravi-electric force,

$$f_{GEa} = \gamma E_a = -e^{2(\psi+\phi)} G_a + C_a. \tag{90}$$

This shows the relation of gravi-electric field or force to both gravitational and centrifugal forces. In the stationary case also we have

$$n_a n^b F_{bc} \tau^c \equiv 0. \tag{91}$$

Then it follows

$$\begin{aligned} f_{GHa} &\equiv \gamma v \sqrt{-g}\varepsilon_{abcd} n^d \tau^c H^d, \\ &= \gamma v F_{ac} \tau^c, \\ &= \left[\frac{C_a}{2} + e^{2(\psi+\alpha)}\tilde{\omega}^2 G_a - Z_a\right]. \end{aligned} \tag{92}$$

Hence gravi-magnetic force is related to all the three inertial forces—gravitational, centrifugal and Coriolis.

5.2 *Gravi-electric and Gravi-magnetic fields with respect to comoving frame*

In the previous section we have defined gravi-electric and gravi-magnetic fields with respect to the irrotational congruence. Similarly these fields can be defined with respect to the four velocity u^a of the particle as follows.

Gravi-electric field:

$$\tilde{E}^a = F^{ab} u_b, \tag{93}$$

Gravi-magnetic field:

$$\tilde{H}^a = \tilde{F}^{ab} u_b. \tag{94}$$

Where $\tilde{F}^{ab}$ is dual to F^{ab} as before. The equation of motion takes the form

$$a^a = \tilde{E}^a. \tag{95}$$

Precession frequency can be written simply as

$$\omega^a = \tilde{H}^a. \tag{96}$$

Following Honig, Schücking & Vishveshwara (1974), Frenet-Serret parameters κ,τ_1 and τ_2 can be expressed in terms of gravi-electric and gravi-magnetic fields.

$$\kappa = |\tilde{E}| \tag{97}$$

where

$$|\tilde{E}| = \sqrt{-\tilde{E}^a\tilde{E}_a}, \tag{98}$$

$$\tau_1 = \frac{|\tilde{P}|}{|\tilde{E}|}, \tag{99}$$

where

$$\tilde{P}^a = \varepsilon^{abcd}\tilde{E}_b\tilde{H}_a u_d = \tilde{E}\times\tilde{H}, \tag{100}$$

$$|\tilde{P}| = \sqrt{-\tilde{P}^a\tilde{P}_a} \tag{101}$$

and

$$\tau_2 = -\frac{\tilde{H}^a\tilde{E}_a}{|\tilde{E}|}. \tag{102}$$

Frenet-Serret tetrad components can also be expressed in terms of

$$\begin{aligned} e^a_{(1)} &= \frac{\tilde{E}^a}{|\tilde{E}|}, \\ e^a_{(2)} &= \frac{\tilde{P}^a}{|\tilde{P}|}, \\ e^a_{(3)} &= \frac{\varepsilon^{abcd}\tilde{E}_b\tilde{P}_c u_d}{\tilde{P}^r\tilde{E}_r}. \end{aligned} \tag{103}$$

In reference (Honig, Schücking & Vishveshwara 1974), these expressions had been derived for charged particle motion in a constant electromagnetic field. We have now demonstrated the exact analogues in the case of gravi-electric and gravi-magnetic fields. The one-to-one correspondence is indeed remarkable.

All this can be translated easily to the specific example of black holes since the required expressions have been given already.

6. Conclusion

The geometric structure and the physical phenomena associated with black holes offer a striking example of the general relativistic effects engendered by strong gravitational fields. Furthermore, rotation plays a pivotal role in distinguishing the properties of the Kerr black hole from those of the Schwarzschild black hole. In comparing and contrasting their properties and the consequent effects, the Killing fields admitted by

the two spacetimes provide an elegant, simple and yet a powerful basis for detailed analysis. This is utilized in defining fundamental concepts and formalisms as in the definition of the global rest frame. Again, the Killing symmetries provide a covariant method for treating gravi-electromagnetism, gyroscopic precession and inertial forces. They are interrelated and can be synthesized in an appealing manner. There are many other topics in black hole physics that carry the stamp of rotation: radiation, thermodynamics, Mach's principle, astrophysical applications such as accretion and so on. All this is way beyond the scope of the present article.

References

Abramowicz, M. A., Nurowski, P., Wex, N. 1995, *Class Quantum. Grav.,* **12**, 1467.
Abramowicz, M.A., Nurowski,P., Wex,N.1993, *Class Quantum. Grav.,* **10**, L183.
Abramowicz, M. A., Prasanna, A. R. 1990, *Mon. Not. R. Astr. Soc,* **245**, 720.
Abramowicz, M. A. 1990, *Mon. Not. R. Astr. Soc,* **245**, 733.
Bardeen, J. M.1970, *Astrophys. J,* **162**, 71
Greene, R. D., Schucking, E. L., Vishveshwara, C. V. 1975, *J. Math. Phys.,* **16**, 153
Honig, E., Schucking, E. L., Vishveshwara, C. V. 1974, *J. Math. Phys.,* **15**, 774
Iyer, B. R., Vishveshwara, C V. 1993, *Phys. Rev.,* **D48**, 5706.
Prasanna, A. R. 1991, *Phys. Rev.,* **D43**, 1418
Rajesh Nayak, K., Vishveshwara, C. V. 1998, *GRG,* **30**, 593
Rajesh Nayak, K., Vishveshwara, C. V. 1997, *GRG,* **29**, 291
Rajesh Nayak, K., Vishveshwara, C. V. 1996, *Class Quantum. Grav.,* **13**, 1173.

J. Astrophys. Astr. (1999) **20**, 121–129

Plenty of Nothing: Black Hole Entropy in Induced Gravity

V. P. Frolov, *Theoretical Physics Institute*, *Department of Physics*, *University of Alberta*, *Edmonton, Canada T6G 2J1*
email: frolov @phys. ualberta. ca

D. V. Fursaev, *Theoretical Physics Institute*, *Department of Physics, University of Alberta*, *Edmonton, Canada T6G 2J1* and *Joint Institute for Nuclear Research, Bogoliubov Laboratory of Theoretical Physics*, *141 980 Dubna*, *Russia*
e-mail: fursaev@cv.jinr. dubna. su

Abstract. We demonstrate how Sakharov's idea of induced gravity allows one to explain the statistical-mechanical origin of the entropy of a black hole. According to this idea, gravity becomes dynamical as the result of quantum effects in the system of heavy constituents of the underlying theory.The black hole entropy is related to the properties of the vacuum in the induced gravity in the presence of the horizon. We obtain the Bekenstein-Hawking entropy by direct counting the states of the constituents.

Key words. Black hole—gravity—entropy.

1. Black hole entropy problem

There are physical phenomena that allow a simple description but require tremendous efforts to explain them. Without doubts the origin of black hole entropy is such a problem. A black hole of mass M radiates as a heated body (Hawking 1975) with temperature $T_H = (8\pi GM)_{-1}$ and has entropy (Bekenstein 1972):

$$S^{BH} = \frac{\mathcal{A}}{4G}, \tag{1}$$

where A is the surface area of the black hole ($A=16\pi G^2M^2$), G is Newton's constant, and $c= \hbar=1$. Statistical mechanics relates entropy to the measure of disorder and qualitatively it is the logarithm of the number of microscopically different states available for given values of the macroscopical parameters. Are there internal degrees of freedom that are responsible for the Bekenstein-Hawking entropy S^{BH} ? This is the question that physicists were trying to answer for almost 25 years.

What makes the problem of black hole entropy so intriguing? Before discussing more technical aspects let us make simple estimations. Consider for example a supermassive black hole of 10^9 solar mass. According to equation (1), its entropy is about 10^{95}. This is seven orders of magnitude larger than the entropy of the other matter in the visible part of the Universe. What makes things even more complicated, a black hole is simply an empty space-time with a strong gravitational field and… nothing more. Really, a black hole is "plenty of nothing", or if we put this in a more physical way, the phenomenon we are dealing with is a vacuum in the strong

gravitational field. This conclusion leaves us practically no other choice but to try to relate the entropy of the black hole to properties of the physical vacuum in the strong gravitational field.

The black hole entropy is of the same order of magnitude as the logarithm of the number of different ways to distribute two signs + and – over the cells of Planckian size on the horizon surface. This estimation suggests that a reasonable microscopical explanation of the Bekenstein-Hawking entropy must be based on the quantum gravity, this Holy Grail of the theoretical physics. The superstring theory is the best what we have and what is often considered as the modern version of quantum gravity. Recent observation that the Bekenstein-Hawking entropy can be obtained by counting of string (D-brane) states is a very interesting and important result.

Still there are questions. The string calculations essentially use supersymmetry and are mainly restricted to extreme and near-extreme black holes. Moreover each model requires new calculations. And the last but not the least it remains unclear why the entropy of a black hole is universal and does not depend on the details of the theory at Planckian scales. Note that the black hole thermodynamics follows from the low-energy gravitational theory. That is why one can expect that only a few fundamental properties of quantum gravity but not its concrete details are really important for the statistical-mechanical explanation of the black hole entropy.

2. Sakharov's induced gravity

In the string theory the low-energy gravity with finite Newton's constant arises as the collective phenomenon and is the result of quantum excitations of constituents (strings) of the underlying theory. There is a certain similarity between this mechanism and Sakharov's induced gravity (Sakharov 1968). The low-energy effective action $W[g]$ Φ the induced gravity is defined as a quantum average of the constituent fields propagating in a given external gravitational background g

$$\exp(-W[g]) = \int \mathcal{D}\Phi \exp(-I[\Phi, g]). \tag{2}$$

The Sakharov's basic assumption is that the gravity becomes dynamical only as the result of quantum effects of the constituent fields. The gravitons in this picture are analogous to the phonon field describing collective excitations of a crystal lattice in the low-temperature limit of the theory. Search for the statistical-mechanical origin of the black hole entropy in Sakharov's approach (Jacobson 1994) might help to understand the universality of S^{BH}. In this paper we describe the most important features of the mechanism of generation of the Bekenstein-Hawking entropy in the induced gravity.

Each particular constituent field in equation (2) gives a divergent contribution to the effective action W[g]. In the one loop approximation the divergent terms are local and of the zero order, linear and quadratic in curvature. In the induced gravity the constituents obey additional constraints, so that the divergences cancel each other. It is also assumed that some of the fields have masses comparable to the Planck mass and the constraints are chosen so that the induced cosmological constant vanishes. As a result the effective action W[g] is finite and in the low-energy limit has the form of

the Einstein-Hilbert action

$$W[g] = -\frac{1}{16\pi G}\left(\int_{\mathcal{M}} \mathrm{d}V\, R + 2\int_{\partial\mathcal{M}} \mathrm{d}v K\right) + \cdots, \tag{3}$$

where Newton's constant G is determined by the masses of the heavy constituents The dots in equation (3) indicate possible higher curvature corrections to $W[g_{\mu\nu},]$ which are suppressed by the power factors of m_i^{-2} when the curvature is small. The vacuum Einstein equations $\delta W/\delta g^{\mu\nu} = 0$ are equivalent to the requirement that the vacuum expectation values of the total stress-energy of the constituents vanishes

$$\langle \hat{T}_{\mu\nu} \rangle = 0. \tag{4}$$

The value of the Einstein-Hilbert action (equation 3) calculated on the Gibbons-Hawking instanton determines the classical free energy of the black hole, and hence gives the Bekenstein-Hawking entropy S^{BH}. Sakharov's equality (equation 2) allows one to rewrite the free energy as the Euclidean functional integral over constituent fields on the Gibbons-Hawking instanton with periodic (for bosons) and anti-periodic (for fermions) in the Euclidean time boundary conditions. In this picture constituents are thermally excited and the Bekenstein-Hawking entropy can be expressed in terms of the statistical-mechanical entropy

$$S^{SM} = -\mathrm{Tr}\ \hat{\rho} \ln \hat{\rho}, \quad \hat{\rho} = \frac{e^{-\hat{H}/T_H}}{\mathrm{Tr}\ e^{-\hat{H}/T_H}}. \tag{5}$$

Here the operator H is the total canonical Hamiltonian of all the constituents.

How can an empty space (vacuum) possess thermodynamical properties? The entropy S^{SM} arises as the result of the loss of information about states inside the black hole horizon (Bombelli *et al.* 1986; Frolov & Novikov 1993). In the induced gravity (entanglement) entropy (equation 5) is calculated for the "heavy" constituents. This solves the problems of earlier attempts to explain the Bekenstein-Hawking entropy as the entanglement entropy of physical ("light") fields.

3. Models

3.1 *Models with non-minimal coupling of scalar constituents*

Consider the model of induced gravity (Frolov, Fursayev & Zelnikov 1997; Frolov & Fursaev 1997) that consists of N_s scalar fields ϕ_i with masses $m_{s,i}$ and N_d of Dirac fermions ψ_j with masses $m_{d,j}$ Scalar fields can have non-minimal couplings and are described by actions

$$I_s[\phi_i, g] = -\frac{1}{2}\int \sqrt{-g}\, \mathrm{d}^4x(\phi_i^{,\mu}\phi_{i,\mu} + m_s^2\phi_i^2 + \xi_i R\phi_i^2). \tag{6}$$

The action for Dirac fermions has the following standard form

$$I_d[\psi_j] = \int \sqrt{-g}\, \mathrm{d}^4x\bar{\psi}_j(\gamma^\mu\nabla_\mu + m_{d,j})\psi_j. \tag{7}$$

The corresponding quantum effective action of the model is

$$W[g] = \sum_{i=1}^{N_s} W_s(m_{s,i}) + \sum_{j=1}^{N_d} W(m_{d,j}). \tag{8}$$

$W[g]$ is a functional of the metric $g_{\mu\nu}$ of the background spacetime. The scalar and spinor actions follow immediately from equations (6) and (7),

$$W_s(m_{s,i}) = \tfrac{1}{2} \log\det(-\nabla^2 + m_{s,i}^2 + \xi_i R), \tag{9}$$

$$W_d(m_{d,j}) = -\log\det(\gamma^\mu \nabla_\mu + m_{d,j}). \tag{10}$$

In general, the effective action (9) is ultraviolet divergent quantity. Certain constraints are to be imposed on the parameters of the constituents to eliminate the leading divergencies in $W[g]$. These conditions can be written down with the help of the following two functions

$$p(z) = \sum_{i=1}^{N_s} m_{s,i}^{2z} - 4\sum_{j=1}^{N_d} m_{d,j}^{2z}, \quad q(z) = \sum_{i=1}^{N_s} m_s^{2z}(1 - 6\xi_s) + 2\sum_{j=1}^{N_d} m_{d,j}^{2z} \tag{11}$$

constructed from the parameters of the constituents. Direct calculations show that the induced cosmological constant vanishes and the induced gravitational coupling constant G is finite if the following constraints are satisfied

$$p(0) = p(1) = p(2) = p'(2) = 0, \tag{12}$$

$$q(0) = q(1) = 0. \tag{13}$$

The presence of the non-minimally coupled constituents is important. In this case it is possible to satisfy the constraints on the parameters ms,i, md and ξ_i which guarantee the cancellation of the leading ultraviolet divergencies of the induced gravitational action $W[g]$, equation (2). The induced Newton's constant in this model is

$$\frac{1}{G} = \frac{1}{12\pi}\left(\sum_{i=1}^{N_s}(1 - 6\xi_i)m_{s,i}^2 \ln m_{s,i}^2 + 2\sum_{j=1}^{N_d} m_{d,j}^2 \ln m_{d,j}^2\right). \tag{14}$$

Conditions (12) can be easily satisfied if for example bosons and fermions enter in supersymmetric multiplets. In this case $N_s = 4N_d$ and there exist N_d multiplets. The masses of bosons and a fermion in each supermultiplet are equal one to another. Under these conditions $p(z) = 0$. If in addition we assume that the parameter of the non-minimal coupling ξ has the same value for scalar fields inside the multiplet, the equations (13) and (14) take the form

$$\sum_{j=1}^{N_d} x_j = 0, \quad \sum_{j=1}^{N_d} m_j^2 x_j = 0, \quad \sum_{j=1}^{N_d} m_j^2 \ln(m_j^2) x_j = \frac{3\pi}{G}. \tag{15}$$

Here $x_j = 3/2 - 6\xi_j$. This is a set of linear equations for ξ_j which evidently has solution for $N_d \geq 3$.

The relation between the Bekenstein-Hawking entropy S^{BH} and the statistical-mechanical entropy S^{SM} of the heavy constituents can be found explicitly and has the

form (Frolov, Fursaev & Zelnikov 1987):

$$S^{BH} \equiv \frac{\mathcal{A}}{4G} = S^{SM} - Q. \tag{16}$$

The important property of S^{SM} is that it diverges because both fermions and bosons give positive and infinite contributions to this quantity. An additional term Q in (16) is proportional to the fluctuations of the non-minimally coupled scalar fields ϕ_s on the horizon Σ and is the average value of the following operator

$$\hat{Q} = 2\pi \sum_{i=1}^{N_s} \xi_i \int_{\Sigma} \sqrt{\gamma}\, \mathrm{d}^2 x \hat{\phi}_i^2. \tag{17}$$

The remarkable property of the model is that for the same values of the parameters of the constituents, that guarantee the finiteness of G, the divergences of S^{SM} are exactly cancelled by the divergences of Q. So in the induced gravity the right-hand-side of (16) is finite and reproduces exactly the Bekenstein-Hawking expression.

3.2 *Models with vector fields*

An important role in the above presented model is played by non-minimal coupling of the scalar constituents. Before discussing the statistical-mechanical meaning of the subtraction formula (16) we present another model (Frolov & Fursaev 1998) in which there is no such coupling.

The model consists of N_s minimally coupled scalar fields ϕ_i with masses $m_{s,i,}$ N_d spinors ψ_j with masses $m_{d,j,}$ and N_v vector fields V_k with masses $m_{v,k}$. The classical actions for scalar and spinor felds are given by equations (6) and (7) with the only important difference that the non-minimal couplings of scalar fields vanish, $\xi_i = 0$ The action of the vector field $\mathrm{V}_{k\mu}$ is

$$I_v[V_k] = -\int \sqrt{-g}\mathrm{d}^4 x \left[\frac{1}{4} F_k^{\mu\nu} F_{k\mu\nu} + \frac{1}{2} m_{v,k}^2 V_k^{\mu} V_{k\mu}\right], \tag{18}$$

where $F_{k\mu\nu} = \nabla_\mu V_{k\nu} - \nabla_\nu \mathrm{V}_{k\mu}$.

The corresponding quantum effective action of the model is

$$\Gamma = \sum_{i=1}^{N_s} \Gamma_s(m_{s,i}) + \sum_{j=1}^{N_d} \Gamma(m_{d,j}) + \sum_{k=1}^{N_v} \Gamma(m_{v,k}). \tag{19}$$

As a result of equation of motion, a massive vector field V_μ obeys the condition $\nabla^\mu V_\mu = 0$, which leaves only three independent components. Under quantization this condition can be realized as a constraint so that the effective action for vector fields takes the form

$$\Gamma_v(m_{v,k}) = \tilde{\Gamma}_v(m_{v,k}) - \Gamma_s(m_{v,k}), \tag{20}$$

$$\tilde{\Gamma}_v(m_{v,k}) = \tfrac{1}{2} \log\det(-\nabla^2 \delta_\nu^\mu + R_\nu^\mu + m_{v,k}^2 \delta_\nu^\mu), \tag{21}$$

where R_ν^μ is the Ricci tensor. The functional $\tilde{\Gamma}_v\,(m_{v,k})$represents the effective action for a massive vector field which we will denote as $A_{k,\mu}$. The classical action for $A_{k,\mu}$

which results in (21) is

$$\tilde{I}_v[A_k] = -\frac{1}{2}\int \mathrm{d}V[\nabla^\mu A_k^\nu \nabla_\mu A_{k\nu} + R_{\mu\nu} A_k^\mu A_k^\nu + m_{v,k}^2 A_k^\mu A_{k\mu}]. \tag{22}$$

The field A_k^μ obeys no constraints and carries an extra degree of freedom. The contribution of this unphysical degree of freedom in (21) is compensated by subtracting the action $\Gamma_s(m_{v,k})$ of a scalar field with the mass $m_{v,k}$, see equation (20).

As in the case of the model with non-minimally coupled scalar field, we require vanishing the cosmological constant and cancellation of the divergencies of the induced Newton constant. These conditions can be written down with the help of the following two functions

$$p(z) = \sum_{i=1}^{N_s} m_{s,i}^{2z} - 4\sum_{j=1}^{N_d} m_{d,j}^{2z} + 3\sum_{k=1}^{N_v} m_{v,k}^{2z}, \tag{23}$$

$$q(z) = \sum_{i=1}^{N_s} m_{s,i}^{2z} + 2\sum_{j=1}^{N_d} m_{d,j}^{2z} - 3\sum_{k=1}^{N_v} m_{v,k}^{2z}. \tag{24}$$

It can be shown that the induced cosmological constant vanishes when

$$p(0) = p(1) = p(2) = p'(2) = 0. \tag{25}$$

The induced Newton constant G is finite if

$$q(0) = q(1) = 0. \tag{26}$$

The constraints result in simple relations

$$N_s = N_d = N_v, \quad \sum_{i=1}^{N_s} m_{s,i}^2 = \sum_{j=1}^{N_d} m_{d,j}^2 = \sum_{k=1}^{N_v} m_{v,k}^2. \tag{27}$$

In particular, they show that one cannot construct the theory with finite cosmological and Newton constants from vector and spinor fields only.

The Newton coupling constant is determined by the following expression

$$\frac{1}{G} = \frac{1}{12\pi} q'(1) = \frac{1}{12\pi}\sum_{i=1}^{N} (m_{s,i}^2 \ln m_{s,i}^2 + 2m_{d,i}^2 \ln m_{d,i}^2 - 3m_{v,i}^2 \ln m_{v,i}^2). \tag{28}$$

Here, according to equation (26), we put $N = N_s = N_d = N_v$. From this expression it is easy to conclude that at least some of the constituents must be heavy and have mass comparable with the Planck mass m_{pl}. For simplicity in what follows we assume that all the constituents are heavy.

Let us analyze models where conditions (25) and (26) are satisfied. Equation (27) are trivially satisfied when all fields are in supersymmetric multiplets. However in such supersymmetric models $p(z) = q(z) \neq 0$ (because masses of the fields in the same supermultiplet coincide) and the induced gravitational constant vanishes. A nontrivial induced gravity theory can be obtained if the supersymmetry is partly broken by splitting the masses of the fields in the supermultiplets.

Let us demonstrate this by an example. Consider the model with N massive supermultiplets. Each multiplet consists of one scalar, one Dirac spinor and one vector

field, so that the numbers of Bose and Fermi degrees of freedom coincide[1]. We suggest that masses of vector and spinor fields are equal, $m_{v,i} = m_{d,i} \equiv m_i$, (here i is the number of the multiplet). The masses of the scalar partners are assumed to be $m_{s,i}$—$(1+x_i)mi$, where x_i; is a dimensionless coefficient. The case when $|xi| \ll 1$ corresponds to slightly broken supersymmetry. For this case

$$p(z) = q(z) = \sum_{i=1}^{N} m_i^{2z}[(1+x_i)^{2z} - 1] \simeq 2z \sum_{i=1}^{N} x_i m_i^{2z}. \tag{29}$$

Now equations (25), (26), and (27) take the simple form

$$\sum_{i=1}^{N} x_i m_i^2 = 0, \quad \sum_{i=1}^{N} x_i m_i^4 = 0, \tag{30}$$

$$\sum_{i=1}^{N} x_i m_i^4 \ln m_i^2 = 0, \quad \frac{1}{G} \simeq \frac{1}{6\pi} \sum_{i=1}^{N} x_i m_i^2 \ln m_i^2. \tag{31}$$

This is a system of linear equations which for $N \geq 4$ has nontrivial solutions.

The calculations give

$$\begin{aligned} S^{SM} = &\frac{1}{48\pi} \sum_{i=1}^{N} [m_{s,i}^2 \ln m_{s,i}^2 + 2m_{d,i}^2 \ln m_{d,i}^2 + 3m_{v,i}^2 \ln m_{v,i}^2] \mathcal{A} \\ &+ \frac{1}{8\pi} \left[cN\mu^2 - \ln \mu^2 \sum_{i=1}^{N} m_{v,i}^2 \right] \mathcal{A}. \end{aligned} \tag{32}$$

Here c = 1n (729/256). The Noether charge in this model arises due to the interaction of the vector fields $A_{k\mu}$ with curvature (see (22) and it is of the form (Frolov & Fursaev 1998):

$$Q = -\pi \int_{\Sigma} \sum_{i=1}^{N} \langle \hat{A}_{i\mu} \hat{A}_{i\nu} \rangle P^{\mu\nu} \mathrm{d}\sigma. \tag{33}$$

Here $P^{\mu\upsilon}$ is a projector onto a two-dimensional surface orthogonal to the bifurcation surface of horizons Σ, and $d\sigma$ is the surface element of Σ. The average $(\hat{A}_{i\mu}\hat{A}_{i\nu})$ is understood as a regularized quantity. The quantity Q has a meaning of Wald's Noether charge associated with nonminimal interaction terms of the vector field.

By using the Pauli-Villars regularization one finds that in the Rindler approximation

$$Q = \frac{\mathcal{A}}{8\pi} \left(cN\mu^2 - \ln \mu^2 \sum_{i=1}^{N} m_{v,i}^2 + \sum_{i=1}^{N} m_{v,i}^2 \ln m_{v,i}^2 \right). \tag{34}$$

This result allows one to show that the Bekenstein-Hawking entropy S^{BH} in induced gravity is the difference of statistical-mechanical entropy S^{SM}, see equation (32), and the Noether charge Q. As can be easily seen, the divergences of S^{SM} are exactly canceled by the divergences of the charge Q, so that one gets the finite

[1] Supersymmetric models with free massive scalar, spinor and vector fields are discussed,for instance, by Lopuszanski & Wolf (1981).

expression

$$S^{BH} \equiv \frac{\mathcal{A}}{4G} = S^{SM} - Q. \tag{35}$$

This result coincides with (16) for the model with non-minimally coupled scalar fields.

4. Statistical-mechanical interpretation of black hole entropy in induced gravity

What is the origin of the compensation mechanism in equations (16) and (35) and what is the statistical-mechanical meaning of the subtraction in this relation? The answer to both questions is in the properties of the operator $\hat{Q}$ of the Noether charge (Wald 1993). In statistical-mechanical computations we consider fields localized in the black hole exterior. The charge $\hat{Q}$ determines the difference between the energy E in the external region B, defined by means of the stress-energy tensor and the canonical energy $\hat{H}$ in the same region

$$\hat{E} = \int_{\mathcal{B}} \hat{T}_{\mu\nu}\zeta^{\mu}\mathrm{d}\sigma^{\nu} = \hat{H} - T_H\hat{Q}. \tag{36}$$

Here ζ^{μ} is the timelike Killing vector field. This formula is valid for both models (for its proof, see Frolov & Fursaev (1997; 1998).

Two energies, $\hat{E}$ and $\hat{H}$, play essentially different roles. The canonical energy is the value of the Hamiltonian $\hat{H}$ which is the generator of translations of the system along ζ^{μ} and which enters definition (5) of the statistical-mechanical entropy S^{SM}. The energy E is the contribution of the constituents to the black hole mass. The number $\nu(E)\Delta E$ of microscopically different physical states of the constituents in the energy interval ΔE near $E = 0$ determines the degeneracy of the black hole mass spectrum. Since the Killing vector ζ^{μ} vanishes at the bifurcation surface Σ of the Killing horizons, the Hamiltonian H is degenerate. One can add to the system an arbitrary number of soft modes, i.e. modes with zero frequencies, without changing the canonical energy. On the other hand, only soft modes contribute to the average of the Noether charge Q. According to equation (36), this removes the degeneracy of the energy E. As a result the infinite number of thermal states of the constituents reduces to the finite number of physical states of the black hole and S^{SM} reduces to the Bekenstein-Hawking value. The corresponding number density $\nu(E=0)$ of physical states is $\exp S^{BH}$ (Frolov & Fursaev 1997). There is a similarity between this mechanism and gauge theories, soft modes playing the role of the pure gauge degrees of freedom.

5. Discussion

Let us note that the concrete models of the induced gravity may differ from the ones considered here, and may contain, for example, finite or infinite number of fields of higher spins. However our consideration indicates that it is quite plausible that the same mechanism of black-hole entropy generation still works

Our discussion can be summarized as follows. The entropy S^{BH} of a black hole in induced gravity is the logarithm of the number of different states of the constituents obeying the condition $\hat{T}\langle_{\mu\nu}\rangle = 0$. The statistical-mechanical explanation of the entropy

becomes possible only when one appeals to a more deep underlying theory in which vacuum is equipped with an additional "fine" structure. The vacuum in the induced gravity is a ground state of "heavy" constituents. "Light" particles are just low energy collective excitations over this vacuum. The black hole entropy S^{BH} acquires its statistical-mechanical meaning because of this additional fine structure of the vacuum of the underlying theory. Two important elements that make the statistical-mechanical picture self-consistent and universal are a representation of gravity as an induced phenomenon and the ultraviolet finiteness of the induced gravitational couplings. These are main lessons the induced gravity teaches us.

It should be emphasized, that the induced gravity approach does not pretend to explain the black hole entropy from the first principles of the fundamental theory of quantum gravity (such as the string theory) but it allows one to demonstrate the universality of the entropy and its independence of the concrete details of such theory. It gives us a hint that only a few quite general properties of the fundamental theory are really required for the statistical-mechanical explanation of the black hole entropy.

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada. One of the authors (V.F.) is grateful to the Killam Trust for its financial support.

References

Bekenstein, J. D. 1972, *Nuov. Cim. Lett,* **4**, 737 (1972); 1973, *Phys. Rev. D,* **7,** 2333.
Bombelli, L., Koul, R., Lee, J., Sorkin, R. 1986, *Phys. Rev. D,* **34**, 373.
Frolov, V. P., Fursaev, D. V. 1997, *Phys. Rev. D,* **56**, 2212.
Frolov, V. P., Fursaev, D. V. 1998, *Phys. Rev. D.* **58**, 124009.
Frolov, V. P., Fursaev, D. V., Zelnikov, A. I. 1997, *Nucl. Phys. B,* **486**, 339.
Frolov, V., Novikov, I. 1993, *Phys. Rev. D,* **48**, 4545.
Hawking, S. W., 1975, *Comm. Math. Phys.,* **43**, 199.
Jacobson, T. 1994, *Black Hole Entropy and Induced Gravity,* preprintgr-qc/9404039.
Lopuszanski, J. T., Wolf, M. 1981, *Nucl. Phys. B,* **184**, 133 (1981).
Sakharov, A. D. 1968, *Sov. Phys. Doklady,* **12**, 1040; 1976, *Theor. Math. Phys.,* **23**, 435.
Wald, R. M. 1993, *Phys. Rev. D,* **48**, R3427.

J. Astrophys. Astr. (1999) **20**, 131–148

Hawking Radiation in String Theory

Sumit R. Das, *Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400 005, India*

Abstract. We review some of the progress in understanding the statistical basis of black hole thermodynamics in string theory. The emphasis is on the "derivation" of Hawking radiation from the unitary decay of near extremal D-brane states. We also review recent progress in understanding Schwarzschild black holes by relating them to D-brane black holes via "boosts" in M-theory.

1. Introduction

In classical gravity a black hole is a classical solution of the equations of motion with a region of spacetime which is causally disconnected from the asymptotic region, the region being bounded by an *event horizon.* In the presence of quantum matter, a black hole, however, radiates nearly thermally—as shown by Hawking in 1975. Accordingly there are standard thermodynamic quantities associated with a black hole. Remarkably these turn out to be rather universal *geometric* quantities. The entropy for large black holes (refered to as the Beckenstein-Hawking entropy S_{BH}) is, in any number of dimensions

$$S_{BH} = \frac{A_H}{4G} \tag{1}$$

where A_H is the area of the event horizon and G is the Newton's constant. The temperature is

$$T_H = \frac{\kappa}{2\pi} \tag{2}$$

where κ is the surface gravity at the horizon. The rate of emission of particles at some frequency ω is given by

$$\Gamma(\omega) = \frac{\sigma(\omega)}{e^{\beta_H \omega} \pm 1} \frac{\mathrm{d}^d k}{(2\pi)^d} \tag{3}$$

where $\sigma(\omega)$ is the absorption cross-section at that frequency and $\beta_H = 1/T_H$. Even though hawking radiation is a rather ordinary process involving pair creation of particles at the event horizon, it leads to the disturbing possibility of a fundamental lack of unitarity and loss of information.

Ever since its discovery, black hole thermodynamics has been a major puzzle. Usually a normal "hot" body appears "hot" simply because there are a large number of microstates for a given macrostates and usual measurements do not keep track of

these microstates. The question is whether black hole is simply such an ordinary hot body and the black hole entropy is related to the number of microstates by Boltzmann relation

$$S_{BH} = \log \Omega. \tag{4}$$

If so, puzzling consequences of black hole thermodynamics can be avoided.

It is clear that in order to arrive at a microscopic theory of black holes, one has to work in a consistent *quantum* theory of gravity. This is similar to the situation in usual thermodynamics: here we have a microscopic description through statistical mechanics only in a proper quantum theory.

The only known consistent quantum theory of gravity is string theory. Indeed, over the past year there has been pathbreaking progress in *deriving* black hole thermodynamics from properties of string states which describe these black holes. In these notes, I will summarize this development. Rather then trying to be exhaustive in the treatment, I will concentrate on the simplest examples and emphasize the physical concepts involved.

2. Black holes as quantum states

To illustrate the relationship between black holes and quantum states consider the simplest black hole, e.g. the Schwarzschild metric in standard four dimensional general relativity

$$ds^2 = -\left(1 - \frac{r_0}{r}\right)dt^2 + \frac{dr^2}{(1 - r_0/r)} + r^2 d\Omega. \tag{5}$$

The gravitational radius is $r_0 = 2GM$ where M is the mass. The horizon is at $r = r_0$ and the curvature at the horizon is $| R | \sim 1/r_0^2$.

Since the low energy limit of superstring theory contains general relativity, this is of course a solution of low energy string theory. The Newton's constant is then related to the string coupling constant g by

$$G = \frac{8\pi^6 \alpha' g^2}{V_6} \tag{6}$$

where $T_s = 1/2\ \pi\alpha'$ is the string tension and V_6 is the volume of the six dimensional compact space which is used to compactify the ten dimensional theory to four dimensions. Let us consider $V_6 \sim \alpha'^3$ (i.e. of the string scale).

In string theory such an object of mass M should be describable in terms of a quantum state. However, just as in any other quantum theory such a state is difficult to describe when the coupling is strong, while it should be easy to describe this state when the coupling is weak. For a given mass, weak string coupling means a small r_0 and hence a tiny black hole. At zero coupling there is in fact no gravitational field and the horizon is singular. The gravitational field builds up as we turn on the effects of coupling. This, however, is the regime where we should not trust the semiclassical description of a black hole in terms of a classical solution.

Indeed, it is clear from (5) that the classical limit corresponds to $g \to 0$ with r_0 held fixed. When r_0 is large in string units the curvature at the horizon is weak and one can trust the semiclassical approximations which lead to the standard formulae of black hole thermodynamics in the previous section. As we just mentioned, this is however the regime where it is difficult to describe the quantum state of the black hole.

Some years ago Susskind (1993) suggested that such Schwarzschild black holes should be described by highly excited states of a single string. Such states are known to be highly degenerate and the corresponding entropy at weak coupling is proportional to the mass of the state in string units

$$S \sim M\sqrt{\alpha'} \tag{7}$$

which appears to contradict the fact that the entropy of a Schwarzschild hole is proportional to the square of the mass in Planck units

$$S_{BH} \sim GM^2 \sim g^2M^2\alpha'. \tag{8}$$

However, as we pointed out, the semiclassical answer (8) is expected to hold for $g^2M\sqrt{\alpha'} \gg 1$ whereas the degeneracy for some given M in (7) is known when $g^2M\sqrt{\alpha'} \ll 1$. One has to find out whether the effects of mass renormalization for finite coupling can indeed change the behavior in (7) to that in (8) (Susskind 1993; Russo & Susskind 1995).

This is however a rather difficult thing to do. In fact it is only recently that a kind of "correspondence principle" has been found which shows that one can indeed obtain the right *qualitative* behavior (Horowitz & Polchinski 1996). The point is: one can only expect the two behaviors to match at some specified value of the coupling. The authors (Horowitz & Polchinski 1996) take this to be the coupling at which the horizon curvature is of the string scale. Then $g^2M \sim 1$ in string units and (8) gives an entropy proportional to the mass in string units, as required by (7). The proportionality constant cannot be fixed by this principle.

It was pointed out by Duff & Rahmfeld (1995) and Sen (1995a, 1995b) that the kind of black holes to look at are a class of *extremal* black holes. These are the BPS saturated states in string theory. Supersymmetry then ensures that the masses of these states are not renormalized. Thus the degeneracy for a given mass level at weak coupling is the degeneracy at strong coupling. Thus we have an accurate way of determining the entropy from the string side. The states considered by Duff & Rahmfeld (1995) and Sen (1995a, 1995b) were highly excited elementary string states. The corresponding black holes have, however, zero horizon radii so that the semiclassical entropy is in fact zero! Of course the curvatures are very large at the horizon and one would expect large stringy corrections to general relativity. Sen (1995 a, 1995b) argued that these stringy corrections should lead us to consider the area of the *stretched horizon* rather than the classical horizon as the semiclassical entropy and found that the string theory entropy is indeed proportional to the area of the stretched horizon. The exact proportionality constant could not be determined because of the inherent ambiguity in defining the stretched horizon.

The classic work of Polchinski (1995) showed how to describe a class of string solitons–those carrying RR charges–at weak string couplings in terms of worldsheet dynamics of D-branes. Very soon, Strominger & Vafa (1996) used a specific configuration of D-branes to describe an extremal BPS saturated five dimensional black

hole with a *finite* horizon radius. They showed that the semiclassical entropy of these holes, as given by (1) agrees *precisely* with the entropy obtained from the degeneracy of the D-brane states. The latter can be computed at weak coupling, but the answer is exact since we are dealing with BPS states—so that the answer can be extrapolated reliably to the strong coupling regime where the semiclassical approximation is good. The result was extended to four dimensional extremal black holes with four large charges by Maldacena & Strominger (1996); Johnson, Khuri & Myers (1996); Maldacena (1996); Balasubramanian and Larson (1996), to spinning black holes in Breckenridge, Myers, Peet & Vafa (1996); Breckenridge *et al* (1996).

Extremal BPS black holes are stable objects which do not radiate. The low energy modes of D-branes excited away from extremality were studied first by Das and Mathur (1996), for the case of the D-string, where a simple physical picture of Hawking radiation resulting from interactions of these excitations was proposed. It was shown by Callan & Maldacena (1996), and Horowitz & Strominger (1996) that similar excitations are responsible to take the five dimensional black hole away from extremality and account for the non-extremal Hawking-Beckenstein entropy exactly. Callan & Maldacena (1996) also showed that the rate of Hawking radiation is proportional to the area of the horizon. Dhar, Mandal & Wadia (1996) explained why this black hole is indeed black in the classical limit and the semiclassical radiation rate was calculated and found to be exactly equal to the horizon area. Remarkably, it was found by Das & Mathur (1996a, 1996b) that the precise low energy radiation rate from the excited D-brane system does not depend on the details of the microscopic theory and is in fact equal to the area of the horizon—thus in exact agreement with the semiclassical result. This was extended to the four dimensional black hole by Gubser & Klebanov (1996). Even more remarkably, at slightly higher energies, the grey body factors were also shown to match in Maldacena & Strominger (1997) for the five dimensional black hole and subsequently by Gubser & Klebanov (1996) for four dimensional black holes.

3. The five dimensional black hole

Let us first describe the five dimensional black hole in the low energy supergravity theory coming from the type IIB superstring. We will follow Maldacena & Strominger (1997). We consider compactification of the theory on a $T^5 = T^4 \times S^1$ along the directions $x^5 \cdots x^9$ with the S^1 along x^5. The volume of the T^4 is V and the radius of S^1 is R. The non-compact four spatial directions are $x^1 \cdots x^4$. The ten dimensional string metric is given by

$$\begin{aligned} ds^2 =& f_1^{-1/2} f_5^{-1/2} \left[-dt^2 + dx_5^2 + \frac{r_0^2}{r^2} (\cosh\sigma dt + \sinh\sigma dx_5)^2 \right] \\ &+ f_1^{1/2} f_5^{-1/2} [dx_6^2 + \cdots dx_9^2] \, f_1^{1/2} f_5^{1/2} \left[\frac{dr^2}{1 - r_0^2/r^2} + r^2 d\Omega_3^2 \right]. \end{aligned} \tag{9}$$

This has a dilaton field

$$e^{-2\phi} = \frac{f_5}{f_1}. \tag{10}$$

The functions f_1 and f_5 are given by

$$f_1(r) = 1 - \frac{r_1^2}{r^2} \quad r_1^2 = r_0^2 \sinh^2 \alpha,$$
$$f_5(r) = 1 - \frac{r_5^2}{r^2} \quad r_1^2 = r_0^2 \sinh^2 \gamma. \tag{11}$$

This represents a 1-brane along S^1 and a 5-brane along the T^5, so that there is also an antisymmetric tensor field with field strength

$$H = 2r_5^2 \epsilon_3 + 2r_1^2 e^{-2\phi} *_6 \epsilon_3 \tag{12}$$

where * denotes the six dimensional Hodge dual. This field strength corresponds to a one brane charge Q_1 and a five brane charge Q_5 which are given by

$$r_1^2 = \frac{16\pi^4 g\alpha'^3 Q_1}{V} \quad r_5^2 = g\alpha' Q_5. \tag{13}$$

We also define

$$r_N^2 = r_0^2 \sinh^2 \sigma. \tag{14}$$

It is clear from the metric that this solution has both left and right moving waves along the S^1. The total momentum (which is quantized) is then given by

$$P = \frac{N}{R} = \frac{VR}{2g^2} r_0^2 \sinh 2\sigma. \tag{15}$$

To obtain the entropy and the temperature for this classical solution one has to convert to Einstein metric. The results are

$$S_{BH} = \frac{2\pi R V r_0^3}{g^2} (\cosh\alpha \cosh\gamma \cosh\sigma),$$
$$T_H^{-1} = 2\pi r_0 (\cosh\alpha \cosh\gamma \cosh\sigma). \tag{16}$$

The extremal limit is

$$r_0 \to 0 \ \alpha, \gamma, \sigma \to \infty \quad r_1, r_5, r_N = \text{finite}. \tag{17}$$

In this limit

$$S_{BH}^{\text{ext}} = 2\pi\sqrt{Q_1 Q_5 N} \quad T_H = 0. \tag{18}$$

In the extremal situation the momentum along x^5 is entirely in one direction.

We will consider departures from extremality of a specific kind, viz. when r_1 and r_5 are fixed to their extremal values by still taking α and γ to infinity, but σ is finite. Now there is both a left and right moving momentum and the various quantities appear as sums of left and rightmoving quantities

$$S = S_L + S_R \frac{1}{T_H} = \frac{1}{2}\left(\frac{1}{T_L} + \frac{1}{T_R}\right) \tag{19}$$

where

$$S_{L,R} = \frac{\pi^2 r_1 r_5 r_0}{4G_5} e^{\pm\sigma},$$
$$T_{L,R} = \frac{r_0 e^{\pm\sigma}}{2\pi r_1 r_5}. \tag{20}$$

The total energy is

$$E = \frac{RQ_1}{g} + \frac{RVQ_5}{g} + E_L + E_R \tag{21}$$

where

$$E_L = \frac{N}{R} + \frac{VRr_0^2 e^{-2\sigma}}{4g^2},$$
$$E_R = \frac{VRr_0^2 e^{-2\sigma}}{4g^2}. \tag{22}$$

In (26) the first two terms are contributions from the masses of the 1-brane and the 5-brane.

When dealing with this nonextremal solution we will sometimes restrict ourselves to the *dilute gas regime.* Consider a left moving wave of amplitude A and wavelength λ. Then in this regime $A \ll \lambda$. Since the mass of the 1-brane is $Q_1R/g\alpha'$ and the typical velocity is A/λ one has

$$\text{Energy} = \frac{Q_1 R}{g\alpha'} \left(\frac{A}{\lambda}\right)^2 = \frac{N}{R}. \tag{23}$$

Using the definitions of r_1 and r_N one sees that the dilute gas regime has

$$r_1 \gg r_N. \tag{24}$$

T-duality along the four directions $(x^6 \cdots x^9)$ interchange r_1 and r_5 so that one must have $r_5 \gg r_N$ as well. A similar condition for right-movers gives $r_0 \ll r_1$, r_5, so that finally the dilute gas condition

$$r_0, r_N \ll r_1, r_5. \tag{25}$$

4. Absorption by black holes

In the first section we found that to find the luminosity of Hawking radiation we need to find the classical absorption coefficient of a black hole. In this section we will restrict our attention to absorption of massless neutral scalars. Such scalar fields satisfy the minimally coupled Klein-Gordon equation in the black hole background. We will also be interested in the result at low energies.

It turns out that to the lowest order in energy expansion, the absorption cross-section of such scalars is completely universal (Das, Gibbons & Mathur 1997). For *any* spherically symmetric black hole in any number of dimensions, the absorption cross-section for a minimally coupled scalar is exactly equal to the area of the horizon

at low energies, i.e.

$$\sigma(\omega) = A_H \tag{26}$$

plus higher order corrections in ω

It was in fact known since the 70's that the low energy cross-section in four dimensional Schwarzschild holes is equal to the area. For the five dimensional hole, this result was first obtained by Dhar, Mandal & Wadia (1996) and by Das & Mathur (1996). That this is in fact a general universal result valid for all spherically symmetric holes in any number of dimensions was realized much later by Das, Gibbons & Mathur (1997).

For the five dimensional hole, it is possible to obtain $\sigma(\omega)$ at higher energies in the following interesting domain. One considers the dilute gas regime, given by (31) and energies

$$\omega r_5 \ll 1 \frac{\omega}{T_{L,R}} \sim O(1). \tag{27}$$

Then it was shown by Maldacena & Strominger (1997) that

$$\sigma(\omega) = 2\pi r_1^2 r_5^2 \frac{\pi\omega}{2} \frac{e^{\omega/T_H} - 1}{(e^{\omega/T_L} - 1)(e^{\omega/T_R} - 1)}. \tag{28}$$

5. The D-brane model: Thermodynamics

In Type IIB string theory the five dimensional black hole is described by Q_5 5D-branes wrapped around $x^5 \cdot x^9$, Q_1 lD-branes wrapped around x^5 and some quantized momentum N/R along x^5 . The states of these system are described in terms of open strings whose ends are stuck on the branes: (1,1) strings with both ends on the lD-brane, (5,5) strings with both ends on the 5Dbrane and (1,5) and (5,1) strings which have one end on a 1Dbrane and the other end on a 5Dbrane.

In one regime of parameters the low energy modes are the (1,5) and (5,1) strings in their lowest oscillator level which can freely move up and down the x^5 direction (Callan & Maldacena 1996): this gives a supersymmetric field theory of $4Q_1Q_5$ massless bosons and $4Q_1Q_5$ fermions in 1 + 1 dimensions. Essentially the problem boils down to that of the low energy modes of a D-string; however the polarization of the states are restricted to lie on the T^4 as a result of the binding between the 1branes and 5-branes. The excitations of a D-string were studied by Das & Mathur (1996), where it was found that S-duality requires that the low energy spectrum of a D-string with RR charge n_w wound around a compact direction of radius R are in fact the excitations of a single long string wrapped n_w times around the compact direction rather than n_w individual strings each wrapped singly. The result of this is that the individual open string modes can have *fractional* momenta n/n_wR though the total momentum must be integer quantized.

It was shown by Maldacena & Susskind (1996) that a similar phenomenon happens in the 1brane 5brane system. The *effective string* is a single long string wound Q_1Q_5 times around x^5. We thus have a field theory of 4 bosons and 4 fermions on a circle of radius Q_1Q_5R. The way to understand this in terms of the underlying Yang-Mills theory is to realise that it is entropically favorable to have Wilson loops of the gauge

field resulting in twisted boundary conditions on the fields (Maldacena 1996; Hashimoto 1996; Hassan & Wadia 1997).

We have to study the statistical mechanics of a collection of f bosons and f fermions in a one dimensional periodic box of size L, with total energy E and total momentum P. For large values of E and P we can treat the system in a canonical ensemble by introducing a temperature $T = 1/\beta$ and a chemical potential α (Das & Mathur 1996) If there are n_r particles with energy e_r and momentum p_r the partition function is

$$Z = \sum_{\{n_r\}} \exp\left[-\beta \sum n_r e_r - \alpha \sum n_r p_r\right] \tag{29}$$

and β, α are determined by requiring

$$E = -\frac{\partial \log Z}{\partial \beta} \quad P = -\frac{\partial \log Z}{\partial \alpha}. \tag{30}$$

The partition function can be easily evaluated and derive E and P as functions *of* α, β

$$\begin{aligned} E &= \frac{fL\pi}{8}\left[\frac{1}{(\beta+\alpha)^2} + \frac{1}{(\beta-\alpha)^2}\right], \\ P &= \frac{fL\pi}{8}\left[\frac{1}{(\beta+\alpha)^2} - \frac{1}{(\beta-\alpha)^2}\right]. \end{aligned} \tag{31}$$

The entropy is then given by

$$S = \log Z + \alpha P + \beta E = \frac{fL\pi}{4}\left[\frac{1}{(\beta+\alpha)} + \frac{1}{(\beta-\alpha)}\right]. \tag{32}$$

The results display the expected left-right splitting, $E = E_L + E_R$, $P = P_L + P_R$, $S = S_L + S_R$ which are best written in terms of

$$\beta_L = \beta + \alpha \quad \beta_R = \beta - \alpha. \tag{33}$$

Eliminating β and a one gets

$$\frac{1}{\beta_i} = T_i = \sqrt{\frac{8E_i}{fL\pi}} = \frac{4S_i}{fL\pi} \tag{34}$$

where i stands for L or R.

The extremal state corresponds to either purely left or purely right movers. For example a purely left moving state has $\beta_R = \infty$ or $T_R = 0$ and hence $T = 0$. This has $E = E_L = P = P_L = N/R$ and $S = S_L$. For the system under consideration one has $L = 2\pi Q_1 Q_5 R$ and $f = 4$. Substituting this in (41) gives the result

$$S_{\text{ext}} = 2\pi\sqrt{Q_1 Q_5 N} \tag{35}$$

in *exact* agreement with the semiclassical Beckenstein-hawking entropy in the previous section.

A slightly nonextremal state may be obtained by adding a few right movers to this extremal state. Recalling that individual modes can have fractional momentum we can now have

$$E_L = P_L = \frac{N}{R} + \frac{n}{Q_1 Q_5 R},$$
$$E_R = -P_R = \frac{n}{Q_1 Q_5 R}. \tag{36}$$

The entropy is now

$$S = 2\pi[\sqrt{Q_1 Q_5 N + n} + \sqrt{n}] \tag{37}$$

again in exact agreement with the semiclassical answer for $n \ll N$ (Callan & Maldacena 1996; Horowitz & Strominger 1996. Since the temperature is the derivative of the entropy with respect to the energy, it is clear that the temperature of the gas would also agree with the Hawking temperature of the nonextremal hole.

6. D-brane decay

A non-extremal state containing both left and right movers will decay by left and right moving open strings combining into a closed string which can now escape from the brane configuration. At low energies, this decay can be computed from the low energy action of the massless modes of the open strings coupled to background massless closed string modes. Since the initial state is highly degenerate and the degeneracy depends on the energy, the outgoing closed string state will have a thermal distribution with a temperature equal to that of the gas of open strings. In the following we will describe the computation of such a decay cross-section into closed string modes which are components of the ten dimensional graviton with polarizations purely along the brane direction. From the non-compact five dimensional view points these are scalars. When the state has a zero momentum along the brane direction, the scalars do not have any Kaluza-Klein charge. We will describe in detail the emission of neutral scalars which would obey the minimally coupled massless Klein-Gordon equation in the noncompact five dimensional spacetime. Extension of these results to charged scalars as well as scalars which are not minimally coupled will be dealt with briefly later. We will also restrict our attention to the five dimensional black hole. The treatment below follows Das & Mathur (1996).

The low energy effective action of the 1brane 5brane system is given by the excitations of an effective string with polarizations along the directions x^I, $I = 6, \ldots 9$. In the presence of a background metric $G^E_{\mu\nu}$ in the Einstein frame and a dilaton field ϕ the action is given by a Nambu-Goto action

$$S = \frac{T_{\text{eff}}}{2} \int d^2\xi e^{-\phi} \sqrt{\det(e^{\phi/2} G^E_{\mu\nu} \partial_m X^\mu \partial_n X^\nu)}. \tag{38}$$

Now we fix a static gauge

$$X^0 = \tau \quad X^5 = \sigma \tag{39}$$

and use the fact that the only polarizations which are relevant are those which are along the brane directions. Expanding the determinant we have

$$S = \frac{T_{\rm eff}}{2}\int {\rm d}x^0 {\rm d}x^5 e^{-\phi}\Big[1+\frac{1}{2}\,e^{\phi/2}G^E_{IJ}\partial_+X^I\partial_-X^J \\ -\frac{1}{8}\,e^{\phi}G^E_{IJ}G^E_{IJ}G^E_{KL}(\partial_+X^I\partial_+X^J)(\partial_-X^K\partial_-X^L)+\cdots\Big] \tag{40}$$

together with fermionic terms.

First consider the case where the dilaton is set to zero. Let h_{IJ} denote the traceless part of the deviation of $G^E_{\mu\nu}$ from the flat metric. Then the dominant interaction term between the bosonic open string modes X^I and the h_{IJ} are given by

$$\sqrt{2}\kappa\int {\rm d}x^0 {\rm d}x^5[h_{IJ}\partial_+X^I\partial_-X^J]. \tag{41}$$

Here κ is related to the ten dimensional gravitational constant by $\kappa^2 = 8\pi G_{10}$. In terms of the string coupling

$$\kappa^2 = 64\pi^7 g^2\alpha'^4. \tag{42}$$

In (42) we have rescaled the fields X^I so that they have a standard kinetic term. Then the effective string tension disappears from this interaction term which also contains only two open string fields. The field h_{IJ} has also been normalized to give rise to a standard kinetic term in the bulk action for the metric, which necessitates the factor $\sqrt{2}\kappa$

Consider a process where two open string states with momenta

$$p = (p_0;0,0,0,0;p_5,0,0,0,0), \\ q = (q_0;0,0,0,0;q_5,0,0,0,0), \tag{43}$$

collide to form a h_{IJ} mode with momentum $k = (k_0, \ldots k_9)$. Note that the open strings can move only along x^5 while the closed string state can move in any direction. Then the decay rate for this process is given by

$$\Gamma(p,q;k) = (2\pi)^2 L\delta(p_0+q_0-k_0)\delta(p_5+q_5-k_5)\,\frac{2\kappa^2(p\cdot q)^2}{(2p_0L)(2q_0L)(2k_0VLV_4)}\,\frac{{\rm d}^4k}{(2\pi)^4}. \tag{44}$$

The delta functions impose energy conservation and momentum conservation along the x^5 direction (momentum is not conserved in the other directions because of Dirichelt conditions). The open string modes are normalized on the circle along while the closed string field is normalized in the entire space with volume VLV_4 where V_4 is the volume of the noncompact four spatial dimensions $x^1\cdots x^4$.

To obtain the total cross-section for production of a scalar one has to now average over all initial states. Since these states are drawn from a thermal ensemble, this means that the decay rate is

$$\Gamma(k) = \left(\frac{L}{2\pi}\right)^2\int_{-\infty}^{\infty}{\rm d}p_5\int_{-\infty}^{\infty}{\rm d}q_5\,\rho(p_0,p_5)\rho(q_0,q_5)\Gamma(p,q,;k) \tag{45}$$

where $\rho(p_0, p_5)$ denotes the bose distribution functions discussed in the previous section. The integral may be evaluated easily. For neutral closed string states $k_5 = 0$ and the answer is

$$\Gamma(k) = 2\pi r_1^2 r_5^2 \frac{\pi\omega}{2} \frac{1}{(e^{\omega/T_L} - 1)(e^{\omega/T_R} - 1)}. \tag{46}$$

We have expressed the answer in terms of the parameters in the classical solution by using the length of the effective string

$$L = 2\pi Q_1 Q_5 R = \frac{8\pi^4 r_1^2 r_5^2 VR}{\kappa^2}. \tag{47}$$

The resulting absorption cross-section

$$\sigma(\omega)^{\text{dbrane}} = 2\pi r_1^2 r_5^2 \frac{\pi\omega}{2} \frac{e^{\omega/T_H} - 1}{(e^{\omega/T_L} - 1)(e^{\omega/T_R} - 1)} \tag{48}$$

is then easily seen to be in exact agreement with the classical answer in (34). In the very low energy limit one has $\sigma^{\text{dbrane}} = A_H$, a result first shown by Das & Mathur (1996). The general grey body factor (56) was shown to agree with the classical answer by Maldacena & Strominger (1996). These results for grey body factors have been extended to four dimensional holes as well.

It is important to note that the value of the effective string tension T_{eff} was irrelevant in the above calculation. This is because the interaction term involves only two open string fields and once these are normalized by absorbing $\sqrt{T_{\text{eff}}}$ the factor disappears from the interaction terms as well.

A more detailed test of the effective string model is provided the emission of *fixed scalars.* These are scalars which do not obey the minimally coupled Klein-Gordon equation (Ferrara, Kallosh & Strominger 1995; Ferrara & Kallosh 1996; Gibbons, Kallosh & Kol 1996; Kol & Rajaraman 1996). An example is the size of the T^4, i.e. the trace of the metric G_{IJ} . Let us write $G_{IJ} = e^{2v}\delta_{IJ}$ in the string frame. Then in the Einstein frame $G^E_{IJ} = e^{v-\phi/2}\delta_{IJ}$ and the *five* dimensional dilaton is given by $\phi 5 = \phi - 2v$. It is then easily seen from (48) that the field *v,* which is a fixed scalar, does not couple to two open strings. The lowest order coupling is to four open strings and comes from the third term in (48). This interaction term depends on t_{eff} after the kinetic terms are properly normalized. In Callan *et al* (1996) the cross-section was computed both classically as well as from the D-brane model and found to agree exactly, provided one has L given by (55) and the tension given by

$$T_{\text{eff}} = \frac{1}{2\pi\alpha' Q_5}. \tag{49}$$

The agreement of the fixed scalar cross-section has been also extended to the four dimensional case (Klebanov & Krasnitz 1996).

7. An apparent puzzle

In a sense these spectacular results are puzzling. String states are expected to describe black holes when (gQ) is large, where g is the string coupling and Q is a typical

charge of the hole. This product (gQ) is in fact the open string coupling. The D-brane calculations are, however, performed at weak open string coupling. For extremal BPS states there are well-known non-renormalization theorems which ensure that the degeneracy of states do not change as we increase the coupling. But for non-BPS states there are no such obvious theorems.

To appreciate the point consider the metric for the five dimensional near extremal black hole described by (11). It is clear from this classical solution that the classical limit of the string theory corresponds to $g \rightarrow 0$ with gQ_1 ,gQ_5,g^2N held fixed (Maldacena & Strominger 1997). In fact we have large black holes (compared to string scale) when gQ_1, gQ_5, $g^2N > 1$ and small holes when gQ_1, gQ_5 , $g^2N < 1$. It is in the latter regime that the D-brane description is good.

In the dilute gas regime $r_N \ll r_1$, r_5 the size of the black hole is controlled by (gQ_1) and (gQ_5) which are the effective open string coupling constants. The full classical solution can be obtained by summing over an infinite number of string diagrams which does not contain any closed string loop, but contains all terms with closed strings terminating on an aribtrary number of branes. Each such insertion carries a factor gQ which has to be held finite. In other words we have to sum over all open string loops. Closed string loops do not contain any factor of the charge Q and are therefore suppressed. This perturbation expansion is a description of the black hole expanded around flat spacetime with the curvature emerging as a result of summing over open string loops.

The question is: why is it that the absorption or emission cross sections calculated in tree level open string theory agree in detail with the semiclassical black hole answers. Why is it that open string loops do not alter the result.

8. The issue of loop corrections

A little thought shows that the situation is not as puzzling as it first appears. Consider for example absorption by extremal black holes which have a single length scale in the problem. Examples are fat black holes with $r_1 = r_5 = r_N = R$ or extremal branes, like the 3brane. Let us call this length scale l. In the classical solution the string coupling can enter only through this length scale l which is typically given by the form

$$l^{(d-3)} \sim gQ\alpha'^{(d-3/2)} \tag{50}$$

where d denotes the number of noncompact dimensions. It is then clear that the classical absorption cross-section has to be of the form

$$\sigma_{\text{class}} \sim l^{d-2} F\left(\omega^{(d-3)} gQ\alpha'^{(d-3/2)}\right). \tag{51}$$

On general grounds we expect that this classical answer should agree with the D-brane answer when gQ is large. However the above expression shows that for sufficiently small ω one may have the factor $\omega^{(d-3)}gQ\alpha'^{(d-3/2)}$ small even if gQ is large so that one may imagine performing a Taylor expansion of the function F, which then becomes a power series expansion in the string coupling g as well (Klebanov 1997). The spectacular success of the tree level D-brane calculations of the absorption cross-section then means that the *lowest order* term in this expansion has been shown to agree with the *lowest order* term in D-brane open string perturbation theory.

The puzzle regarding this agreement of D-brane and classical calculations may be now restated as follows: In the classical limit a higher power of the string coupling comes with a higher power of the energy in a specific way dictated by (59). On the other hand, on the D-brane side these higher powers of coupling are to be obtained in open string perturbation theory and there is no *α priori* reason why this should also involve higher powers of energy in precisely the same way.

9. Some one loop results

General results for such "low energy non-renormalization" properties are not known at the moment. For some processes, Maldacena (1997) has given arguments based on low energy effective actions on the brane. However we have seen that processes involving higher dimension operators on the effective string also agree with semi classical results: these cannot be analyzed using the low energy Yang-Mills action on the brane. For the case of parallel branes, like the three-brane (Klebanov 1997; Gubser, Klebanov & Tseytlin 1997) where the low energy absorption also agrees with the semiclassical result, one may directly analyze one loop corrections to the tree level diagrams using standard string perturbation theory. In (Das 1997) a class of such corrections were studied in detail and it was found that for all such processes the one (open string) loop corrections vanish at low energies. Furthermore, the nonvanishing loop corrections appear with exactly the powers of energy as dictated by the expansion of the semiclassical answer in (59) (Das 1997).

10. Schwarzschild black holes

All the results described above are for charged black holes. Recently there has been some progress in understanding the thermodynamics and Hawking radiation properties of neutral black holes in string theory, by embedding them in M-theory.

By now, there is considerable evidence that ten dimensional string theory is a dimensional reduction of an eleven dimensional theory, usually called M-theory. While little is known about M-theory, it is known that the low energy theory is in fact eleven dimensional supergravity. Type IIA theory is most directly related to M-theory: if we consider M theory compactified on a circle of radius *R,* the dimensionally reduced ten dimensional theory is Type IIA superstring theory. The string coupling g_s and the string length l_s is given in terms of the eleven dimensional Planck length *lp* and *R* by the relations

$$g_s = \left(\frac{R}{l_p}\right)^{3/2} \quad l_s = \left(\frac{l_p^3}{R}\right)^{1/2}. \tag{52}$$

Thus for large *R* the string theory is strongly coupled and an *eleven* dimen sional Lorentz invariance becomes apparent. States which carry momentum along the x^{11} direction are described as states which carry RR charges in string theory, i.e. 0-branes.

Many of the charges of the black holes described above are in fact RR charges. Thus it appears that states with different RR charges are in fact related to each other

by Lorentz boosts in the 11th direction. Thus one should be able to relate properties of Schwarzschild black holes to those of charged black holes by performing such a boost (Banks *et al* 1997).

A little thought shows that things are not so straightforward since the direction of the boost is in fact a compact direction. The momentum in this direction is quantized in units of $1/R$ and the quantized momentum counts the number of waves which fit in this circle. This number is a geometric quantity and cannot be changed by change of a reference frame.

Nevertheless, it turns out that neutral and charged black holes are related to each other not by genuine boosts, but by maps which are boosts in the covering space of the compact cricle.

We will consider 11-dimensional M-theory compactified on a space $T^p \times S^1$. The torus has sides L_i, $i = 1, \cdots 6$ and the circle, which will be taken to be along the x^{11} direction, has a radius R. In this space a "Schwarzschild string" is a product of a Schwarzschild black hole in the noncompact $(10 - p)$ dimensions and flat space along the $(p + 1)$ compact dimensions. (This is thus a $(p + 1)$ black brane. We call this a string since it extends along x^{11}.) The 11-dimensional metric is

$$ds_{11}^2 = -\left(1 - \left(\frac{r_0}{r}\right)^n\right)dt^2 + \frac{dr^2}{(1 - (r_0/r)^n)} + r^2 d\Omega_{n+1} + dz^2 + \sum_{i=1}^{p}(dx^i)^2, \qquad (53)$$

where $n = 7 - p$ and $r^2 = \sum_{i=p+1}^{9}(x^i)^2$ The map which relates the Schwarzschild string to a charged hole consists of a boost in the covering space in which x^{11} is noncompact, but the other ρ directions are still compact

$$z' = z\cos h\alpha + t\sin h\alpha, \quad t' = t\cos h\alpha + z\sin h\alpha. \qquad (54)$$

The boosted coordinate z' has to be then compactified on a radius R' which is related to R by a Standard Lorentz contraction

$$R' = R/\cos h\alpha. \qquad (55)$$

By standard Kaluza-Klein reduction along z' (as in Home, Horowitz & Steif 1992) the resulting metric then represents an RR charged black hole in $(10 - p)$ dimensions. The relationship (63) ensures that the energy and momenta transform correctly and the semiclassical entropy is kept invariant under the boosts (Das *et al* 1997).

The above transformation is not a symmetry of the theory and thus relates two physically inequivalent configurations. It does provide a well defined map which relates these. In fact, it was shown in Das *et al* (1997) that absorption cross sections and hence Hawking radiation rates are related using elementary properties of Lorentz transformations, when both R and R' are sufficiently large, i.e. in the long string limit. If $\sigma(\omega', q', \vec{k}, ; A')$ denotes the absorption cross-section of some particle of energy ω, momentum q along x^{11}, and transverse momentum $\vec{k}$ by the black hole A, then the absorption cross-section $\sigma'(\omega', q', \vec{k}; A')$ of the transformed particle with energy momentum given by (w', q', k) by the transformed black hole A' is given by

$$\sigma'(\omega', q', \vec{k}'; \mathcal{A}') = \frac{\omega}{\omega'}\sigma(\omega, q, \vec{k}; \mathcal{A}),$$
$$q' = q\cos h\alpha + \omega\sin h\alpha, \quad \omega' = q\sin h\alpha + \omega\cos h\alpha. \qquad (56)$$

The emission rate $\Gamma(k)$ is related to σ by the usual relation

$$\Gamma(\omega, q, \vec{k}) = \frac{\sigma(\omega, q, \vec{k})}{e^{\xi} \mp 1} \frac{\mathrm{d}^d \vec{k}}{(2\pi)^d}; \quad \xi = (\omega - q\phi)/T, \tag{57}$$

where T is the temperature of the black hole, and ϕ is the potential of the Kaluza-Klein gauge field at the horizon (in 10d language) and d is the number of transverse dimensions. The relation (65) does not depend on *how* the cross-section is calculated, but follows from the fact that decay rates decrease by a time dilation factor. Hence if one had a microscopic derivation of the cross-section for the charged black hole one may use this to obtain a microscopic derivation of the cross-section for the neutral black hole.

If we perform T-duality transformations along the p compact direction we have a product of a standard D-p brane black hole in string theory with the M-theory circle.

Remarkably combining such "boosts" with T-duality symmetries, it is possible to relate a four dimensional Schwarzschild black hole with three sets of 5-branes in M-theory intersecting along x^{11} and carrying some momentum along x^{11} (Argurio, Englert & Houart 1998). Start with a spacetime which is a 4d Schwarzschild hole with seven more compact directions which are taken to be x^i where $i = 1 \cdots 6, 11$. Now perform the following operations.

1. Boost along x^{11} by parameter α_1
2. T-dualize along (1234).
3. Boost along x^{11} by parameter α_2.
4. T-dualize along (1256).
5. Boost along x^{11} by parameter α_3.
6. T-dualize along (1234).
7. Boost along x^{11} by parameter α_4.
8. T-dualize along (1256).

At the end of these steps we have the configuration with three sets of intersecting fivebranes mentioned above. This is a description of a four dimensional black hole with four charges in M theory (Cvetic & Youm 1996; Cvetic & Tseytlin 1996). The entropy is the same at all stages. At every step we will denote the string coupling by g_n, the string length by l_n, the radii of the torus by $L_i^{(n)}$ and the x^{11} radius by R_n where $n = 1 \cdots 6$. The first seven steps were used by Argurio, Englert & Houart (1998). For us, however, the last step is important.

Since the entropy of the latter near the extremal hole is known to follow from a microscopic calculation (Maldacena & Strominger 1996; Johnson, Khuri & Myers 1996; Horowitz, Lowe & Maldacena 1996; Balasubramanian & Larsen 1996; Tseytlin 1996; Gauntlett, Kastor & Traschen 1996; Klebanov & Tseytlin 1996), and thus we get the entropy of the neutral hole (Argurio, Englert & Houart 1998). Furthermore since absorption cross-sections are also related in a well-defined fashion, we might hope to understand Hawking radiation from the four dimensional Schwarzschild hole using the microscopic model of the three sets of five branes. This is indeed the case (Das, Mathur & Ramadevi 1998).

Consider the emission of a minimally coupled scalar from the 4d Schwarzschild hole carrying no momentum along x^{11}. As we perform the above steps this emitted particle acquires a x^{11} momentum—i.e. an RR charge in the language of string theory.

In general, the emitted particle gets endowed with various kinds of brane charges, just as the black hole and it would be hopeless to perform a microscopic calculation of this. Remarkably there is a drastic simplification at low energies. This is because at the end of the second step the emitted particle is an *extremal* D-4 brane with a small transverse motion. If this extremal D4 brane was at rest this would be an M5 brane at rest in M-theory with x^{11} as one of the longitudinal direction. This object is invariant under boosts along x^{11}. The crucial point is that to lowest order in the transverse velocity this continues to be the case. As a result after the second step, till the 7th step the emitted particle continues to be a 4D brane with no other charge : the wrapping of this D4 brane of course changes as a result of the intermediate steps. No other kind of charge is present at lowest order of the energy. This means that the metric produced by this emitted particle remains that of a D4 brane—however the energy and the equivalent x^{11} momentum change under boosts by trivial Lorentz contraction relations. The final step converts this particle into a 0-brane in string theory language—or a particle with x^{11} momentum in M-theory. The microscopic answer for the low energy cross-section in the near extremal case is already known from the calculation by Gubser & Klebanov (1996).

Using the relationship between the cross-sections under "boosts" mentioned above and that under T-duality symmetries we can now obtain a prediction for the semi classical cross-section for emission of such a 0-brane, starting from the universal cross-section of neutral massless particles in the 4D Schwarzschild hole. The result is (Das, Mathur & Ramadevi 1998)

$$\sigma_8 = A_8 \frac{\omega_8 - q_8 \tan h\alpha_1}{\omega_8} \tag{58}$$

where ω_8 and q_8 are the energy and the equivalent x^{11} momentum at the end of these steps α_1 is the boost parameter of the first boost in the above list and A_8 is the area of the (two-dimensional) horizon of the final black hole. The near-extremal case corresponds to $\alpha_i \to \infty$ .The result for σ_8 in (58) in this case is in *exact* agreement with the microscopic calculation of Gubser & Klebanov (1996).

11. Outlook

String theory has successfully provided a microscopic basis for black hole entropy and Hawking radiation for a class of black holes. While we understand the near-extremal holes best, we are beginning to get encouraging results for Schwarzschild holes as well—which indicates that the same mechanism is at work for these neutral holes.

It is remarkable that all the ingredients of string theory went into this: extra dimensions, supersymmetry and duality. And the content of the theory had to be just right in order to reproduce the semiclassical results exactly in the relevant domain. This gives a fair degree of confidence that we are on the right track.

It is disturbing, though, that we do not yet have a picture of the rich space-time structure – particularly the rich physics of the horizon–in this approach. As we discussed nontrivial space time structure arises in this approach as higher order effects in the coupling. For all the cases described above we find that the weak coupling microscopic results can be extrapolated to strong coupling where the solitons are better

described as semiclassical black holes—and these microscopic results are in spectacular agreement with the semiclassical answers. For many cases we also understand why this extrapolation works. Yet we do not have a good physical picture at strong coupling. It is likely that such a physical picture is needed before we can say unambiguously that there is no information loss problem for black holes.

References

Argurio, R., Englert, R., Houart, L. 1998, hep-th / 9801053.
Balasubramanian, V., Larsen, F. 1996, *Phys. Lett.,* **B478**, 199; hep-th / 9604189.
Banks, T., Fischler, W., Klebanov, I., Susskind, L. 1997, hep-th / 9709091.
Breckenridge, J., Myers, R., Peet, A., Vafa, C. 1996, hep-th / 9602065.
Breckenridge, J., Lowe, D., Myers, R., Peet, A., Strominger, A., Vafa, C. 1996, *Phys. Lett.,* **B381**, 423.
Callan, C, Gubser, S., Klebanov, I., Tseytlin, A. 1996, hep-th / 961O172.
Callan, C. G., Maldacena, J. 1996, *Nucl. Phys.,* **B475**, 645; hep-th / 9602043.
Cvetic, M., Youm, D. 1996, *Phys. Rev.,* **D53**, 584; hep-th / 9507090.
Cvetic, M., Tseytlin, A. 1996, *Phys. Lett,* **B366**, 95; hep-th / 9510097.
Das, S. R. 1997, *Phys. Rev.,* **D56**, 3582.
Das, S. R. 1997, hep-th / 97 (talk at Strings 97).
Das, S. R., Gibbons, G., Mathur, S. D. 1997, *Phys. Rev. Lett.,* **78**, 417; hep-th / 9609052.
Das, S. R., Mathur, S. D. 1996, *Phys. Lett,* **B375**, 103; hep-th / 9601152.
Das, S. R., Mathur, S. D. 1996a, *Nucl. Phys.,* **B478**, 561; hep-th / 9606185.
Das, S. R., Mathur, S. D. 1996b, *Nucl. Phys.,* **B482**, 153; hep-th / 9607149.
Das, S. R., Mathur, S. D., Kalyana Rama, S., Ramadevi, P. 1997, *Nucl. Phys.,* **B527**, 187; hep-th / 9711003.
Das, S. R., Mathur, S. D., Ramadevi, P. 1998, *Phys. Rev.* **D59**, 084001; hep-th / 9803078.
Dhar, A., Mandal, G., Wadia, S. R. 1996, *Phys. Lett,* **B388**, 51; hep-th / 9605234.
Duff, M., Rahmfeld, J. 1995, *Phys. Lett,* **B345**, 441; hep-th / 9406105.
Ferrara, S., Kallosh, R. 1996, hep-th / 9602136; hep-th / 9603090.
Ferrara, S., Kallosh, R., Strominger, A. 1995, *Phys. Rev.,* **D52**, 5412; hep-th / 9508072.
Gauntlett, J., Kastor, D., Traschen, J. 1996, hep-th / 9604179.
Gibbons, G., Kallosh, R., Kol, B. 1996, *Phys. Rev. Lett.,* **77**, 4992; hep-th / 9607108.
Gubser, S., Klebanov, I. 1996, *Phys. Rev. Lett,* **77**, 4491; hep-th / 9609076.
Gubser, S., Klebanov, I., Tseytlin, A. 1997, *Nucl. Phys.,* **B499**, 217; hepth / 9703040.
Hashimoto, A. 1996; hep-th / 9610250.
Hassan, S. F., Wadia, S. 1997; hep-th / 9703163.
Home, J., Horowitz, G., Steif, A. 1992, *Phys. Rev. Lett,* **68,** 568.
Horowitz, G., Lowe, D., Maldacena, J. 1996, *Phys. Rev. Lett,* **77**, 430; hep-th / 9603195.
Horowitz, G., Martinec, E. 1997, hep-th / 9710217.
Horowitz, G., Polchinski, J. 1996, hepth / 9612146.
Horowitz, G., Strominger, A. 1996, *Phys. Rev. Lett,* 77, 2368; hep-th / 9602051.
Johnson, C., Khuri, R., Myers, R. 1996, *Phys. Lett* **B378**, 78; hep-th / 9603061.
Klebanov, I. 1997, *Nucl. Phys.,* **B496**, 231; hepth / 9702076.
Klebanov, L, Krasnitz, M. 1996, hep-th / 961205.
Klebanov, L., Susskind, L. 1997, hep-th / 9709108.
Klebanov, I., Tseytlin, A. 1996, hep-th / 9604166 and hep-th / 9607107.
Kol, B., Rajaraman, A. 1996, hep-th / 9608126.
Maldacena, J. 1996; hep-th / 9607235.
Maldacena, J. 1997, *Phys. Rev.,* **D55**, 7645.
Maldacena, J., Strominger, A. 1996, *Phys. Rev. Lett,* **77**, 428; hep-th / 9603060.
Maldacena, J, Strominger, A. 1997, *Phys. Rev. D,* **55** 861; hep-th / 9609026.
Maldacena, J., Susskind, L. 1996, *Nucl. Phys,* **B475**, 679; hep-th / 9604042.
Mathur, S. D. 1996, hep-th / 9609053

Polchinski, J. 1995, *Phys. Rev. Lett.,* **75,** 4724; hep-th / 9510017.
Russo, J., Susskind, L. 1995, *Nucl. Phys.,* **B437**, 611.
Sen, A. 1995a, *Nucl. Phys.,* **B440**, 421; hepth / 9411187;
Sen, A. 1995b, *Mod. Phys. Lett,* **Al0**, 2081.
Strominger, A., Vafa, C. 1996, *Phys. Lett.,* **B379**, 99; hep-th / 9601029.
Susskind, L. 1993, hep-th / 9309145.
Tseytlin, A. 1996, hep-th / 9604035.

J. Astrophys. Astr. (1999) **20**, 149–164

On M-Theory

Hermann Nicolai, *Max-Planck-Institut, fiir Gravitationsphysik, Albert-Einstein-Institut, Am Mühlenberg 1, D-14476 Golm, Germany*

Abstract. This contribution gives a personal view on recent attempts to find a unified framework for non-perturbative string theories, with special emphasis on the hidden symmetries of supergravity and their possible role in this endeavor. A reformulation of $d = 11$ supergravity with enlarged tangent space symmetry SO(1, 2) × SO(16) is discussed from this perspective, as well as an ansatz to construct yet further versions with SO(1, 1) × $SO(16)^{\infty}$ and possibly even $SO(1, 1)_{+}$ × $ISO(16)^{\infty}$ tangent space symmetry. It is suggested that upon "third quantization", dimensionally reduced maximal supergravity may have an equally important role to play in this unification as the dimensionally reduced maximally supersymmetric $SU(\infty)$ Yang Mills theory.

Key words. Supergravity—hidden symmetries—superstrings.

1. Introduction

Many theorists now believe that there is a unified framework for all string theories, which also accomodates $d = 11$ supergravity (Cremmer, Julia & Scherk 1978). Much of the evidence for this elusive theory, called "M-Theory" (Witten 1995; Townsend 1995), is based on recent work on duality symmetries in string theory which suggests that all string theories are connected through a web of non-perturbative dualities (Font *et al.* 1990; Rey 1991; Sen 1993,94; Schwarz & Sen 1994; Duff & Khuri 1994; Giveon *et al.* 1994; Hull & Townsend 1995; Witten 1995; Kachru & Vafa 1995; Schwarz 1995, 96; Duff 1996; Horava 1996). Although it is unknown what M-theory really is, we can probably assert with some confidence

(i) that it will be a pregeometrical theory, in which space-time as we know it will emerge as a secondary concept (which also means that it makes little sense to claim that the theory "lives" in either ten or eleven dimensions), and
(ii) that it should possess a huge symmetry involving new and unexplored types of Lie algebras (such as hyperbolic Kac Moody algebras), and perhaps other exotic structures such as quantum groups. In particular, the theory should be background independent and should be logically deducible from a vast generalization of the principles underlying general relativity.

According to a widely acclaimed recent proposal (Banks *et al.* 1997) M-Theory "is" the $N \to \infty$ limit of the maximally supersymmetric quantum mechanical $SU(N)$ matrix model (Claudson & Halpern 1985; Flume 1985; Baake, Reinicke & Rittenberg 1985) (see deWit (1997), Banks (1997) and Bigatti & Susskind (1997) for recent reviews, points of view and comprehensive lists of references). This model had already

appeared in an earlier investigation of the $d = 11$ supermembrane (Bergshoeff, Sezgin & Townsend 1987; 1988) in a flat background in the light cone gauge (deWit, Hoppe & Nicolai 1988). Crucial steps in the developments leading up to this proposal were the discovery of Dirichlet p-branes and their role in the description of non-perturbative string states (Polchinski 1995) and the realization that the dynamics of an ensemble of such objects is described by dimensionally reduced supersymmetric Yang Mills theories (Witten 1996; Polchinski 1996). Although there are a host of unsolved problems in matrix theory, two central ones can perhaps be singled out: one is the question whether the matrix model admits massless normalizable states for any N (see Fröhlich & Hoppe 1997; Yi 1997; Sethi & Stern 1997; Porrati & Rozenberg 1997; Hoppe 1997; Green & Gutperle 1997; Halpern & Schwarz 1997) for recent work in this direction); the other is related to the still unproven existence of the $N \to \infty$ limit. This would have to be a weak limit in the sense of quantum field theory, requiring the existence of a universal function $g = g(N)$ (the coupling constant of the $SU(N)$ matrix model) such that the limit $N \to \infty$ exists for all correlators. The existence of this limit would be equivalent to the renormalizability of the supermembrane (deWit, Hoppe & Nicolai 1988). However, even if these problems can be solved eventually, important questions remain with regard to the assertions made above: while matrix theory is pregeometrical in the sense that the target space coordinates are replaced by matrices, thus implying a kind of non-commutative geometry, the hidden exceptional symmetries of dimensionally reduced supergravities discovered long ago (Julia 1979) are hard to come by (see Elitzur *et al.* (1997) and references therein).

In the first part of this contribution, I will report on work (Melosch & Nicolai 1997), which was motivated by recent advances in string theory as well as the possible existence of an Ashtekar-type canonical formulation of $d = 11$ supergravity. Although at first sight our results, which build on the earlier work of (deWit & Nicolai 1985; Nicolai 1987), may seem to be of little import for the issues raised above, I will argue that they could actually be relevant, assuming (as we do) that the success of the search for M-Theory will crucially depend on the identification of its underlying symmetries, and that the hidden exceptional symmetries of maximal supergravity theories may provide important clues as to where we should be looking. Namely, as shown in (deWit & Nicolai 1985; Nicolai 1987) the local symmetries of the dimensionally reduced theories can be partially "lifted" to eleven dimensions, indicating that these symmetries may have a role to play also in a wider context than that of dimensionally reduced supergravity. The existence of alternative versions of $d = 11$ supergravity, which, though equivalent on-shell to the original version of (Cremmer, Julia & Scherk 1978), differ from it off-shell, suggests the existence of a novel kind of "exceptional geometry" for $d = 11$ supergravity and the bigger theory containing it. This new geometry would be intimately tied to the special properties of the exceptional groups, and would be characterized by relations such as (3)–(5) below, which have no analog in ordinary Riemannian geometry. The hope is, of course, that one may in this way gain valuable insights into what the (surely exceptional) geometry of M-Theory might look like, and that our construction may provide a simplified model for it. After all, we do not even know what the basic physical concepts and mathematical "objects" (matrices, BRST string functionals, spin networks,...?) of such a theory should be, especially if it is to be a truly pregeometrical theory of quantum gravity.

The second part of this paper discusses the infinite dimensional symmetries of $d = 2$ supergravities (Julia 1981; Julia 1982, 1984; Breitenlohner & Maison 1987; Breitenlohner, Maison & Gibbons 1988; Nicolai 1991; Julia & Nicolai 1996; Bernard & Julia 1997) and an ansatz that would incorporate them into the construction of (Melosch & Nicolai 1997; deWit & Nicolai 1985; Nicolai 1987). The point of view adopted here is that the fundamental object of M-Theory could well be a kind of "Unendlichbein" belonging to an infinite dimensional coset space (Ashtekar 1986), which would generalize the space $GL(4, \mathbf{R})/SO(1, 3)$ of general relativity. This bein would be acted upon from the right side by a huge extension of the Lorentz group, containing not only space-time, but also internal symmetries, and perhaps even local supersymmetries. For the left action, one would have to appeal to some kind of generalized covariance principle. An intriguing, but also puzzling, feature of the alternative formulations of $d = 11$ supergravity is the apparent loss of manifest general covariance, as well as the precise significance of the global E_{11-d} symmetries of the dimensionally reduced theories. This could mean that in the final formulation, general covariance will have to be replaced by something else.

The approach taken here is thus different from and arguably even more speculative than current ideas based on matrix theory, exploiting the observation that instead of dimensionally reducing the maximally extended *rigidly* supersymmetric theory to one dimension, one might equally well contemplate reducing the maximally extended *locally* supersymmetric theory to one (light-like ≡ null) dimension. While matrix theory acquires an infinite number of degrees of freedom only in the $N \to \infty$ limit, the chirally reduced supergravity would have an infinite number from the outset, being one half of a field theory in two dimensions. The basic idea is then that upon quantization the latter might undergo a similarly far-reaching metamorphosis as the quantum mechanical matrix model, its physical states being transmuted into "target space" degrees of freedom as in string theory (Nicolai 1987). This proposal would amount to a third quantization of maximal ($N = 16$) supergravity in two dimensions, where by "third quantization" I mean that the quantum treatment should take into account the gravitational degrees of freedom on the worldsheet, i.e. its (super)moduli for arbitrary genus. The model can be viewed as a very special example of $d = 2$ quantum cosmology; with the appropriate vertex operator insertions the resulting multiply connected $d = 2$ "universes" can be alternatively interpreted as multistring scattering diagrams (Mandelstam 1973; Giddings & Wolpert 1987). One attractive feature of this proposal is that it might naturally bring in E_{10} as a kind of non-perturbative spectrum generating (rigid) symmetry acting on the third quantized Hilbert space, which would mix the worldsheet moduli with the propagating degrees of freedom. A drawback is that these theories are even harder to quantize than the matrix model (see, however, Nicolai, Korotkin & Samtleben (1997) and references therein).

2. SO(1, 2) × SO(16) invariant supergravity in eleven dimensions

In (deWit & Nicolai 1985; Nicolai 1987), new versions of $d = 11$ supergravity (Cremmer, Julia & Scherk 1985) with local $SO(1, 3) \times SU(8)$ and $SO(1, 2) \times SO(16)$ tangent space symmetries, respectively, have been constructed. Melosch & Nicolai (1997) develop these results further (for the $SO(1, 2) \times SO(16)$ invariant version of

(Nicolai 1987) and also discusses a hamiltonian formulation in terms of the new variables. In both versions the supersymmetry variations acquire a polynomial form from which the corresponding formulas for the maximal supergravities in four and three dimensions can be read off directly and without the need for complicated duality redefinitions. This reformulation can thus be regarded as a step towards the complete fusion of the bosonic degrees of freedom of $d = 11$ supergravity (i.e. the elfbein E_M^A and the antisymmetric tensor A_{MNP}) in a way which is in harmony with the hidden symmetries of the dimensionally reduced theories.

For lack of space, and to exhibit the salient features as clearly as possible I will restrict the discussion to the bosonic sector. To derive the SO(1, 2) × SO(16) invariant version of (Nicolai 1987; Melosch & Nicolai 1997) from the original formulation of $d = 11$ supergravity, one first breaks the original tangent space symmetry SO(1,10) to its subgroup SO(1, 2) × SO(8) through a partial choice of gauge for the elfbein, and subsequently enlarges it again to SO(1, 2) × SO(16) by introducing new gauge degrees of freedom. The symmetry enhancement of the transverse (helicity) group SO(9) ⊂ SO(1,10) to SO(16) requires suitable redefinitions of the bosonic and fermionic fields, or, more succinctly, their combination into tensors w r.t. the new tangent space symmetry. The construction thus requires a 3 + 8 split of the $d = 11$ coordinates and indices, implying a similar split for all tensors of the theory. It is important, however, that the dependence on all eleven coordinates is retained throughout.

The elfbein and the three-index photon are thus combined into new objects covariant w r.t. to the new tangent space symmetry. In the special Lorentz gauge preserving SO(1, 2) × SO(8) the elfbein takes the form

$$E_M^A = \begin{pmatrix} \Delta^{-1} e_\mu^a & B_\mu^m e_m^a \\ 0 & e_m^a \end{pmatrix} \tag{1}$$

where curved $d = 11$ indices are decomposed as $M = (\mu, m)$ with $\mu = 0, 1, 2$ and $m = 3, \ldots, 10$ (with a similar decomposition of the flat indices), and $\Delta := \det e_m^a$. In this gauge, the elfbein contains the (Weyl resealed) dreibein and the Kaluza Klein vector B_μ^m both of which will be kept in the new formulation. By contrast, the internal achtbein is replaced by a rectangular 248-bein (e_{IJ}^m, e_A^m) containing the remaining "matter-like" degrees of freedom, where ($[IJ]$, A) label the 248-dimensional adjoint representation of E_8 in the SO(16) decomposition. This 248-bein, which in the reduction to three dimensions contains all the propagating bosonic matter degrees of freedom of $d = 3$, $N = 16$ supergravity, is defined in a special SO(16) gauge by

$$(e_{IJ}^m, e_A^m) := \begin{cases} \Delta^{-1} e_a^m \Gamma^a_{\alpha\dot\beta} & \text{if } [IJ] \text{ or } A = (\alpha\dot\beta) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where the SO(16) indices IJ or A are decomposed w.r.t. the diagonal subgroup SO(8) ≡ $(\mathrm{SO}(8) \times \mathrm{SO}(8))_{\mathrm{diag}}$ of SO(16) (see Nicolai (1987) for details). Being the inverse densitized internal achtbein contracted with an SO(8) Γ-matrix, this object is very much analogous to the inverse densitized triad in the framework of Ashtekar's reformulation of Einstein's theory (Ashtekar 1986). Note that, due to its rectangularity, there does not exist an inverse for the 248-bein (nor is one needed for the supersymmetry variations and the equations of motion!). In addition we need the composite fields (Q_μ^{IJ}, P_μ^A) and (Q_m^{IJ}, P_m^A), which together make up an E_8 connection in

eleven dimensions and whose explicit expressions in terms of the $d = 11$ coefficients of anholonomity and the four-index field strength F_{MNPQ} can be found in (Nicolai 1987).

The new geometry is encoded into algebraic constraints between the vielbein components, which are without analog in ordinary Riemannian geometry because they rely in an essential way on special properties of the exceptional group E_8. We have

$$e^m_A e^n_A - \tfrac{1}{2} e^m_{IJ} e^n_{IJ} = 0 \tag{3}$$

and

$$\Gamma^{IJ}_{AB}(e^m_B e^n_{IJ} - e^n_B e^m_{IJ}) = 0 \qquad \Gamma^{IJ}_{AB} e^m_A e^n_B + 4 e^m_{K[I} e^n_{J]K} = 0 \tag{4}$$

where $\Gamma^I_{A\dot{A}}$ are the standard SO(16) Γ-matrices and $\Gamma^{IJ}_{AB} \equiv (\Gamma^{[I}\,\Gamma^{J]})_{AB}$, the minus sign in (3) reflects the fact that we are dealing with the maximally non-compact form $E_{8(+8)}$. While the SO(16) covariance of these equations is manifest, it turns out, remarkably, that they are also covariant under E_8. Obviously, (3) and (4) correspond to the singlet and the adjoint representations of E_8. More complicated are the following relations transforming in the 3875 representation of E_8

$$\begin{aligned} e^{(m}_{IK} e^{n)}_{JK} - \tfrac{1}{16}\delta_{IJ} e^m_{KL} e^n_{KL} &= 0, \\ \Gamma^K_{\dot{A}B} e^{(m}_B e^{n)}_{IK} - \tfrac{1}{14}\Gamma^{IKL}_{\dot{A}B} e^{(m}_B e^{n)}_{KL} &= 0, \\ e^{(m}_{[IJ} e^{n)}_{KL]} + \tfrac{1}{24} e^m_A \Gamma^{IJKL}_{AB} e^n_B &= 0. \end{aligned} \tag{5}$$

The 248-bein and the new connection fields are subject to a "vielbein postulate" similar to the usual vielbein postulate stating the covariant constancy of the vielbein w r.t. to generally covariant and Lorentz covariant derivative:

$$\begin{aligned} (\partial_\mu - B^n_\mu \partial_n) e^m_{IJ} + \partial_n B_\mu{}^n e^m_{IJ} + \partial_n B_\mu{}^m e^n_{IJ} + 2\, Q_\mu{}^K{}_{[I} e^m_{J]K} + P^A_\mu \Gamma^{IJ}_{AB} e^B_m &= 0, \\ (\partial_\mu - B^n_\mu \partial_n) e^m_A + \partial_n B_\mu{}^m e^n_A + \partial_n B_\mu{}^n e^m_A + \tfrac{1}{4} Q^{IJ}_\mu \Gamma^{IJ}_{AB} e^m_B - \tfrac{1}{2}\Gamma^{IJ}_{AB} P^B_\mu e^m_{IJ} &= 0, \\ \partial_m e^n_{IJ} + 2\, Q_m{}^K{}_{[I} e^n_{J]K} + P^A_m \Gamma^{IJ}_{AB} e^n_B &= 0, \\ \partial_m e^n_A + \tfrac{1}{4} Q^{IJ}_m \Gamma^{IJ}_{AB} e^n_B - \tfrac{1}{2}\Gamma^{IJ}_{AB} P^B_m e^n_{IJ} &= 0. \end{aligned} \tag{6}$$

Like (3)–(5), these relations are E_8 covariant. It must be stressed, however, that the full theory of course does not respect E_8 invariance. A puzzling feature of (6) is that the covariantization w.r.t. an affine connection is "missing" in these equations, even though the theory is still invariant under $d = 11$ coordinate transformations. One can now show that the supersymmetry variations of $d = 11$ supergravity can be entirely expressed in terms of these new variables (and their fermionic partners).

The reduction of $d = 11$ supergravity to three dimensions yields $d = 3$, $N = 16$ supergravity (Marcus & Schwarz 1983), and is accomplished rather easily, since no duality redefinitions are needed any more, unlike in (Cremmer & Julia 1979). The propagating bosonic degrees of freedom in three dimensions are all scalar, and combine into a matrix $\mathcal{V}(x)$, which is an element of a non-compact $E_{8(+8)}$/SO(16) coset space, and whose dynamics is governed by a non-linear σ-model coupled to $d = 3$ gravity. The identification of the 248-bein with the a-model field $\mathcal{V} \in E_8$ is

given by

$$e^m_{IJ} = \tfrac{1}{60}\,\mathrm{Tr}(Z^m \mathcal{V} X^{IJ} \mathcal{V}^{-1}) \qquad e^m_A = \tfrac{1}{60}\,\mathrm{Tr}(Z^m \mathcal{V} Y^A \mathcal{V}^{-1}) \tag{7}$$

where X^{IJ} and Y^A are the compact and non-compact generators of E_8, respectively, and where the Z_m for $m = 3, \ldots, 10$ are eight non-compact commuting generators obeying $\mathrm{Tr}(Z^m Z^n) = 0$ for all m and n (the existence of eight such generators is a consequence of the fact that the coset space $E_{8(+8)}/S0(16)$ has real rank 8 and therefore admits an eight-dimensional maximal flat and totally geodesic submanifold (Helgason 1962). This reduction provides a "model" for the exceptional geometry, where the relations (3)–(6) can be tested by means of completeness relations for the E_8 Lie algebra generators in the adjoint representation. Of course, this is not much of a test since all dependence on the internal coordinates is dropped in (7), and the terms involving B^m_μ disappear altogether. It would be desirable to find other "models" with non-trivial dependence on the internal coordinates. The only example of this type so far is provided by the S^7 truncation of $d = 11$ supergravity for the SO(1, 3) × SU(8) invariant version of $d = 11$. supergravity (deWit & Nicolai 1987).

3. More symmetries

The emergence of hidden symmetries of the exceptional type in extended supergravities (Cremmer & Julia 1978) was a remarkable and, at the time, quite unexpected discovery. It took some effort to show that the general pattern continues when one descends to $d = 2$ and that the hidden symmetries become infinite dimensional (Julia 1981; Julia 1982, 1984; Breitenlohner & Maison 1987; Breitenlohner Maison & Gibbons 1988; Nicolai 1991; Julia & Nicolai 1996; Bernard & Julia 1997) generalizing the Geroch group of general relativity (Geroch 1972; Kinnersley & Chitre 1977). As we will see, even the coset structure remains, although the mathematical objects one deals with become a lot more delicate. The fact that the construction described above works with a 4 + 7 and 3 + 8 split of the indices suggests that we should be able to go even further and to construct versions of $d = 11$ supergravity with infinite dimensional tangent space symmetries, which would be based on a 2 + 9 or even a 1 + 10 split of the indices. This would also be desirable in view of the fact that the new versions are "simple" only in their internal sectors. The general strategy is thus to further enlarge the internal sector by absorbing more and more degrees of freedom into it, such that in the final step corresponding to a 1 + 10 split, only an einbein is left in the low dimensional sector. Although the actual elaboration of these ideas has to be left to future work, I will try to give at least a flavor of some anticipated key features.

3.1 *Reduction to two dimensions*

Let us first recall some facts about dimensional reduction of maximal supergravity to two dimensions. Following the empirical rules of dimensional reduction one is led to predict $E_9 = E_8^{(1)}$ as a symmetry for the dimensional reduction of $d = 11$ supergravity to two dimensions (Julia 1981). This expectation is borne out by the existence of a linear system for maximal $N = 16$ supergravity in two dimensions (Nicolai

1987; Nicolai & Warner 1989) (see Maison (1978); Belinski & Zakharov (1978); Breitenlohner & Maison (1987) for the bosonic theory). The linear system requires the introduction of an extra "spectral" parameter t, and the extension of the σ-model matrix $\mathcal{V}(x)$ to a matrix $\widehat{\mathcal{V}}(x; t)$ depending on this extra parameter t, as is generally the case for integrable systems in two dimensions. An unusual feature is that, due to the presence of gravitational degrees of freedom, this parameter becomes coordinate dependent, i.e. we have $t = t(x; w)$, where w is an integration constant, sometimes referred to as the "constant spectral parameter" whereas t itself is called the "variable spectral parameter".

Here, we are mainly concerned with the symmetry aspects of this system, and with what they can teach us about the $d = 11$ theory itself. The coset structure of the higher dimensional theories has a natural continuation in two dimensions, with the only difference that the symmetry groups are infinite dimensional. This property is manifest from the transformation properties of the linear system matrix $\widehat{\mathcal{V}}$, with a global affine symmetry acting from the left, and a local symmetry corresponding to some "maximal compact" subgroup acting from the right:

$$\widehat{\mathcal{V}}(x;t) \longrightarrow g(w)\widehat{\mathcal{V}}(x;t)h(x;t). \tag{8}$$

Here $g(w) \in E_9$ with affine parameter w, and the subgroup to which $h(x; t)$ belongs is characterized as follows (Julia 1982; Breitenlohner & Maison 1987). Let τ be the involution characterizing the coset space $E_{8(+8)}/\mathrm{SO}(16)$: then $h(t) \in \mathrm{SO}(16)^{\infty}_{\varepsilon}$ is defined to consist of all τ^{∞} invariant elements of E_9, where the extended involution τ^{∞} is defined by $\tau^{\infty}(h(t)) = \tau h(\varepsilon t^{-1})$, with $\varepsilon = +1$ (or -1) for a Lorentzian (Euclidean) worldsheet. For $\varepsilon = 1$, which is the case we are mainly interested in, we will write $\mathrm{SO}(16)^{\infty}$ $\mathrm{SO}(16)^{\infty}_{\varepsilon}$. We also note that $\mathrm{SO}(16)^{\infty}_{\varepsilon}$ is different from the affine extension of SO(16) for either choice of sign.

What has been achieved by the coset space description is the following: by representing the "moduli space of solutions" $\mathcal{M}$ (of the bosonic equations of motion of $d = 11$ supergravity with nine commuting space-like Killing vectors) as

$$\mathcal{M} = \frac{\text{solutions of field equations}}{\text{diffeomorphisms}} = \frac{E_9}{\mathrm{SO}(16)^{\infty}} \tag{9}$$

we have managed to endow this space, which a priori is very complicated, with a group theoretic structure, that makes it much easier to handle. In particular, the integrability of the system is directly linked to the fact that $\mathcal{M}$ possesses an infinite dimensional "isometry group" E_9. The introduction of infinitely many gauge degrees of freedom embodied in the subgroup $\mathrm{SO}(16)^{\infty}$ linearizes and localizes the action of this isometry group on the space of solutions. Of course, in making such statements, one should keep in mind that a mathematically rigorous construction of such spaces is a thorny problem. This is likewise true for the infinite dimensional groups* and their associated Lie algebras; the latter being infinite dimensional vector spaces, there are myriad ways of equiping them with a topology. We here take the liberty of ignoring

* For instance, the Geroch group can be defined rigorously to consist of all maps from the complex w plane to $SL(2, \mathbf{R})$ with meromorphic entries. With this definition, one obtains all multisoliton solutions of Einstein's equations, and on this solution space the group acts transitively by construction.

these subleties, not least because these spaces ultimately will have to be "quantized" anyway.

There is a second way of defining the Lie algebra of SO $(16)^{\infty}_{\varepsilon}$ which relies on the Chevalley-Serre presentation. Given a finite dimensional non-compact Lie group G with maximal compact subgroup H, a necessary condition for this prescription to work is that dim $H = \frac{1}{2}$ (dim G – rank G), and we will subsequently extend this prescription to the infinite Lie group. Let us first recall that any (finite or infinite dimensional) Kac Moody algebra is recursively defined in terms of multiple commutators of the Chevalley generators subject to certain relations (Bourbaki 1968; Kac 1930). More specifically, given a Cartan matrix Au and the associated Dynkin diagram, one starts from a set of $sl(2, \mathbf{R})$ generators $\{e_i, f_i, h_I\}$, one for each node of the Dynkin diagram, which in addition to the standard $sl(2, \mathbf{R})$ commutation relations

$$\begin{aligned} [h_i, h_j] &= 0 \qquad [e_i, f_j] = \delta_{ij} h_j, \\ [h_i, e_j] &= A_{ij} e_j \quad [h_i, f_j] = -A_{ij} f_j \end{aligned} \tag{10}$$

Are subject to the multilinear Serre relations

$$[e_i, [e_i, \ldots [e_i, e_j] \ldots]] = 0 \qquad [f_i, [f_i, \ldots [f_i, f_j] \ldots]] = 0 \tag{11}$$

where the commutators are $(1 - A_{ij})$-fold ones. The Lie algebra is then by definition the linear span of all multiple commutators which do not vanish by virtue of these relations.

To define the subalgebra $\mathrm{SO}(16)^{\infty}_{\varepsilon}$, we first recall that the Chevalley involution θ is defined by

$$\theta(e_i) = -f_i \qquad \theta(f_i) = -e_i \qquad \theta(h_i) = -h_i. \tag{12}$$

This involution, like the ones to be introduced below, leaves invariant the defining relations (18) and (19) of the Kac Moody algebra, and extends to the whole Lie algebra via the formula $\theta([x, y]) = [\theta(x), \theta(y)]$. It is not difficult to see that, for E_8 (and also for $sl(n, \mathbf{R})$), we have $\tau = \theta$, and the maximal compact subalgebras defined above correspond to the subalgebras generated by the multiple commutators of the θ invariant elements $(e_i - f_i)$ in both cases. The trick is now to carry over this definition to the affine extension, whose associated Cartan matrix has a zero eigenvalue. To do this, however, we need a slight generalization of the above definition; for this purpose, we consider involutions ω that can be represented as products of the form

$$\omega = \theta \cdot s \tag{13}$$

where the involution s acts as

$$s(e_i) = s_i e_i \qquad s(f_i) = s_i f_i \qquad s(h_i) = h_i \tag{14}$$

with $s_i = \pm 1$. It is important that different choices of s_i do not necessarily lead to inequivalent involutions (the general problem of classifying the involutive automorphisms of infinite dimensional Kac Moody algebras has so far not been completely solved, see e.g. (Levstein 1988; Bausch & Rousseau 1989)[†]). In particular

[†] I am very grateful to C. Daboul for helpful discussions on this topic.

for E_9, which is obtained from E_8 by adjoining another set $\{e_0, f_0, h_0\}$ of Chevalley generators, we take $s_i = 1$ for all $i \geq 1$, whereas $s_0 = \varepsilon$, with ε as before, i.e. $\epsilon = +1$ (or -1) for Lorentzian (Euclidean) worldsheet. Thus, on the extended Chevalley generators,

$$\omega(e_0) = -\varepsilon f_0 \qquad \omega(f_0) = -\varepsilon e_0 \qquad \omega(h_0) = -h_0. \tag{15}$$

With this choice, the involution ω coincides with the involutions defined before for the respective choices of ε, i.e. $\omega = \tau^\infty$, and therefore the invariant subgroups are the same, too. For $\varepsilon = 1$, the involution ω defines an infinite dimensional "maximal compact" subalgebra consisting of all the negative norm elements w.r.t. to the standard bilinear form

$$\langle e_i | f_j \rangle = \delta_{ij} \qquad \langle h_i | h_j \rangle = A_{ij} \tag{16}$$

(the norm of any given multiple commutator can be determined recursively from the fundamental relation $\langle [x, y] | z \rangle = \langle x | [y, z] \rangle$). The notion of "compactness" here is thus algebraic, not topological: the subgroup $SO(16)^\infty$ will not be compact in the topological sense (recall the well known example of the unit ball in an infinite dimensional Hilbert space, which is bounded but not compact in the norm topology). On the other hand, for $\varepsilon = -1$, the group $SO(16)^\infty_\varepsilon$ is not even compact in the algebraic sense, as $e_0 + f_0$ has positive norm. However, this is in accord with the expectation that $SO(16)^\infty_\varepsilon$ should contain the (non-compact) group SO(1,8) rather than SO(9) if one of the compactified dimensions is time-like.

3.2 2 + 9 *split*

Let us now consider the extension of the results described in section 2 to the situation corresponding to a 2 + 9 split of the indices. Elevating the local symmetries of $N = 16$ supergravity from two to eleven dimensions would require the existence of yet another extension of the theory, for which the Lorentz group SO(1,10) is replaced by $SO(1, 1) \times SO(16)^\infty$; the subgroup $SO(16)^\infty$ can be interpreted as an extension of the transverse group SO(9) in eleven dimensions. Taking the hints from (1), we would now decompose the elfbein into a zweibein and nine Kaluza Klein vectors B^m_μ (with $m = 2, \ldots, 10$). The remaining internal neunbein would have to be replaced by an "Unendlichbein" $(e^m_{IJ}(x;\ t),\ e^m_A(x;\ t))$, depending on a spectral parameter t, necessary to parametrize the infinite dimensional extension of the symmetry group. However, in eleven dimensions, there is no anolog of the dualization mechanism, which would ensure that despite the existence of infinitely many dual potentials, there are only finitely many physical degrees of freedom. This indicates that if the construction works it will take us beyond $d = 11$ supergravity.

Some constraints on the geometry can be deduced from the requirement that in the dimensional reduction to $d = 2$, there should exist a formula analogous to (7), but with $\mathcal{V}$ replaced by the linear system matrix $\widehat{\mathcal{V}}$, or possibly even the enlarged linear system of (Julia & Nicolai 1996). Evidently, we would need a ninth nilpotent generator to complement the Z^ms of (7); an obvious candidate is the central charge generator c, since it obeys $\langle c | c \rangle = \langle c | Z^m \rangle = 0$ for all $m = 3, \ldots, 10$. The parameter t, introduced somewhat ad hoc for the parametrization of the unendlichbein, must obviously coincide with the spectral parameter of the d 2 theory, and

the generalized "unendlichbein postulate" should evidently reduce to the linear system of $d = 2$ supergravity in this reduction. To write it down, we need to generalize the connection coefficients appearing in the linear system. The latter are given by

$$\mathcal{Q}_\mu^{IJ} = Q_\mu^{IJ} + \cdots \qquad \mathcal{P}_\mu^A = \frac{1+t^2}{1-t^2} P_\mu^A + \frac{2t}{1-t^2} \varepsilon_{\mu\nu} P^{\nu A} + \cdots \tag{17}$$

with Q_μ^{IJ} and $P_{A\mu}$ as before; the dots indicate t dependent fermionic contributions which we omit. A very important difference with section 2, where the tangent space symmetry was still finite dimensional, is that the Lie algebra of $SO(16)^\infty$ also involves the P's, and not only the Q's. More specifically, from the t dependence of the dimensionally reduced connections in (17) we infer that theconnections ($\mathcal{Q}_M^{IJ}(x;\ t)$, $\mathcal{P}_M^A(x;\ t)$) constitute an $SO(16)^\infty$ (and not an E_9) gauge connection. This means that the covariantizations in the generalized vielbein postulate are now in precise correspondence with the local symmetries, in contrast with the relations (6) which look E_8 covariant, whereas the full theory is invariant only under local SO(16).

To write down an ansatz, we put

$$\mathcal{D}_\mu := \partial_\mu - B_\mu^n \partial_n + \cdots \tag{18}$$

where the dots stand for terms involving derivatives of the Kaluza Klein vector fields. Then the generalization of (6) should read

$$\begin{aligned}
&\mathcal{D}_\mu e_{IJ}^m(t) + 2\,\mathcal{Q}_\mu{}^K{}_{[I}(t) e_{J]K}^m(t) + \mathcal{P}_\mu^A(t)\Gamma_{AB}^{IJ} e_B^m(t) = 0,\\
&\mathcal{D}_\mu e_A^m(t) + \tfrac{1}{4}\mathcal{Q}_\mu^{IJ}(t)\Gamma_{AB}^{IJ} e_B^m(t) - \tfrac{1}{2}\Gamma_{AB}^{IJ}\mathcal{P}_\mu^B(t) e_{IJ}^m(t) = 0,\\
&\partial_m e_{IJ}^n(t) + 2\,\mathcal{Q}_m{}^K{}_{[I}(t) e_{J]K}^n(t) + \mathcal{P}_m^A(t)\Gamma_{AB}^{IJ} e_B^n(t) = 0,\\
&\partial_m e_A^n(t) + \tfrac{1}{4}\mathcal{Q}_m^{IJ}(t)\Gamma_{AB}^{IJ} e_B^n(t) - \tfrac{1}{2}\Gamma_{AB}^{IJ}\mathcal{P}_m^B(t) e_{IJ}^n(t) = 0.
\end{aligned} \tag{19}$$

Of course, the challenge is now to find explicit expressions for the internal components $\mathcal{Q}_m^J(x;\ t)$ and $\mathcal{P}^A{}_m(x;\ t)$, such that (19) can be interpreted as a $d = 11$ generalization of the linear system of dimensionally reduced supergravity. Another obvious question concerns the fermionic partners of the unendlichbein: in two dimensions, the linear system matrix contains all degrees of freedom, including the fermionic ones, and the local $N = 16$ supersymmetry can be bosonized into a local $SO(16)^\infty$ gauge transformation (Nicolai & Warner 1989). Could this mean that there is a kind of bosonization in eleven dimensions or M-Theory? This idea may not be as outlandish as it sounds because a truly pregeometrical theory might be subject to a kind of "pre-statistics", such that the distinction between bosons and fermions arises only through a process of spontaneous symmetry breaking.

4. Yet more symmetries?

In 1982, B. Julia conjectured that the dimensional reduction of maximal supergravity to one dimension should be invariant under a further extension of the E-series, namely (a non-compact form of) the hyperbolic Kac Moody algebra E_{10} obtained by adjoining another set $\{e_{-1}, f_{-1}, h_{-1}\}$ of Chevalley generators to those of E_9 (Julia

1985)‡. As shown in Nicolai (1992), the last step of the reduction requires a null reduction if the affine symmetry of the $d = 2$ theory is not to be lost. The reason is that the infinite dimensional affine symmetries of the $d = 2$ theories always involve dualizations of the type

$$\partial_\mu \varphi = \varepsilon_{\mu\nu} \partial^\nu \tilde{\varphi} \tag{20}$$

(in actual fact, there are more scalar fields, and the duality relation becomes non-linear, which is why one ends up with infinitely many dual potentials for each scalar degree of freedom). Dimensional reduction w r.t. to a Killing vector ξ^μ amounts to imposing the condition $\xi^\mu \partial_\mu \equiv 0$ on *all* fields, including dual potentials. Hence,

$$\xi^\mu \partial_\mu \varphi = 0, \quad \xi^\mu \partial_\mu \tilde{\varphi} \equiv \eta^\mu \partial_\mu \varphi = 0 \tag{21}$$

where $\eta^\mu \equiv \varepsilon^{\mu\nu}\ \xi_\nu$. If ξ^μ and η^μ are linearly independent, this constraint would force all fields to be constant, which is clearly too strong a requirement. Hence we must demand that ξ^μ and η^μ are collinear, which implies

$$\xi^\mu \xi_\mu = 0, \tag{22}$$

i.e. the Killing vector must be null. Starting from this observation, it was shown in Nicolai (1992) that the Matzner Misner $sl(2, \mathbf{R})$ symmetry of pure gravity can be formally extended to an $sl(3, \mathbf{R})$ algebra in the reduction of the vierbein from four to one dimensions. Combining this $sl(3, \mathbf{R})$ with the Ehlers $sl(2, \mathbf{R})$ of ordinary gravity, or with the E_8 symmetry of maximal supergravity in three dimensions, one is led to the hyperbolic algebra $\mathcal{F}_3$ (Feingold & Frenkel 1993) for ordinary gravity, and to E_{10} for maximal supergravity. The transformations realizing the action of the Chevalley generators on the vierbein components can be worked out explicitly, and the Serre relations can be formally verified (Nicolai 1992) (for E_{10}, this was shown more recently in Mizoguchi (1970).

There is thus some evidence for the emergence of hyperbolic Kac Moody algebras in the reduction to one null dimension, but the difficult open question that remains is what the configuration space is on which this huge symmetry acts. This space is expected to be much bigger than the coset space (9). Now, already for the $d = 2$ reduction there are extra degrees of freedom that must be taken into account in addition to the propagating degrees of freedom. Namely, the full moduli space involveing all bosonic degrees of freedom should also include the moduli of the zweibein, which are not contained in (9). For each point on the worldsheet, the zweibein is an element of the coset space GL(2, **R**)/SO(1, 1); although it has no local degrees of freedom any more, it still contains the global information about the conformal structure of the world sheet $\sum$. Consequently, we should consider the Teichmüller space

$$\mathcal{T} = \frac{\{e_\mu^\alpha(x) | x \in \Sigma\}}{\mathrm{SO}(1,1) \times \mathrm{Weyl}(\Sigma) \times \mathrm{Diff}_0(\Sigma)} \tag{23}$$

as part of the configuration space of the theory (see Verlinde (1990) for a detailed description of $\mathcal{T}$). In fact, we should even allow for arbitrary genus of the worldsheet,

‡The existence of a maximal dimension for supergravity (Nahm 1978) would thus be correlated with the existence of a "maximally extended" hyperbolic Kac Moody algebra, which might thus explain the occurrence of maximum spin 2 for massless gauge particles in nature.

and replace $\mathcal{T}$ by the "universal Teichmüller space" $\widetilde{\mathcal{T}}$. This infinite dimensional space can be viewed as the configuration space of non-perturbative string theory (Friedan & Shenker 1987). For the models under consideration here, however, even $\widetilde{\mathcal{T}}$ is not big enough, as we must also take into account the dilaton ρ and the non-propagating Kaluza Klein vector fields in two dimensions. For the former, a coset space description was proposed in Julia & Nicolai (1996). On the other hand, the Kaluza Klein vectors and the cosmological constant they could generate in two dimensions have been largely ignored in the literature. Even if one sets their field strengths equal to zero (there are arguments that the Geroch group, and hence infinite duality symmetries, are incompatible with a nonzero cosmological constant in two dimensions), there still remain topological degrees of freedom for higher genus world sheets.

The existence of inequivalent conformal structures is evidently important for the null reductions, as the former are in one-to-one correspondence with the latter. Put differently, the inequivalent null reductions are precisely parametrized by the space (23). The extended symmetries should thus not only act on one special null reduction (set of plane wave solutions of Einstein's equations), but relate different reductions. Indeed, it was argued in Mizoguchi (1997) that, for a toroidal worldsheet, the new $sl(2, \mathbf{R})$ transformations associated with the over-extended Chevalley generators change the conformal structure, but only for non-vanishing holonomies of the Kaluza Klein vector fields on the worldsheet. This indicates that the non-trivial realization of the hyperbolic symmetry requires the consideration of non-trivial worldsheet topologies. The dimensionally reduced theory thereby retains a memory of its two-dimensional ancestor. It is therefore remarkable that, at least for isomonodromic solutions of Einstein's theory, the $d = 2$ theory exhibits a factorization of the equations of motion akin to, but more subtle than the holomorphic factorization of conformal field theories (Korotkin & Nicolai 1995). In other words, there may be a way to think of the $d = 2$ theory as being composed of two chiral halves just as for the closed string. Consequently, a truncation to one null dimension may not be necessary after all if the theory factorizes all by itself.

In summary, what we are after here is a group theoretic unification of all these moduli spaces that would be analogous to (9) above, and fuse the matter and the topological degrees of freedom. No such description seems to be available for (23) (or $\widetilde{\mathcal{T}}$), and it is conceivable that only the total moduli space $\widetilde{\mathcal{M}}$ containing both $\mathcal{M}$ and $\widetilde{\mathcal{T}}$ as well as the dilaton and the Kaluza Klein, and perhaps even the fermionic, degrees of freedom is amenable to such an interpretation. Extrapolating the previous results, we are thus led to consider coset spaces E_{10}/H with $\mathrm{SO}(16)^{\infty} \subset H \subset E_{10}$. As before, the introduction of the infinitely many spurious degrees of freedom associated with the gauge group H would be necessary in order to "linearize" the action of E_{10}.

What are the choices for H? One possibility would be to follow the procedure of the foregoing section, and to define $H = \mathrm{SO}(16)^{\infty\infty} \subset E_{10}$ in analogy with $\mathrm{SO}(16)^{\infty} \subset E_9$ by taking its associated Lie algebra to be the linear span of all ω invariant combinations of E_{10} Lie algebra elements. To extend the affine involution to the full hyperbolic algebra, we would again invoke (13), setting $\varepsilon = +1$ in (15) (since we now assume the worldsheet to be Lorentzian), which leaves us with the two choices $s_{-1} = \pm 1$. For $s_{-1} = +1$ we would get the "maximal compact" subalgebra of E_{10}, corresponding to the compactification of ten spacelike dimensions. A subtlety here is that a definition in terms of the standard bilinear form is no longer possible,

unlike for affine and finite algebras, as this would now also include part of the Cartan subalgebra of E_{10}: due to the existence of a negative eigenvalue of the E_{10} Cartan matrix, there exists a negative norm element $\sum_i n_i h_i$ of the Cartan subalgebra, which would have to be excluded from the definition of H (cf. the footnote on p. 438 of (Julia & Nicolai 1996). The alternative choice $s_{-1} = -1$ would correspond to reduction on a 9 + 1 torus.

However, for the null reduction advocated here, physical reasoning motivates us to propose yet another choice for H. Namely, in this case, H should contain the group ISO(9) $\subset$ SO(1, 10) leaving invariant a null vector in eleven dimensions (Julia & Nivolai 1995). To identify the relevant parabolic subgroup of E_{10}, which we denote by ISO(16)$^{\infty}$, we recall (Nicolai 1992) that the over-extended Chevalley generators correspond to the matrices

$$e_{-1} = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad f_{-1} = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad h_{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix} \tag{24}$$

in a notation where we only write out the components acting on the 0, 1, 2 components of the elfbein, with all other entries vanishing. Evidently, we have $h_{-1} = d - c_-$ with

$$d = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \quad c_- = -\frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{25}$$

where d is the scaling operator on the dilaton ρ, and c_- is the central charge, alias the "level counting operator" of E_{10}, obeying $[c_-, e_{-1}] = -e_{-1}$ and $[c_-, f_{-1}] = + f_{-1}$ (and having vanishing commutators with all other Chevalley generators). Writing

$$c_\pm := -\frac{1}{2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \pm \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{26}$$

we see that the first matrix on the right scales the conformal factor, generating Weyl transformations (called Weyl($\sum$) in (23)) on the zweibein, while the second generates the local SO(1,1) Lorentz transformations. In a lightcone basis, these symmetries factorize on the zweibein, which decomposes into two chiral einbeine. Consequently, Weyl transformations and local SO(1, 1) can be combined into two groups SO(1, 1)$_\pm$ with respective generators $c_\pm$, and which act separately on the chiral einbeine. One of these, SO(1, 1)$_-$ (generated by c_-), becomes part of E_{10}. The other, SO(1, 1)$_+$, acts on the residual einbein and can be used to eliminate it by gauging it to one. Since $c_\pm$ acts in the same way on the conformal factor, we also recover the result of Julia (1982).

We wish to include both ISO(9) and SO(1, 1)$_-$ into the enlarged local symmetry H = ISO(16)$^{\infty}$, and thereby unify the longitudinal symmetries with the "transversal" group SO(16)$^{\infty}$ discussed before. Accordingly, we define ISO(16)$^{\infty}$ to be the algebra generated by the SO (16)$^{\infty}$ Lie algebra together with c_- and e_{-1}, as well as all their nonvanishing multiple commutators. The "classical" configuration space of M-Theory should then be identified with the coset space

$$\widetilde{\mathcal{M}} = \frac{E_{10}}{\mathrm{ISO}(16)^{\infty}}\,. \tag{27}$$

Of course, we will have to worry about the fate of these symmetries in the quantum theory. Indeed, some quantum version of the symmetry groups appearing in (27) must be realized on the Hilbert space of third quantized $N = 16$ supergravity, such that E_{10} becomes a kind of spectrum generating (rigid) symmetry on the physical states, while the gauge group $\mathrm{ISO}(16)^{\infty}$ gives rise to the constraints defining them. Because "third quantization" here is analogous to the transition from first quantized string theory to string field theory, the latter would have to be interpreted as multi-string states in some sense (cf. Witten (1986)) for earlier suggestions in this direction; note also that the coset space (27) is essentially generated by half of E_{10}, so there would be no "anti-string states". According to Font *et al.* (1990); Rey (1991); Sen (1993, 94); Schwarz & Sen (1994); Duff & Khuri (1994); Giveon *et al.* (1994); Hull & Townsend (1995); Witten (1995); Kachru & Vafa (1995); Schwarz (1995, 96); Duff (1996); Horava (1996) the continuous duality symmetries are broken to certain discrete subgroups over the integers in the quantum theory. Consequently, the quantum configuration space would be the left coset

$$\widetilde{\mathcal{F}} = E_{10}(\mathbf{Z})\backslash\widetilde{\mathcal{M}}$$

and the relevant partition functions would have to be new kinds of modular forms defined on $\widetilde{\mathcal{F}}$. However, despite recent advances (Bakas 1996; Sen 1995), the precise significance of the (discrete) "string Geroch group" remains a mystery, and it is far from obvious how to extend the known results and conjectures for finite dimensional duality symmetries to the infinite dimensional case (these statements apply even more to possible discrete hyperbolic extensions; see, however, (Mizoguchi 1997; Gebert & Mizoguchi 1997). Moreover, recent work (Korotkin & Samtleben 1997) confirms the possible relevance of quantum groups in this context (in the form of "Yangian doubles").

Returning to our opening theme, more should be said about the 1 + 10 split, which would lift up the $\mathrm{SO}(1, 1)_{+} \times \mathrm{ISO}(16)^{\infty}$ symmetry, and the "bein" which would realize the exceptional geometry alluded to in the introduction, and on which $\mathrm{IS0}(16)^{\infty}$ would act as a generalized tangent space symmetry. However, as long as the 2 + 9 split has not been shown to work, and a manageable realization is not known for either E_{10} or $\mathrm{ISO}(16)^{\infty}$, we must leave the elaboration of these ideas to the future. It could well prove worth the effort.

Acknowledgements

The results described in section 2 are based on work done in collaboration with S. Melosch. I would also like to thank C. Daboul, R. W. Gebert, H. Samtleben and P. Slodowy for stimulating discussions and comments.

References

Ashtekar, A. 1986, *Phys. Rev. Lett.*, **57**, 2244.
Baake, M., Reinicke, P., Rittenberg, V. 1985, *J. Math. Phys.*, **26**, 107.

Bausch, J., Rousseau, G. 1989, *Algêbres de Kac-Moody eines: autornorphismes et formes reelles*, Rev. de 1'Institut **E**. Cartan 11.
Bakas, I. 1996, hep-th/941 1118, hep-th/9606030.
Banks, T., Fischler, W., Shenker, S. H., Susskind, L. 1997, *Phys. Rev*., **D55**, 5112.
Banks, T. 1997, hep-th/9710231.
Belinskii, S., Zakharov, V. 1978, *Sov. Phys. JETP*, **48**, 985.
Bergshoeff, E., Sezgin, E., Townsend, P. K. 1987, *Phys. Lett*., **B189**, 75; 1988, *Ann. of Phys*., **185**, 330.
Bernard, D., Julia, B. 1997, hep-th/9712254.
Bigatti, L., Susskind, L. 1997, hep-th/9712072.
Breitenlohner, P., Maison, D. 1987, *Ann. Inst. H. Poincare*, **46**, 215.
Breitenlohner, P., Maison, D., Gibbons, G. W. 1988, *Comm. Math. Phys*., **120**, 295.
Bourbaki, N. 1968, *Groupes et Algebres de Lie*, chapters 4–6 (Hermann, Paris).
Claudson, M., Halpern, M. 1985, *Nucl. Phys*., **B250**, 689.
Cremmer, E., Julia, B., Scherki, J. 1978, *Phys. Lett*., **76B**, 409.
Cremmer, E., Julia, B. 1978, *Phys. Lett*., **B80**, 48; 1979, *Nucl. Phys*., **B159**, 141.
de Wit, B., Nicolai, H. 1985, *Phys. Lett*., **B155**, 47; 1986, *Nucl. Phys*., **B274**, 363.
de Wit, B., Nicolai, H. 1987, *Nucl. Phys*., **281** 211.
de Wit, B. 1997, hep-th/9701169.
de Wit, B., Hoppe, J., Nicolai, H. 1988, *Nucl. Phys*., **B305**, 545.
Duff, M. J., Khuri, R. 1994, *Nucl. Phys*., **B411**, 473.
Duff, M. J. 1996, *Int. J. Mod. Phys*., **All**, 5623.
Elitzur, S., Giveon, A., Kutasov, D., Rabinovici, E. 1997, hep-th/9707217.
Feingold, A., Frenkel, I. B. 1983, *Math. Ann*., **263**, 87.
Flume, R. 1985, *Ann. Phys*., **164**, 189.
Friedan, D., Shenker, S. 1987, *Nucl. Phys*., **B281**, 509.
Fröhlich, J., Hoppe, J. 1997, hep-th/9701119.
Font, A., Ibáñez, L., Lüst, D., Quevedo, F. 1990, *Phys. Lett*., **B249**, 35.
Geroch, R. 1972, *J. Math. Phys*., **13**, 394.
Gebert, R. W., Mizoguchi, S. 1997, hep-th/9712078.
Giddings, S., Wolpert, S. 1987, *Commun. Math. Phys*., **109**, 177.
Giveon, A., Porrati, M., Rabinovici, E. 1994, *Phys. Rep*., **244**, 77.
Green, M. B., Gutperle, M. 1997, hep-th/9711107.
Halpern, M. B., Schwartz, C. 1997, hep-th/9712133.
Helgason, S. 1962, *Differential Geometry and Symmetric Spaces* (Academic Press).
Horava, P., Witten, E. 1996, *Nucl. Phys*., **B460**, 506.
Hoppe, J. 1997, hep-th/9709132.
Hull, C. M., Townsend, P. K. 1995, *Nucl. Phys*., **B438**, 109.
Julia, B. 1982, in *Unified Theories and Beyond*, Proc. 5th Johns Hopkins Workshop on Current Problems in Particle Theory, Johns Hopkins University, Baltimore, 1984 in *Vertex Operators in Mathematics and Physics* (eds.) Lepowsky, J., Mandelstam S. and Singer, I. (Springer Verlag)
Julia, B. 1981, in *Superspace and Supergra*^{y}*ity*, (eds.) S. W. Hawking and M. Rocek (Cambridge University Press).
Julia, B. 1985, in *Lectures in Applied Mathematics*, Vol. 21, 355.
Julia, B., Nicolai, H. 1995, *Nucl. Phys*., **B439**, 291.
Julia, B., Nicolai, H. 1996, *Nucl. Phys*., **B482**, 431.
Kac, V. G. 1990, *Infinite Dimensional Lie Algebras*, 3rd edition (Cambridge University Press).
Kachru, S., Vafa, C. 1995, *Nucl. Phys*., **B450**, 69.
Kinnersley, W., Chitre, D. M. 1977, *J. Math. Phys*., **18**, 1538.
Korotkin, D., Nicolai, H. 1995, *Phys. Rev. Lett*., **74**, 1272.
Korotkin, D., Samtleben, H. 1997, hep-th/9710210.
Levstein, F. 1988, *J. of Algebra*, **114**, 489.
Mandelstam, S. 1973, *Nucl. Phys*., **B64**, 205.
Maison, D. 1978, *Phys. Rev. Lett*., **41**, 521.
Marcus, N., Schwarz, J. H. 1983, *Nucl. Phys*., **B228**, 145.
Melosch, S., Nicolai, H. 1997, hep-th/9709227.

Melosch, S. PhD Thesis, in preparation.
Mizoguchi, **S.** 1997, hep-th/9703160.
Nahm, W. 1978, *Nucl. Phys*., **B135**, 149.
Nicolai, H. 1987, *Phys. Lett*., **B187,** 363.
Nicolai, H. 1991, in *Recent Aspects of Quantum Fields*, (eds.) H. Mitter and H. Gausterer (Springer Verlag)
Nicolai, H. 1987, *Phys. Lett*., **B194,** 402.
Nicolai, H., Warner, N. P. 1989, *Comm. Math. Phys*., **125,** 384.
Nicolai, H., Korotkin, D., Samtleben, H. 1997, in *Quantum Fields and Quantum Space Time*, Series B: Physics Vol. 364, (eds.) G. t' Hooft, A. Jaffe, G. Mack, P. Mitter and R. Stora (Plenum Press), hep-th/9612065.
Nicolai, H. 1992, *Phys. Lett*., **B276,** 333.
Nicolai, H. 1994, *Nucl. Phys.,* **B414,** 299.
Polchinski, J. 1995, *Phys. Rev. Lett*., **75,** 4724.
Polchinski, J. 1996, hep-th/9611050.
Porrati, M., Rozenberg, A. 1997, hep-th/9708119.
Rey, S. J. 1991, *Phys. Rev*., **D43,** 526.
Samtleben, H., Korotkin, D. 1998, *Phys. Rev. Lett*., **80,** 14.
Schwarz, J., Sen, A. 1994, *Nucl. Phys*., **B411,** 35.
Schwarz, J. H. 1995, *Phys. Lett*., **B360,** 13; 1996, *Phys. Lett*., **B367,** 97.
Sen, A. 1993, *Phys. Lett*., **B303,** 22; 1994, *Int. J. Mod. Phys*., **A9,** 3707.
Sen, A. 1995, *Nucl. Phys.,* **B447,** 62.
Sethi, S., Stern, M. 1997, hep-th/9705046.
Townsend, P. K. 1995, *Phys. Lett*., **B350,** 184; hep-th/9612121.
Verlinde, H. 1990, *Nucl. Phys*., **B337,** 652.
Witten, E. 1986, *Mt. J. Mod. Phys*., **Al,** 39.
Witten, E. 1995, *Nucl. Phys.,* **B443,** 85.
Witten, E. 1996, *Nucl. Phys*., **B460,** 335.
Yi, P. 1997, hep-th/9704098.

J. Astrophys. Astr. (1999) **20**, 165–185

Observational Evidence for Massive Black Holes in the Centers of Active Galaxies

J. M. Moran & L. J. Greenhill, *Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge Mass 02138 USA*

J. R. Herrnstein, *National Radio Astronomy Observatory*

Abstract. Naturally occurring water vapor maser emission at 1.35 cm wavelength provides an accurate probe for the study of accretion disks around highly compact objects, thought to be black holes, in the centers of active galaxies. Because of the exceptionally fine angular resolution, 200 microarcseconds, obtainable with very long baseline interferometry, accompanied by high spectral resolution, < 0.1 km s^{-1}, the dynamics and structures of these disks can be probed with exceptional clarity. The data on the galaxy NGC 4258 are discussed here in detail. The mass of the black hole binding the accretion disk is $3{\cdot}9 \times 10^7$ $M_\odot$. Although the accretion disk has a rotational period of about 800 years, the physical motions of the masers have been directly measured with VLBI over a period of a few years. These measurements also allow the distance from the earth to the black hole to be estimated to an accuracy of 4 per cent. The status of the search for other maser/black hole candidates is also discussed.

Key words. Masers—black holes—accretion disks—very long baselines interferometry.

1. Introduction

The observational evidence for the existence of supermassive black holes (10^6 – 10^9 times the mass of the sun, $M_\odot$) in the centers of active galaxies has been accumulate-ing at an ever accelerating pace for the last few decades (e.g., Rees 1998; Blandford & Gehrels 1999). Seyfert (1943) first drew attention to a group of galaxies with unusual excitation conditions in their nuclei indicative of energetic activity. Among the twelve galaxies in his list was NGC 4258, which is the subject of much of this paper. Such galaxies, now known as galaxies with active galactic nuclei (AGN), have grown in membership and importance. Ironically, NGC 4258 no longer belongs to the class of Seyfert galaxies by modern classification standards (Heckman 1980), but it is still considered to have a mildly active galactic nucleus. Meanwhile, the study of AGN has become a major field in modem astrophysics. In the 1960s, galaxies with AGN were discovered with intense radio emission arising from jets of relativistic particles often extending far beyond the optical boundaries of the host galaxy. The central engine, the source of energy that powers such jets and other phenomena in the centers of galaxies, has long been ascribed to black holes (e.g., Salpeter 1964;

Blandford & Rees 1992). There are two sources of energy for these phenomena: the gravitational energy from material falling onto the black hole and the spin energy of the black hole itself (Blandford & Znajek 1977).

The direct evidence for black holes in AGN has come principally from observations of the motions of gas and stars in the extended environments of black holes. In the optical and infrared domains, the evidence for black holes from stellar measurements comes from an analysis of the velocity dispersion of stars as a function of distance from the dynamical centers of galaxies. In the case of our own Galactic center, the proper motions (angular velocities in the plane of the sky) of individual stars can be measured. These data show that there is a mass of about 2.6×10^6 $M_\odot$ within a volume of radius 0·01 pc (Genzel *et al.* 1997; Ghez *et al.* 1998). In addition, measurements by the Hubble Space Telescope of the velocity field of hydrogen gas in active galaxies indicate the presence of massive centrally condensed objects. Reviews of these data have been written by Kormendy & Richstone (1995), Faber (1999), Ho (1999), and others.

In the X-ray portion of the spectrum, there is compelling evidence for black holes in AGN from the detection of the highly broadened iron Kα line at 6·4 keV. The line is broadened by the gravitational redshift of gas as close as 3 Schwarzschild radii from the black hole. An example of an iron line profile in the galaxy MCG-6-30-15 is shown in Fig. 1 (Tanaka *et al.* 1995). The linear extent of the emission region cannot be determined directly by the X-ray telescope, so it is not possible to estimate directly the mass of the putative black hole. Detailed analysis of the line profile suggests that the black hole is spinning (e.g., Bromley, Miller & Pariev 1998).

In the radio regime, a new line of inquiry has given unexpectedly clear and compelling evidence for black holes: the discovery of water masers orbiting highly massive and compact central objects. With the aid of very long baseline interferometry (VLBI), which provides angular resolution as fine as 200 microarcseconds (μas) at a wavelength of 1.3 cm and spectral resolution of 0·1 kms^{-1} or less, the structure of accreting material around these central objects can be studied in detail. This paper describes the observations and the significance of these measurements of water masers in AGN. We begin with a brief description of cosmic masers and the interferometric techniques used to observe them.

2. Cosmic masers

Intense maser action in cosmic molecular clouds was discovered in 1965 (Weaver *et al.* 1965) from observations of OH, and later from H_2O, SiO, and CH_3OH. In the case of water vapor, the commonly observed masers emit in the 6_{16}–5_{23} transition at 22235 MHz (1.35 cm wavelength). Most masers have been found to be associated with one of two types of objects: newly formed stars or evolved stars (e.g., Reid & Moran 1988; Elitzur 1992). Although very distinct, they share the characteristic of having envelopes of outflowing gas and dust (silicate material). The pump source in all cases is thought to come in the form of either shock waves or infrared radiation. More recently, masers have been found in the spiral arms of nearby galaxies and in AGN.

Cosmic masers are similar to their laboratory counterparts on earth in that their intense radiation is produced by population inversion. However, cosmic masers are one-pass amplifiers and have little temporal or spatial coherence. The intensity of

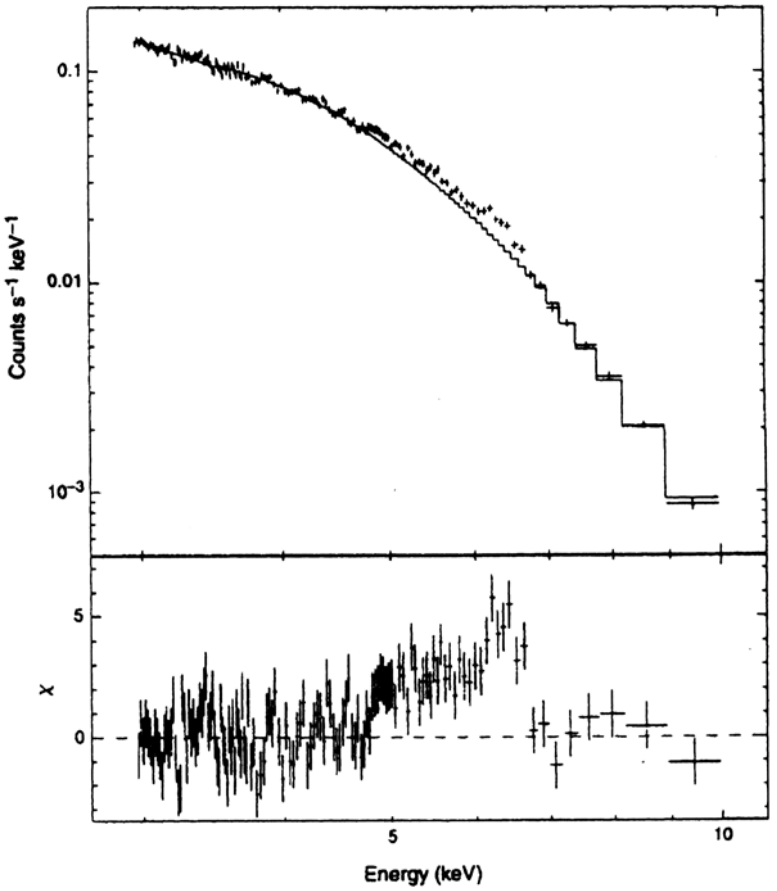

Figure 1. The X-ray spectrum of the galaxy MCG-6-30-15, observed by the Japanese ASCA satellite. The top panel shows the total spectrum with a model of the continuum emission fitted to the data outside the range of 5–7 keV. The bottom panel shows the residuals, which reveal a broad spectral feature attributed to the Fe Kα line at 6.4 keV. The line has a width of 100,000 km s^{-1}. The most extreme redshifted part is thought to arise from gas at a radius of about 3 Schwarzschild radii. (From Tanaka *et al.* 1995).

cosmic masers varies, often erratically, on timescales of hours to years. The underlying electric field is a Gaussian random process (Moran 1981). In an unsaturated maser (in astronomical terminology), the pumping is sufficiently strong that the microwave intensity does not affect the level populations, and the intensity increases exponentially through the masing medium. The input signal can be either a background source or the maser's own spontaneous emission. In a saturated maser, one pump photon is needed for each microwave photon, and in a one-dimensional maser medium where beaming can be neglected, the intensity increases linearly with distance. Maser emission is expected to be beamed. Most masers are thought to be saturated, and this condition requires the least pump power. However, this assertion is difficult to prove observationally.

Consider a simple geometry for a maser, a filamentary tube, shown in Fig. 2. The boundaries of the filament can be defined in terms of either gas density or the region

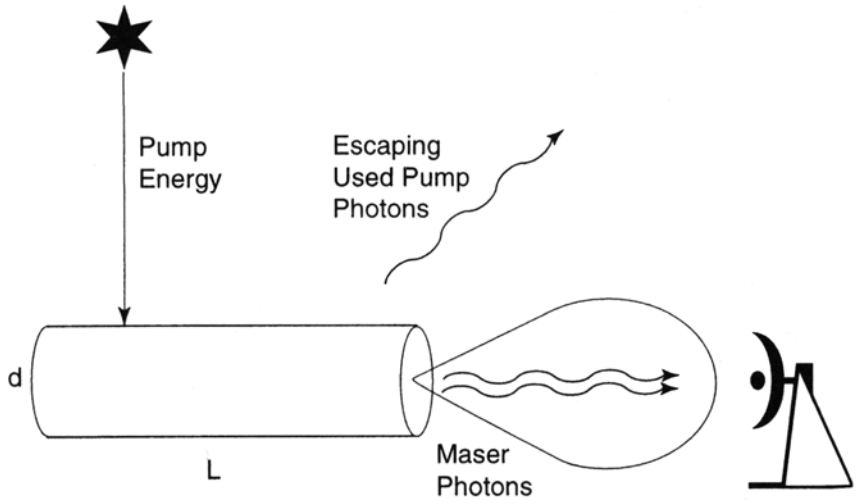

Figure 2. A cartoon of a simple filamentary maser. The pump energy can be supplied by either shock waves or radiation. Pump cycles usually involve excitation to infrared rotational levels, followed by de-excitation to the upper level of the maser transition. In order to avoid thermalization, the infrared photons emitted during the pump cycle must escape from the masing medium, which favors a geometry that is thin in at least one dimension. In a filamentary geometry, the masing medium is defined either by the physical extent of the masing gas or by the volume over which the gas is coherent, that is, where the variation in the velocity projected along the line of sight is less than the thermal line width. The maser emission is beamed into a cone of angular opening equal to *d/L*. If the maser amplifies a background source, the radiation will be beamed in the forward direction. If it amplifies its own spontaneous emission, then beamed maser emission will emerge from both ends, and weaker emission from the sidewalls.

where the line-of-sight velocity is constant to within the thermal line width. The gas in the tube is predominantly molecular hydrogen, with trace amounts of water vapor (about one part in 10^5) and other constituents. If this maser medium is saturated, then the luminosity is given by

$$L = h\nu n \Delta P V; \tag{1}$$

where h is Planck's constant, v is the frequency, n is the population density in pumping level, ΔP is the differential pump rate per molecule, and V is the volume of the masing cloud. This is the most luminosity a maser of given pump rate and volume can produce. The emission will be beamed along the major axis of the filament into an angle

$$\beta \cong \frac{d}{L}, \tag{2}$$

where d is the cross-sectional diameter, and L is the length of the filament. The beam angle and the length of the maser are not directly observable. Since the volume of the maser is approximately d^2L, the observed flux density from a maser beamed toward the earth is

$$F_\nu = \frac{1}{2} h\nu n \frac{\Delta P}{\Delta \nu} \frac{V}{D^2 \Omega} = \frac{1}{2} h\nu n \frac{\Delta P}{\Delta \nu} \frac{L^3}{D^2}, \tag{3}$$

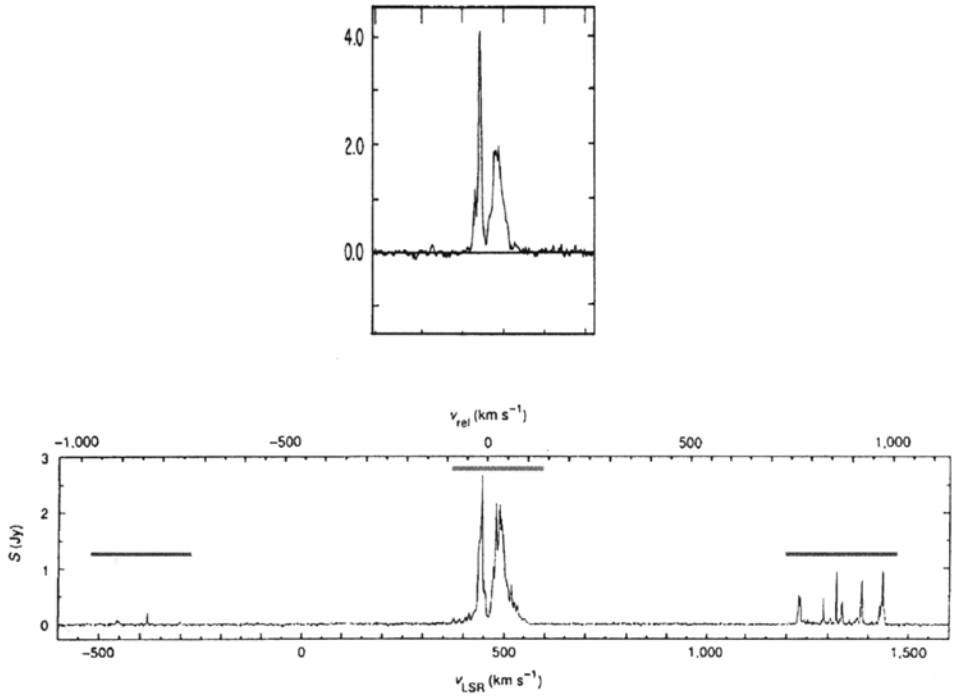

Figure 3. The top panel shows the maser spectrum as first discovered in NGC 4258 at velocities near the systemic velocity of the galaxy (Claussen, Heiligman & Lo 1984). The lower spectrum shows the observations of Nakai, Inoue & Miyoshi (1993) over a much broader velocity range. The velocities are computed from the Doppler effect and are based on a rest frequency of 22235.080 MHz. The effects of the motions of the earth and sun with respect to the local standard of rest have been removed. The bars indicate velocity ranges of emission.

where Δv is the line width, D is the distance between the maser and the observer, and Ω is the beam angle of the emission, $\sim \beta^2$. The maximum allowable hydrogen number density is about 10^{10} molecules per cubic centimeter, above which the maser levels become thermalized by collisions. This maximum allowable density ($\rho_c = 3 \times 10^{-13}$ gcm^{-3}, an important parameter in much of the following discussion) along with the maximum allowable pump rate, which equals the Einstein A-coefficient for infrared transitions linking the maser levels, limit the luminosity of a maser of given volume.

Water masers outside our Galaxy were first discovered in the spiral arms of the nearby galaxy M33 by Churchwell *et al.* (1977). Their properties were found to be similar to masers found in Galactic star-forming regions. Much more luminous water masers were found in the AGN associated with NGC 4945 and the Circinus galaxy by dos Santos & Lepine (1979) and Gardner & Whiteoak (1982), respectively. Claussen, Heiligman & Lo (1984) and Claussen & Lo (1986) conducted surveys and found five additional masers associated with AGN, including the one in NGC 4258 (see Fig. 3). They suggested that these masers might arise in gas associated with dust-laden molecular tori that had been proposed to surround black holes by Antonucci & Miller (1985). Nakai, Inoue & Miyoshi (1993), with a powerful new spectrometer of 16,000 channels spanning a velocity range of 3000 km s^{-1}, observed NGC 4258 and

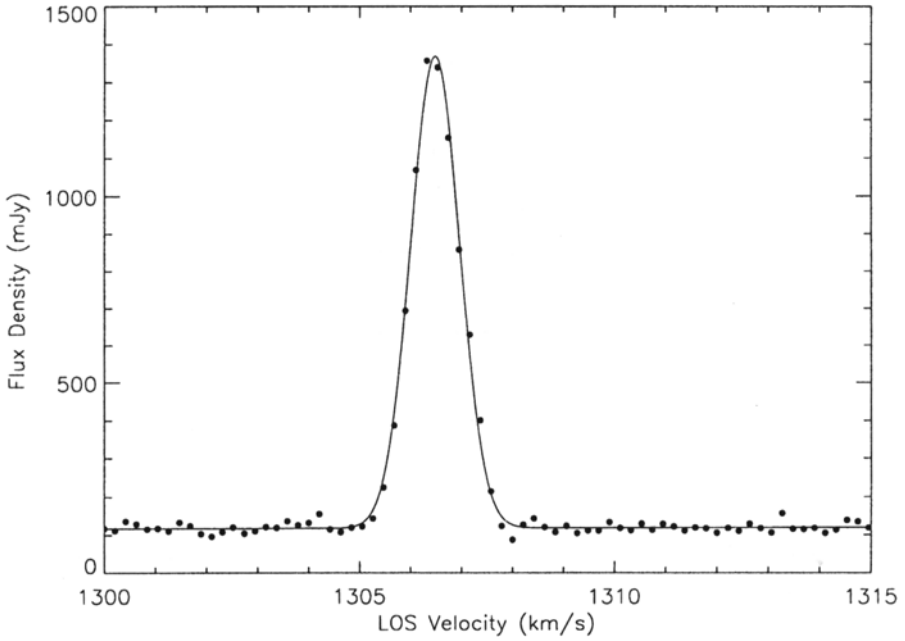

Figure 4. An expanded view of the spectral feature near 1306 km s^{-1} in NGC 4258, which is typical of emission from an isolated maser component (see Fig. 3). The line width is characteristic of gas at 300 K, but the intensity corresponds to that of a blackbody at an equivalent temperature greater than 10^{14} K.

discovered satellite line clusters offset from the systemic velocity by about ±1000 km s^{-1}, which are shown in Fig. 3 (see also Miyoshi 1999).

The compelling reason that the radio emission from the water vapor transition arises from the maser process is straightforward. A typical example of a maser line from a small part of the spectrum of NGC 4258 is shown in Fig. 4. The line width is about 1km s^{-1}, the thermal broadening expected for a gas cloud at 300 K. However, the angular size determined by radio interferometry is less than 100μas, implying that the equivalent blackbody temperature must exceed 10^{14} K. The exceedingly high brightness of the radiation is the principal evidence for the maser process. (Typical molecular lines from molecular clouds have velocities of several tens of km s^{-1}—due to thermal and turbulent broadening—and brightness temperatures of less than 100 K.) Cosmic masers produce very bright spots of radiation but have little else in common with terrestrial masers. It is difficult to use masers to determine physical conditions (e.g., temperature, density) in molecular clouds because of the complexity of the maser process. However, as compact sources of narrowband radiation, masers are ideal probes of the dynamics of their environment.

3. VLBI

The most important tool for the study of the angular structure of masers is very long baseline interferometry (VLBI). Signals from a maser, or from other bright compact radio sources, are converted to a low-frequency baseband and recorded in digital

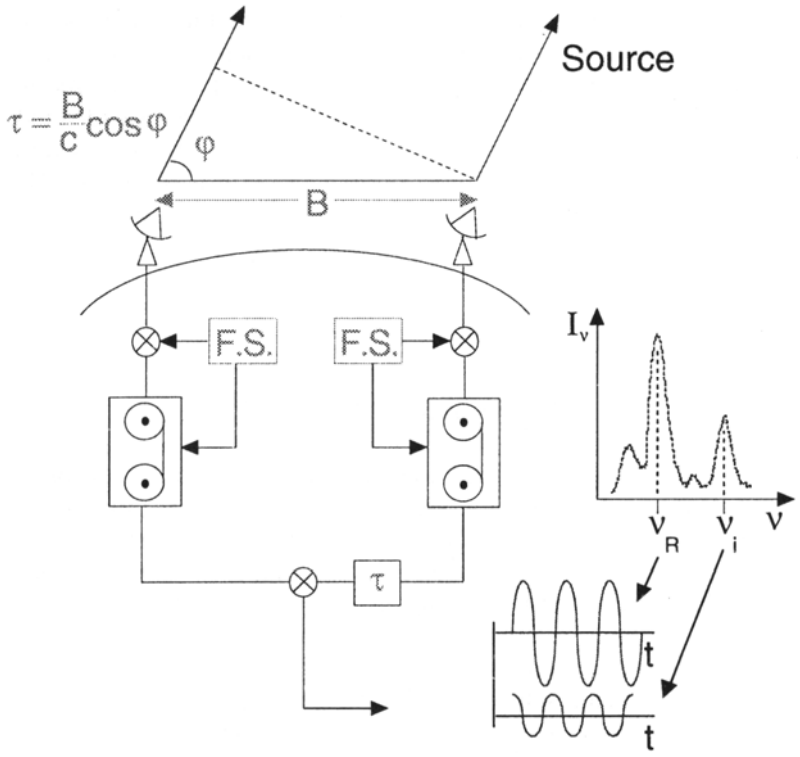

Figure 5. A block diagram of a two-element very long baseline interferometer. It operates as a coherent interferometer. An atomic frequency standard (F.S.) controls the phase of the local oscillator signal at each telescope used to convert the radio frequency signal to a video band for recording on magnetic tape in digital form (without square-law detection) and sampled at the appropriate Nyquist rate. On playback, one of the signals is delayed by τ to compensate for the differential propagation time from the source to the antennas. The signals are correlated and Fourier transformed to produce cross-power spectra, or correlated power as a function of frequency and time.

format on magnetic tape at Nyquist sampling rates of up to about 10^8 samples per second at widely separated telescopes that operate independently. They form a radio version of the classical Michelson stellar interferometer, whose coherence is maintained by the use of atomic frequency standards to preserve the signal phase and timing (see Fig. 5). The received signals (which are proportional to the incident electric fields) from an array of two or more telescopes are cross-correlated pairwise to form cross-correlation functions. Taking advantage of the earth's rotation, the spatial cross-correlation function of the incident electromagnetic field, or visibility, can be measured over a wide range of projected baseline vectors. The image and fringe visibility functions are related through a Fourier transform (see Thompson, Moran & Swenson 1986). The temporal Fourier transform of the cross-correlation function gives the cross-power spectrum of the radiation, or visibility as a function of frequency, so that images at different frequencies can be obtained. The intrinsic angular resolution, θ, of a multielement interferometer is $0{\cdot}7\lambda/B$, where λ is the wavelength, and B is the longest baseline length. For water vapor, $\lambda = 1{\cdot}35$ cm, and B is typically 6000 km, which gives a resolution of 200 μas. The spectral resolution

available is typically about 15 KHz, or about a fifth of the line widths (see Fig. 4). In maser sources, one spectral feature at a particular frequency or velocity can be used as a phase reference for the interferometer, and all other phases referred to it. With this technique, the coherence time of the interferometer can be extended indefinitely, and the relative positions of masers with respect to the reference feature can be measured to a small fraction of the fringe spacing, or intrinsic resolution. The relative position of an unresolved maser component can be measured to an accuracy of about

$$\Delta\theta = \frac{1}{2}\frac{\theta}{SNR}, \tag{4}$$

where *SNR* is the signal-to-noise ratio.

4. The study of NGC 4258

The imaging of the maser in NGC 4258 was one of the first projects undertaken by a dedicated VLBI system known as the Very Long Baseline Array (VLBA) (see Fig. 6) in the spring of 1994. Previous VLBI measurements of the systemic features had shown that they arose from an elongated structure with a velocity gradient along the major axis, highly suggestive of a rotating disk seen edge-on (Greenhill *et al.* 1995a). The VLBA measurements of all the maser components provide convincing evidence for a rotating disk around a massive central object.

The basic observational results on NGC 4258 obtained over the past few years can be summarized as follows (see Table 1 for a list of parameters):

1. The masers appear to trace a highly elongated, although slightly curved, structure (Fig. 7). The high-velocity, redshifted and blueshifted features are offset in position on the left and right sides of the systemic features, respectively. The velocities of

Figure 6. The distribution of the ten elements of the Very Long Baseline Array (VLBA). This network is often augmented with other radio telescopes such as the Very Large Array (a 27-element array operating as a phased array), shown here with the symbol Y, and the 100 m telescope of the Max Planck Institute for Radio Astronomy near Bonn, Germany.

Table 1. Parameters of molecular disk traced by water vapor masers in NGC 4258[a].

Inner radius, R_i	0.14 pc (3.9 mas)
Outer radius, R_o	0.28 pc (8.0 mas)
Inner rotation velocity, $v_\phi(R_i)$	1100 km s^{-1}
Outer rotation velocity, $v_\phi(R_o)$	770 km s^{-1}
Inner rotation period	800 yrs
Outer rotation period	2200 yrs
Position angle of disk (at 3.9 mas radius)	80°
Inclination angle	98°
Position velocity slope	282 km s^{-1} mas^{-1}
Central mass, M	3.9×10^7 M$_\odot$
Disk mass, M_d	$< 10^6$ M$_\odot$
Central mass density, uniform distribution	$> 4 \times 10^9$ M$_\odot$ pc^{-3}
Central mass, Plummer distribution	$> 10^{12}$ M$_\odot$ pc^{-3}
Centripetal acceleration, systemic features	9.3 km s^{-1} yr^{-1}
Centripetal acceleration, high-velocity features	< 0.8 km s^{-1} yr^{-1}
Disk systemic velocity[b], v_0	476 km s^{-1}
Galactic systemic velocity (optical)[b,c]	472 km s^{-1}
Radial drift velocity, v_R	< 10 km s^{-1}
Thickness of disk, H	< 0.0003 pc
Maser beam angle, β	8°
Disk-galaxy angle[d]	119°
Apparent maser luminosity	150 L$_\odot$
Model luminosity[e]	11 L$_\odot$
Distance, D	7.2 ± 0.3 Mpc.

[a] Based on the distance estimate of 7.2 Mpc.
[b] Radio definition, with respect to the local standard of rest. To convert to heliocentric velocity (radio), subtract 8.2 km s–1; to convert to heliocentric (optical), subtract 7.5 km s^{-1}.
[c] From Cecil, Wilson & Tully (1992).
[d] Angle between the spin axis of the molecular disk and the spin axis of the galaxy.
[e] Radiation into a zone within ±4° of the plane of the disk.

the high-velocity features as a function of impact parameter (position along the major axis of the distribution) follow the prediction of Kepler's third law of orbital motion. The systemic features show linear dependence with impact parameter (Miyoshi *et al* 1995).

2. The distribution of maser features in the direction normal to the major axis is too small to be measured at present (see Fig. 7). The upper limit on the ratio of thickness to radius of the disk is 0.0025 (Moran *et al* 1995).
3. The upper limit of any toroidal component of the magnetic field in the masers, derived from searches for Zeeman splitting in the line at 1306 km s^{-1}, is less than 300 mG (Herrnstein *et al.* 1998a).
4. The accelerations (i.e., the linear drift in the line-of-sight velocity with time) of the systemic features are about 9 km s^{-1} yr^{-1} (Haschick, Baan & Peng 1994; Greenhill *et al.* 1995b; Nakai *et al.* 1995). The high-velocity features that have been tracked have accelerations in the range ±0·8 km s^{-1} yr^{-1} (Bragg *et al.* 1999).
5. The high-velocity features show no proper motions with respect to a fixed-velocity component in the systemic range (Herrnstein 1996). The systemic features show proper motions of about 32 μas yr^{-1} (Herrnstein *et al.* 1999).

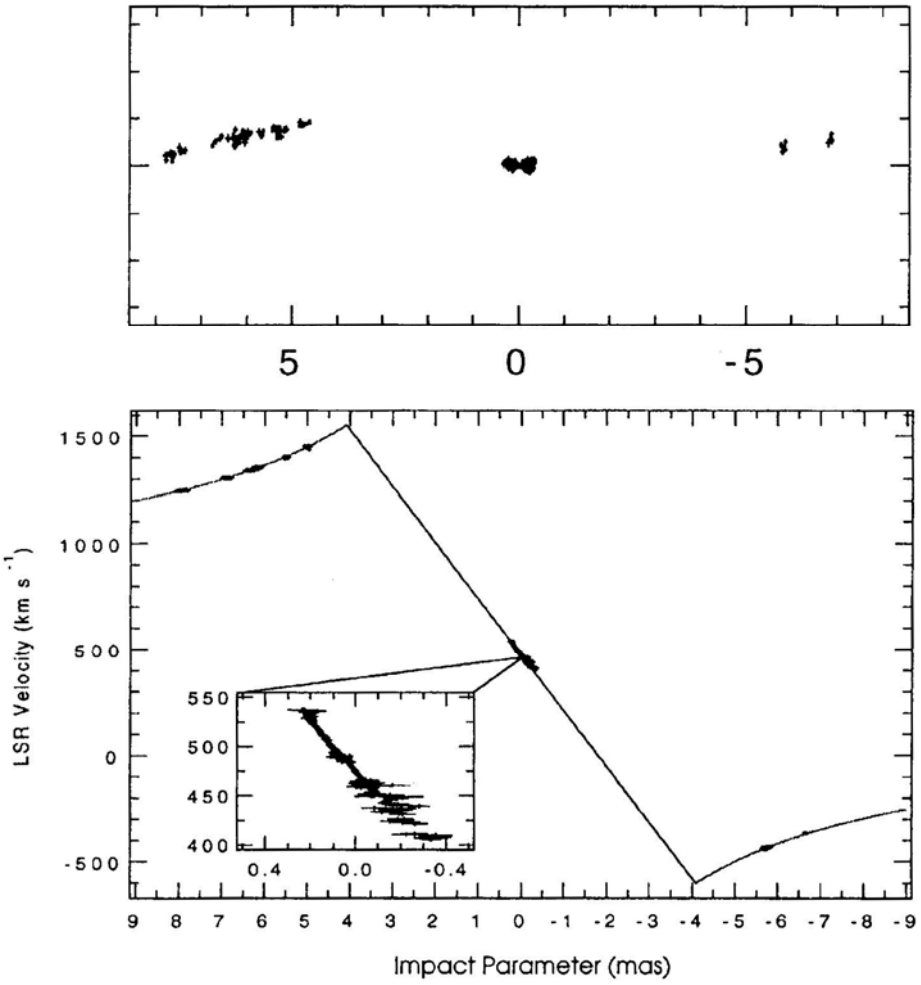

Figure 7. Top: Image of the maser emission from the nucleus of NGC 4258. The ticks on the axes are in milliarcseconds. One milliarcsecond corresponds to 0.035 pc, or 1.1×10^{17} cm, at a distance of 7.2 Mpc. **Bottom:** The line-of-sight velocities of the masers versus position along the major axis. The curved portions of the plot precisely follow a Keplerian dependence. Data from January 1995 (top) and April 1994 (bottom).

6. There is an elongated continuum radio source, which appears to be a jet emanating from the black hole position, parallel to the axis of rotation (Herrnstein *et al.* 1997). There is no 1.35 cm wavelength emission from the position of the black hole (Herrnstein *et al.* 1998b).

There is virtually no doubt that the masers trace a very thin disk in nearly perfect Keplerian motion. Five of six phase-space parameters have been measured for each maser spot, two spatial coordinates and three velocity coordinates. The missing coordinate is the position along the line of sight, which must be inferred from the constraint provided by Kepler's third law.

The approximate placement of the masers in the disk can be understood by considering a simple thin, flat disk viewed edge-on. In this case the line-of-sight

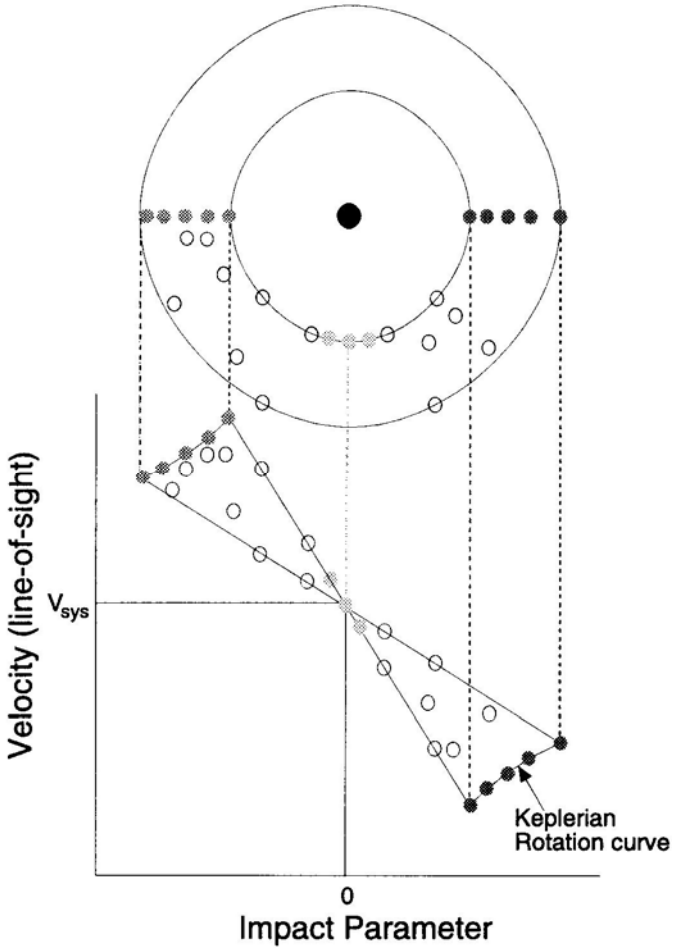

Figure 8. Top: A cartoon model of a flat annular disk viewed edge-on, with randomly distributed maser sources. **Bottom:** Each maser will appear as a point within the "bow tie" boundary in the plot of line-of-sight velocity versus impact parameter. The curved portion of the boundary is populated by masers located along the midline, the diameter perpendicular to the line of sight. This is a pure Keplerian curve, because the velocity vectors lie along the line of sight. Masers at a fixed radius will appear along a straight line. The steep and shallow lines correspond to masers on the inner and outer annular boundaries of the disk.

velocity, v_z, of a maser will be given by

$$v_z - v_0 = \sqrt{\frac{GM}{R}} \sin \phi, \tag{5}$$

where v_0 is the line-of-sight velocity of the central object (i.e., the systemic velocity), G is the gravitational constant, R is the distance of a maser component from the black hole, and ϕ is the azimuth angle in the disk, measured from the line between the black hole and the observer. If the disk were randomly filled with observable masers, one might expect to see a velocity position diagram as shown in Fig. 8. The linear boundaries of the distribution are populated by masers at the inner and outer edges of

the annular disk. The masers on the curved boundaries lie on the midline, where $\phi = 90°$. Hence, the rnasers in NGC 4258 have a very specific distribution: the high-velocity masers lie close to the midline, and the systemic masers lie within a narrow range of radii. From equation (5), the radius of a particular maser can be determined as

$$R = (\mathrm{GM})^{1/3}\left[\frac{b}{v_z - v_0}\right]^{2/3}, \tag{6}$$

where b is the projected distance on the sky along the major axis from the center of the disk (sin $\phi = b/R$). Similarly, positional offsets from the midline, z, of the high-velocity features can be determined by deviations from a Keplerian curve; that is,

$$z = \sqrt{R^2 - b^2}. \tag{7}$$

There is a two-fold ambiguity in the z component of the position for the edge-on disk case. Unambiguous estimates of the positions of the high-velocity features have been derived from the accelerations by Bragg *et al.* (1999), who showed that the masers lie within 15 degrees of the midline.

An expanded plot of the Keplerian part of the velocity curve is shown in Fig. 9. The data fit a Keplerian curve to an accuracy of about 3 km s^{-1} , or less than 1 per cent of the rotation speed. However, there are noticeable deviations from a perfect fit. The estimate of the central mass of the disk derived from this data depends on the distance to the maser, and has a value of 3.9×10^7 $M_\odot$ for a distance of 7.2 Mpc. This mass corresponds to an Eddington luminosity (where radiation pressure from Thomson scattering would balance gravity) of 5×10^{45} erg s^{-1}. Since the total electromagnetic emission appears to be less than $l0^{42}$ erg s^{-1}, the system is highly sub-Eddington.

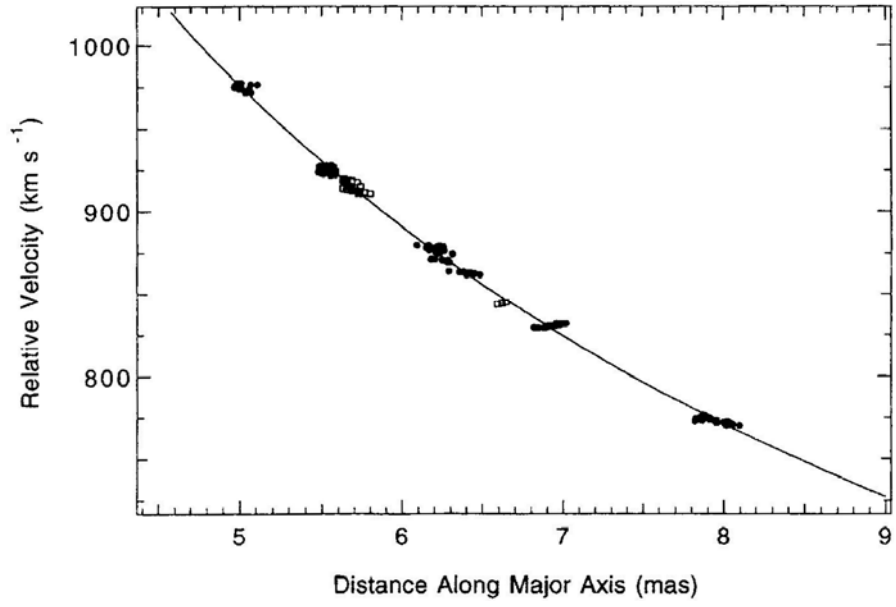

Figure 9. The magnitude of the line-of-sight velocities of the masers, relative to the systemic velocity, versus distance from the dynamical center of NGC 4258. The filled circles are the redshifted masers, and the squares are the blueshifted masers. (Data from April 1994).

Since this binding mass must lie inside the inner radius of the maser disk, the mass density, assuming a spherical mass distribution, is at least 4×10^9 $M_\odot$ pc^{-3} (3×10^{-13} g cm^{-3}). It is unlikely that this mass is in the form of a dense star cluster (Maoz 1995). The average density of stars in the solar neighborhood is about 1 $M_\odot$ pc^{-3}, and the density of the densest known star cluster is about 10^5 $M_\odot$ pc^{-3}. A star cluster will have a mass distribution that decreases monotonically with radius. In order not to disrupt the Keplerian curve, the core mass for a reasonable distribution must have a peak density of at least 1×10^{12} $M_\odot$ pc^{-3} . A cluster of massive stars at this density would evaporate from gravitational interactions on a timescale short with respect to the age of the galaxy, while a cluster of low-mass stars would destroy itself from collisions over a similar timescale. Hence, it is unlikely that the central mass is in the form of a star cluster (see also Begelman & Rees 1978). The best explanation is that the central object is a supermassive black hole, with a Schwarzschild radius (*Rs*) of about 1.2×10^{13} cm. Hence, the masers are distributed in a zone between 40,000 and 80,000 *Rs*. Because the maser clouds are so far from the event horizon, deviations of their motions from the predictions of Newtonian mechanics are small. The gravitational redshift and transverse Doppler shift are about 4 km s^{-1} (detectable), the expected Lense-Thirring precession (see below) is less than about 3° over the maser annulus (possibly detectable), and the apparent shift of the maser positions due to gravitational bending is about 0.1 μas (undetectable).

The disk is remarkably thin. In a disk supported against gravity by pressure (hydrostatic equilibrium), the density distribution is expected to have a Gaussian profile with a thickness, *H*, given by the relation

$$H/R = (c_s^2 + v_a^2)^{1/2}/v_\phi, \tag{8}$$

where c_s is the sound speed and v_a is the Alfvén speed, which characterize thermal and magnetic support pressure, respectively, and v_ϕ is the Keplerian rotational speed. Since $H/R < 0.0025$, the quadrature sum of the sound speed and Alfvén speed is less than 2.5 km s^{-1}. The upper limit on the magnetic field of 300 mG suggests that the Alfvén speed, $B/\sqrt{4\pi\rho}$,*where* ρ is the density, is less than 3 km s^{-1} for $\rho = \rho_c$ (the critical density for quenching maser emission). If the support were completely due to thermal pressure, the temperature would be less than 1000 K.

A proper determination of the positions of the masers on the disk requires that the warp and the inclination of the disk to the line of sight be taken into account. An example of such a disk, slightly warped (in position angle only) and slightly inclined to the line of sight, that fits the maser distribution in position and velocity is shown in Fig. 10.

The distance to the maser of 7.2 ± 0.3 Mpc was determined from analysis of the proper motions and accelerations of the systemic features (Herrnstein *et al.* 1999). Fifteen features were tracked over a period of two years to an accuracy of 0.5–10 μas in relative position and 0.4 km s^{-1} in velocity. The distance estimate is based on simple geometric considerations. The Keplerian curve of the high-velocity masers gives the mass function $GM \sin^2 i/D$, where *i* is the inclination of the disk to the line of sight. The radius, *R*, of the systemic masers (in angular units) is determined from equation (6), based on the slope of the velocity versus impact parameter curve shown in Fig. 7. This fixes the angular velocity, $v\phi$, of the systemic masers under the assumption that the orbits are circular. The expected accelerations and proper motions of the systemic features are v_ϕ^2/R and v_ϕ/D, respectively. The assumption that the

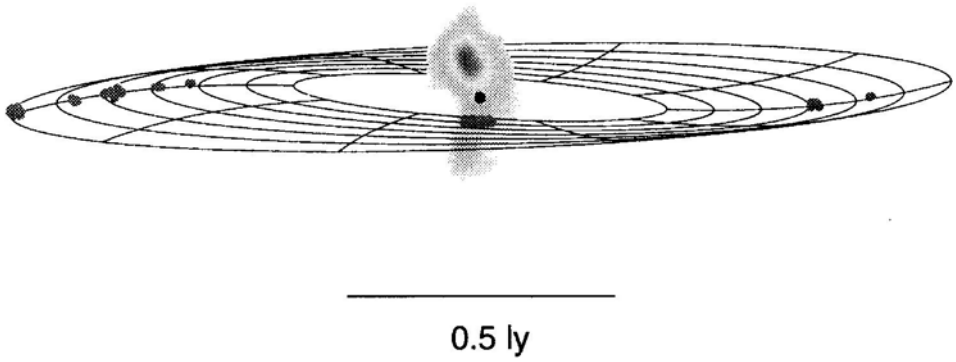

Figure 10. The warped annular disk (wire mesh) modeled to the maser positions, velocities, and accelerations (adapted from Herrnstein, Greenhill & Moran 1996). The black dot in the center marks the dynamical center of the disk. The continuum emission at 1.3 cm is shown in the shaded gray contours. The southern jet may be weaker than the northern jet because of thermal absorption in the disk. The lack of emission at the position of the black hole places constraints on any coronal or advection zone surrounding the black hole (Herrnstein *et al.* 1998b).

orbits are circular is reasonable on theoretical grounds because of viscous relaxation and on observational grounds because the continuum emission arises close to the center of symmetry of the maser distribution.

The distance to the 15 Cepheid variables in NGC 4258 has been estimated to be 8.1 + 0.8 Mpc (Maoz *et al.* 1999). The statistical component of the error is 0.4 Mpc and the systematic error associated with the calibration of the Cepheid distance scale is 0.7 Mpc. The discrepancy between the two distance measurements to NGC4258 may have cosmological implications (Paczynski 1999).

5. Interesting unanswered questions

1. *What is the rate of radial inflow of material through the disk?*

The accretion rate of material onto the black hole is an important parameter that affects our understanding of radiation processes around the black hole. The maser data provide some information about the accretion rate. Key issues are the long timescale needed for material to flow from the disk to the black hole and the assumption that the masers trace all the disk material. It is useful to first estimate the mass of the disk. If we account for systematic effects in addition to the random scatter of 3 km s^{-1}, the deviation from Keplerian motion due to the finite mass of the disk, Δv_ϕ, is less than about 10 km s^{-1} over the radius of the disk. This limits the mass of the disk to less than about $2M\Delta v_\phi / v_\phi$, or about 10^6 $M_\odot$. The density of the molecular gas must be less than ρ_c (10^{10} hydrogen molecules per cubic centimeter). Since $H/R < 0.0025$, the upper limit on mass is 10^5 $M_\odot$. If, in addition, the disk is stable against the effects of self-gravity (Toomre 1964; Binney & Tremaine 1987), then the mass of the disk must be less than $M(H/R)$, or about 10^5 $M_\odot$.

The mass accretion rate of a disk in steady state is given by

$$\dot{M} = 2\pi R \Sigma v_R, \tag{9}$$

where Σ is the surface density of the disk, and v_R is radial drift velocity, which depends on the viscosity of the disk. Unfortunately, v_R is only weakly constrained by the observations (i.e., the possible difference between the optical and radio systemic velocities) to be < 10 km s^{-1}. This provides a crude limit on the accretion rate of 100 $M_\odot$ yr^{-1}. To further constrain the mass accretion rate requires an estimate of the viscosity of the disk. In the standard model of a thin, viscous accretion disk, as formulated by Shakura & Sunyaev (1973), v_R can be written (see Frank, King & Raine 1992) as

$$v_R = \alpha v_\phi \left(\frac{H}{R}\right)^2, \tag{10}$$

where α is the dimensionless viscosity parameter ($0 \leq \alpha \leq 1$). The observational limit on the ratio H/R implies that v_R < 0.006a km s^{-1}. With the limit on mass given by the deviation from Keplerian motion, the accretion rate is less that $10^{-1}\alpha$ $M_\odot$ yr^{-1}. The infall time from the masing region is R/v_R, which from equation (10) can be written as

$$T \sim \frac{1}{\alpha}\left(\frac{c}{c_S}\right)^2\left(\frac{R_S}{c}\right)\left(\frac{R}{R_S}\right)^{1/2}. \tag{11}$$

For NGC 4258, with α = 0.1 and *cs* = 2.5 km s^{-1}, $T = 10^{16}$ s, or about 3×10^8 yrs.

From the magnetic field limit and the assumption of equipartition of magnetic and thermal energy, the upper limit on $\dot{M}$ is also $10^{-1}\,\alpha\,M_\odot$ yr^{-1}. If the maser density is the maximum allowable value, ρ_c, and the maser traces all the material in the disk, then the limit on disk thickness leads to an upper limit on $\dot{M}$ of $10^{-2}\,\alpha$ $M_\odot$ yr^{-1}. Detailed theoretical modeling can give estimates for the accretion rate. For example, a model in which the cause of the outer radial cutoff in maser emission is attributed to the transition from molecular to atomic gas leads to an estimate of $10^{-4}\alpha$ $M_\odot$ yr^{-1}. (Neufeld & Maloney 1995). Gammie, Narayan & Blandford (1999) favor an accretion rate of $10^{-1}\,\alpha\,M_\odot$ yr^{-1}, based on an analysis of the continuum radiation spectrum.

If the accretion rate is high, then the relative weakness of the continuum radiation may be due to the process of advection (Gammie, Narayan & Blandford 1999). On the other hand, if the accretion rate is low, then the weak emission is due to the dearth of infalling material. In this case the gravitational power in the accretion flow may be insufficient to power the jets.

2. *What is the form and origin of the warp?*

The form of the warp is difficult to determine precisely, because the filling factor of the masers in the disk is so small. Better measurements of the positions and directions of motion of the high-velocity features are key to defining the warp more accurately.

The cause of the warp is unknown, but several suggestions have been put forward. Papaloizou, Terquem & Lin (1998) show that the warp could be produced by a binary companion orbiting outside the maser disk. Its mass would need to be comparable to the mass of the disk (< 10^6 $M_\odot$). Alternatively, radiation pressure from the central source will produce torques on a slightly warped disk and will cause the warp to grow (Maloney, Begelman & Pringle 1996). Finally, it is conceivable that in the absence of other torques, the observed warp is due to the Lense-Thirring effect. A maximally

rotating black hole will cause a precession of a nonaligned orbit (weak field limit) of

$$\Omega_{LT} = \frac{2G^2M^2}{c^3R^3}, \tag{12}$$

which can be rewritten in terms of the Schwarzschild radius as

$$\Omega_{LT} = \frac{1}{2}\frac{c}{R_S}\left(\frac{R_S}{R}\right)^3. \tag{13}$$

At the inner radius of the disk *(R/Rs* = 40,000), the precession amounts to 3×10^{-17} s^{-1}. This precession is very small but might be significant over the lifetime of the disk. Equation (11) suggests that the lifetime might be 10^{16} s, which would produce a differential precession of about 10° across the radius of the disk. If the axis of the disk is inclined to the axis of the black hole, then the viscosity of the disk is expected to twist the plane of the innermost part of the disk to the equatorial plane of the black hole (Bardeen & Petterson 1975; Kumar & Pringle 1985).

3. *Do the water masers trace the whole disk?*

The inner and outer radii of the observed masers are undoubtedly due to excitation conditions in the maser. In the vertical direction it is also possible that the masers form in a thin region within a thicker disk with atomic and ionized components. It has also been proposed that the high-velocity features are not indicative of a warped disk but trace material that has been blown off a flat disk (Kartje, Konigl & Elitzur 1999). These proposals are difficult to test.

4. *What are the physical properties of the maser spots?*

The spectrum of the maser has many discrete peaks which correspond to spots of maser emission on the sky. The success of measuring proper motions and accelerations of these masers suggests that they correspond to discrete condensations or density-enhanced regions in the disk. A cartoon of the blobs in a disk is shown in Fig. 11. The blobs in front of the black hole may be visible because they amplify emission from the central region. No masers have been seen on the backside of the disk. On the other hand, the high-velocity masers have no continuum emission to amplify, and we may only see the ones near the midline, where the gradient in the line-of-sight velocity is small. Blobs in the rest of the disk may be radiating in directions away from the earth. The dumpiness of the medium allows us to track the individual masers. If the appearance of spots is due to blobs, or density enhancements in the disk, then the minimum gradient condition would not seem to be necessary. However, intense high-velocity maser spots may occur when blobs at the same velocity line up to form two-stage masers (Deguchi & Watson 1989). This situation forms a highly beamed maser, like the filamentary maser described in section 2. The probability of realizing this situation is greatest along the midline, where the velocity gradient is smallest. All evidence suggests that the masers arise from discrete physical condensations. There have been several suggestions that the apparent motions of the

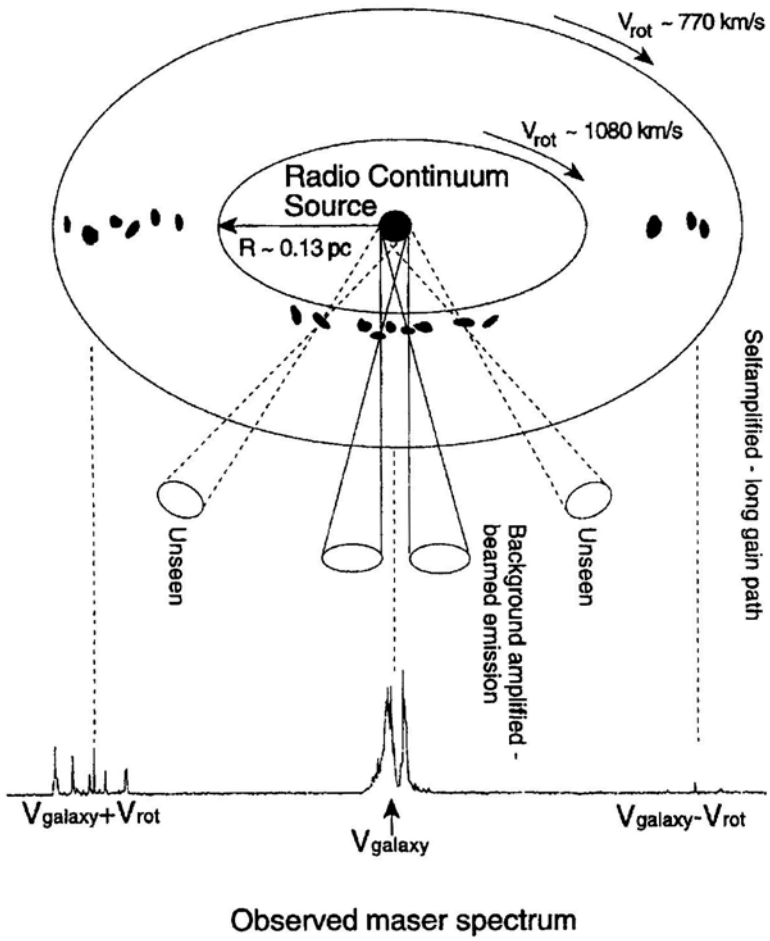

Figure 11. A cartoon of the molecular accretion disk in NGC 4258. The period of rotation at the inner edge of the disk is about 800 years. A particular systemic maser is visible for about 10 years as its radiation beam, estimated to have a width of about 8°, sweeps over the earth. The high-velocity features may be visible for a substantially longer time. (Adapted from Greenhill *et al.* 1995a).

maser spots may be due to a phase effect (e.g., a spiral density wave moving through the disk, Maoz & McKee 1998), but there is no observational evidence for this.

6. Masers in other AGN

At this time (early 1999), 22 masers have been detected among about 700 galaxies searched (e.g., Braatz, Wilson & Henkel 1997). A list of these galaxies with masers is given in Table 2. The yield rate of detections is only about 3 per cent. The major reason for this paucity is probably that the maser disks can only be seen if they are edge-on to the line of sight. If the typical beam angle, β, is 8°, as in NGC 4258, then the probability of seeing a maser is about equal to sin β, or 8 per cent. Braatz, Wilson

Table 2. Known AGN with water maser

Galaxy	Distance (Mpc)	Flux density (Jy)	mas Structure	Disk
M51	3	0.2		
NGC 4945	3.7	4	yes	yes
Circinus	4	4	yes	yes
NGC 4258	7.2	4	yes	yes
NGC 1386	12	0.9	yes	maybe
NGC 3079	16	6	yes	maybe
NGC 1068	16	0.6	yes	yes
NGC 1052	20	0.3	yes	no
NGC 613	20	0.1?	–	–
NGC 5506	24	0.6	–	–
NGC 5347	32	0.1	–	–
NGC 3735	36	0.2	–	–
IC 2560	38	0.4	yes	–
NGC 2639	44	0.1	–	–
NGC 5793	50	0.4	–	–
ESO 103-G035	53	0.7	–	–
Mrk 1210	54	0.2	–	–
IRAS F01063-8034	57	0.2	–	–
Mrk 1	65	0.1	–	–
NGC 315	66	0.05	–	–
IC 1481	83	0.4	–	–
IRAS F22265-1826	100	0.3	yes	no

Note: Extragalactic masers outside AGN are found in NGC 253, M 82, M 33, IC 342, LMC, and SMC.

& Henkel (1997) have shown that most of the known masers are associated with Seyfert II galaxies or LINERs where the accretion disks are thought to be edge-on to the earth.

It is difficult to make VLBI measurements on masers weaker than about 0.5 Jy because of the need to detect the maser within the coherence time of the interferometer. Nine masers have been studied with VLBI. Four of these show strong evidence of disk structure, and two more show probable disk structure. The propertieis of these masers are listed in Table 3. Unfortunately, none of these masers show the simple, well-defined structure that would make them useful for precise study of the physical properties of accretion disks around black holes.

7. Summary

The measurements of the positions and velocities of the masers in the nucleus of NGC 4258 offer compelling evidence for the existence of a supermassive black hole and provide the first direct image of an accretion disk within 10^5 R_s of the black hole. Much more can be learned from this system. A measurement of the disk thickness is important and may require higher signal-to-noise ratios than are achievable currently or VLBI measurements from space. Measurement of the continuum spectrum from the central region is very important to the understanding of the radiation process. Detection of radio emission would require instruments of higher sensitivity. Continued measurements over time of the positions and velocities of the masers

Table 3. Water masers with resolved structures.

Masers without obvious disk structure						
Galaxy	D Mpc	v_o km s^{-1}	Δv km s^{-1}	ΔR pc	Comment	Reference
IRAS 22265 (S0)	100	7570	150	2.4	messy	Greenhill *et al.* 1999a
NGC 1052 (E4)	20	1490	100	0.6	"jet"	Claussen *et al.* 1998
IC 2560	38	2900	30	0.2	velocity gradient	Nakai *et al.* 1998

Masers with disk structure							
Galaxy	D Mpc	v_ϕ km s^{-1}	R_i/R_o pc	M 10^6 M$_\odot$	ρ 10^7 M$_\odot$ pc^{-3}	L_x 10^{42} erg s^{-1}	Reference
NGC 4258	7	1100	0.13/0.26	35	400	0.04	Miyoshi *et al.* 1995
NGC 1068	15	330	0.6/1.2	17	3	40	Greenhill & Gwinn 1997
Circinus	4	230	0.08/0.8	1	40	40	Greenhill *et al.* 1999b
NGC 4945	4	150	0.2/0.4	1	2	1	Greenhill *et al.* 1997
NGC 1386	12	100	−/0.7	2	4	0.02	Braatz *et al.* 1999
NGC 3079	16	150	−/1.0	1	0.2	0.02	Trotter *et al.* 1998, Satoh *et al.* 1998

D = distance, v_o = systemic velocity, Δv = velocity range, ΔR = linear extent, v_ϕ = rotational velocity, R_i/R_o = inner/outer radius of disk, M = central mass, ρ = central mass density, L_x = X-ray luminosity.

will refine the estimates of their proper motions and accelerations, and this will better define the shape of the disk. It is even conceivable that the radial drift velocity will be detected. This work will benefit immensely from new instruments that are in the planning stage for centimeter wavelength radio astronomy. These include the enhanced Very Large Array, the Square Kilometer Array, and space VLBI missions such as ARISE.

We thank Adam Trotter and Ann Bragg for helpful discussions.

References

Antonucci, R. R. J., Miller, J. S. 1985, *Astrophys. J.*, **297**, 621–632.
Bardeen, J. M., Petterson, J. A. 1975, *Astrophys. J.*, **195**, L65–L67.
Begelman, M. C, Rees, M. J. 1978, *Mon. Not. R. Astr. Soc.*, **185,** 847–859.
Binney, J., Tremaine, S. 1987, *Galactic Dynamics* (Princeton: Princeton Univ. Press).
Blandford, R. D., Gehrels, N. 1999, *Physics Today*, **52,** 40–46.
Blandford, R. D., Rees, M. 1992, in *Testing the AGN Paradigm*, (ed.) S. S. Holt, S. G. Neff, C. M. Urry (New York: American Inst, of Physics), 3–19.
Blandford, R. D., Znajek, R. L. 1977, *Mon. Not. R. Astr. Soc.*, **179,** 433–456.
Bragg, A. E., Greenhill, L. J., Moran, J. M., Henkel, C. 1999, *Astrophys. J.*, submitted.
Braatz, J. A., Wilson, A. S., Henkel, C. 1997, *Astrophys. J. Supp.*, **110,** 321–346.
Braatz, J. A. *et al.* 1999, *Astrophys. J.*, in preparation.
Bromley, B. C., Miller, W. A., Pariev, V. I. 1998, *Nature*, **391**, 54–56.
Cecil, G., Wilson, A. S., Tully, R. B. 1992, *Astrophys. J.*, **390,** 365–377.
Churchwell, E., Witzel, A., Huchtmeier, W., Pauliny-Toth, I., Roland, J, Sieber, W. 1977, *Astr. Astrophys.*, **54,** 969–971.
Claussen, M. J., Diamond, P. J., Braatz, J. A., Wilson, A. S., Henkel, C. 1998, *Astrophys. J.*, **500,** L129–L132.
Claussen, M. J., Heiligman, G. M., Lo, K. Y. 1984, *Nature*, **310,** 298–300.
Claussen, M. J., Lo, K.-Y. 1986, *Astrophys. J.*, **308,** 592–599.
Deguchi, S., Watson, W. D. 1989, *Astrophys. J* , **340,** L17–L20.
dos Santos, P. M., Lepine, J. R. D. 1979, *Nature*, **278,** 34–35.
Elitzur, M. 1992, *Astronomical Masers* (Dordrecht: Kluwer).
Faber, S. M. 1999, in Proceedings of the 32nd COSPAR Meeting; The AGN-Galaxy Connection, (ed.) H. R. Schmitt, L. C. Ho, & A. L. Kinney (Advances in Space Research), in press.
Frank, J., King, A., Raine, D. 1992, *Accretion Power in Astrophysics* (Cambridge: Cambridge Univ. Press).
Gammie, C. F., Narayan, R., Blandford, R. 1999, *Astrophys. J.*, **516,** 177–186.
Gardner, F. F., Whiteoak, J. B. 1982, *Mon. Not. R. Astr. Soc.*, **201,** 13p–15p.
Genzel, R., Eckart, A., Ott, T., Eisenhauer, F. 1997, *Mon. Not. R. Astr. Soc.*, **291,** 219–234.
Ghez, A. M., Klein, B. L., Morris, M., Becklin, E. E. 1998, *Astrophys. J.*, **509,** 678–686.
Greenhill, L. J., *et al.* 1999a, in preparation.
Greenhill, L. J., *et al.* 1999b, in preparation.
Greenhill, L. J., Gwinn, C. R. 1997, *Astrophys. Space Sci.*, **248,** 261–267.
Greenhill, L. J., Henkel, C., Becker, R., Wilson, T. L., Wouterloot, J. G. A. 1995b, *Astr. Astrophys.*, **304,** 21–33.
Greenhill, L. J., Jiang, D. R., Moran, J. M., Reid, M. J., Lo, K. Y., Claussen, M. J. 1995a, *Astrophys. J.*, **440,** 619–627.
Greenhill, L. J., Moran, J. M., Hermstein, J. R. 1997, *Astrophys. J.*, **481,** L23–L26.
Haschick, A. D., Baan, W. A., Peng, E. W. 1994, *Astrophys. J.*, **437,** L35–L38.
Heckman, T. M. 1980, *Astr. Astrophys.*, **87,** 152–164.
Herrnstein, J. R. 1996, Ph.D. thesis, Harvard University.
Herrnstein, J. R., Greenhill, L. J., Moran, J. M. 1996, *Astrophys. J.*, **468,** L17–L20.
Herrnstein, J. R., Greenhill, L. J., Moran, J. M., Diamond, P. J., Inoue, M., Nakai, N., Miyoshi, M. 1998b, *Astrophys. J.*, **497,** L69–L73.

Hermstein, J. R., Moran, J. M., Greenhill, L. J., Blackman, E. G., Diamond, P. J. 1998a, *Astrophys. J.*, **508,** 243–247.
Hermstein, J. R., Moran, J. M., Greenhill, L. J., Diamond, P. J., Inoue, M., Nakai, N., Miyoshi, M., Henkel, C., Riess, A. 1999, *Nature*, **400,** 539–541.
Hermstein, J. R., Moran, J. M., Greenhill, L. J., Diamond, P. J., Miyoshi, M., Nakai, N., Inoue, M. 1997, *Astrophys. J.*, **475,** L17–L20.
Ho, L. C. 1999, in *Observational Evidence for Black Holes in the Universe*, (ed.) S. K. Chakrabarti (Dordrecht: Kluwer), 157–186.
Kartje, J. F., Konigl, A., Elitzur, M. 1999, *Astrophys. J.*, **513,** 180–196.
Kumar, S., Pringle, J. E. 1985, *Mon. Not. R. Astr. Soc.*, **213,** 435–442.
Kormendy, J., Richstone, D. 1995, *Ann. Rev. Astr. Astrophys.*, **33,** 581–624.
Maloney, P. R., Begelman, M. C, Pringle, J. E. 1996, *Astrophys. J.*, **472,** 582–587.
Maoz, E. 1995, *Astrophys. J.*, **455,** L131–L134.
Maoz, E., *et al.* 1999, *Nature*, 401, 351–354.
Maoz, E., McKee, C. 1998, *Astrophys. J.*, **494,** 218–235.
Miyoshi, M. 1999, in *Observational Evidence for Black Holes in the Universe*, (ed.) S. K. Chakrabarti (Dordrecht: Kluwer), 141–156.
Miyoshi, M., Moran, J. M., Herrnstein, J. R., Greenhill, L. J., Nakai, N., Diamond, P. D., Inoue, M. 1995, *Nature*, **373,** 127–129.
Moran, J. 1981, *Bull. Am. Astron. Soc.*, **13,** 508.
Moran, J. M., Greenhill, L. J., Hermstein, J. R., Diamond, P. D., Miyoshi, M., Nakai, N., Inoue, M. 1995, *Proc. Nat. Acad. Sci.*, USA, **92,** 11427–11433.
Nakai, N., Inoue, M., Hagiwara, Y., Miyoshi, M., Diamond, P. J. 1998, in *Radio Emission from Galactic and Extragalactic Compact Sources*, (ed.) J. A. Zensus, G. B. Taylor, & J. M. Wrobel (San Francisco: ASP), 237–238.
Nakai, N., Inoue, M., Miyazawa, K., Miyoshi, M., Hall, P. 1995, *Pub. Astron. Soc. Japan*, **47,** 771–799.
Nakai, N., Inoue, M., Miyoshi, M. 1993, *Nature*, **361,** 45–47.
Neufeld, D., Maloney, P. R. 1995, *Astrophys. J.*, **447,** L17–L20.
Paczynski, B. 1999, *Naturei* **401,** 331–332.
Papaloizou, J. C. B., Terquem, C, Lin, D. N. C. 1998, *Astrophys. J.*, **497,** 212–226.
Reid, M. J., Moran, J. M. 1988, in *Galactic and Extragalactic Radio Astronomy,* 2nd ed., (ed.) G. L. Verschuur, K. I. Kellermann (New York: Springer-Verlag), 255–294.
Rees, M. 1998, in *Black Holes and Relativistic Stars*, (ed.) R. M. Wald (Chicago: Univ. Chicago Press), 79–101.
Salpeter, E. E. 1964, *Astrophys. J.*, **140,** 796–800.
Satoh, S., Inoue, M., Nakai, N., Shibataa, K. M., Kameno, S., Migenes, V., Diamond, P. J. 1998, in *Highlights of Astronomy*, **11b,** (ed.) J. Andersen (Dordrecht: Kluwer), 972–973.
Seyfert, C. 1943, *Astrophys. J.*, **97,** 28–40.
Shakura, N. I., Sunyaev, R. A. 1973, *Astr. Astrophys.*, **24,** 337–355.
Tanaka, Y., *et al.* 1995, *Nature*, **375,** 659–661.
Thompson, A. R., Moran, J. M., Swenson, G. W. 1986, *Interferometry and Synthesis in Radio Astronomy* (New York: Wiley Interscience).
Toomre, A. 1964, *Astrophys. J.*, **139,** 1217–1238.
Trotter, A. S., Greenhill, L. J., Moran, J. M., Reid, M. J., Irwin, J. A., Lo, K. Y. 1998, *Astrophys. J.*, **495,** 740–748.
Weaver, H., Williams, D. R. W., Dieter, N. H., Lum, W. T. 1965, *Nature*, **208,** 29–31.

J. Astrophys. Astr. (1999) **20**, 187–196

The Massive Black Hole at the Galactic Center

A. Eckart & R. Genzel, *Max-Planck Institut fur extraterrestrische Physik, D 85740 Garching, Postfach 1603, Germany*
e-mail: eckart@mpe.mpg.de

Abstract. At the dynamic center of the Milky Way high spatial resolution, near-infrared imaging and spectroscopy have made it possible in the last few years to measure stellar velocities down to separations of less than five light days from the compact radio source SgrA* (in the constellation Sagittarius). These measurements make a compelling case for the presence of a compact, central dark mass of 2.6×10^6 solar masses. Simple physical considerations show that this dark mass cannot consist of a stable cluster of stars, stellar remnants, substellar condensations or a degenerate gas of elementary particles. Energy equipartition requires that at least 10^5 solar masses must be associated with SgrA* itself and is enclosed within less than 8 light minutes (equivalent to 15 Schwarzschild radii of a million solar mass black hole). If one accepts these arguments it is hard to escape the conclusions that there must be a massive black hole at the core of the Milky Way.

Key words. Galactic center—massive black holes—SgrA*.

I. Introduction

At a distance of only 8 kpc the Galactic Center is the closest nucleus of a galaxy, 100 to 1000 times closer than the nearest extragalactic systems. It is thus a unique laboratory in which physical processes that are also relevant for nuclei of other galaxies can be studied with the highest angular resolution possible. Right at the center of the nuclear stellar cluster the compact radio source SgrA* is located. This source is the most likely candidate for a massive black hole at the center of our Galaxy. To determine whether massive black holes at the centers of galaxies exist is of considerable importance for the understanding of 'active galactic nuclei' (AGNs, e.g. quasars) and their evolution in the early Universe. In these objects luminosities up 10^{14} (one solar luminosity corresponds to 4×10^2, watts) are produced within a light year or less of the center. Highly collimated jets of relativistic electrons and rapidly varying X- and γ-ray emission provide strong but indirect evidence that AGNs cannot be powered by stars but by the conversion of gravitational energy to radiation in accretion flows onto massive black holes. For a direct proof of the 'black hole' paradigm it is necessary, however, to determine the characteristic mass concentration and show the existence of an event horizon. Probably the most unambiguous method for carrying out such a proof is the determination of the form of the gravitational potential from the velocity field of stars and gas orbiting the hole candidate. Using this technique ground-based and Hubble Space Telescope

observations of the Doppler shifts of spectral lines from gas and stars have indeed shown that many (and perhaps most) nearby galaxies have massive dark mass concentrations in their nuclei (e.g. Kormendy & Richstone 1995). With the exception of detailed radio VeryLongBaselineInterferometry (VLBI) observations of the galaxy NGC 4258 (Myoshi *et al.* 1995) none of these measurements have high enough resolution yet to prove that the central dark mass must be a black hole and could not be, for instance, a dense compact cluster of stellar remnants. In contrast the nucleus of the Milky Way (distance ~ 8kpc corresponding to 26100 light years) is a thousand times closer than the nearest AGN and one hundred thousand times closer than the nearest quasar. It is thus a unique laboratory to test the black hole paradigm. The Galactic Center, however, is heavily extincted by dust so that observations in the visible are impractical. With the advent of sensitive infrared detectors, high resolution images and imaging spectrometers it has recently become possible to study the stellar velocity field at unprecedented resolution and provide the best evidence yet for a massive black hole at the nucleus of a galaxy.

2. Initial evidence for a mass concentration in the Galactic Center

The first indications for a central mass concentration in the Milky Way emerged in the late seventies from spectroscopic observations of a midinfrared fine structure line of Ne^+ (Wollman *et al.* 1977; Lacy *et al.* 1979). These measurements showed unusually large Doppler shifts (± 250km/s) of ionized gas clouds in the central parsec, toward the maximum stellar density. As radio interferometric observations had discovered a compact, nonthermal radio source, SgrA*, in the same region (Balick & Brown 1974), a plausible interpretation - in analogy to quasars - was that the large gas velocities indicate orbital motions in the vicinity of a million solar mass black hole, coincident with SgrA* (LyndenBell & Rees 1971; Lacy *et al.* 1982) Further infrared (and radio) spectroscopic data taken by various groups in the eighties strengthened the gas dynamic evidence for this central mass concentration (e.g. Serabyn & Lacy 1985; Genzel & Townes 1987) but were not considered compelling by many researchers in the field. In addition to gravitational forces, gas may be affected by magnetic fields, radiation pressure, stellar winds and friction with other gas components - all known to be present in the Galactic Center - thus making the interpretation uncertain. Beginning in the late eighties several groups began measuring radial velocities of late type, red giant and supergiants (e.g. Rieke & Rieke 1988; Seligren *et al.* 1990; Krabbe *et al* 1995; Haller *et al.* 1996). These measurements confirm and strengthen the evidence for the presence of a 1 to 3×10^6, central mass that cannot be accounted for solely by the stellar cluster that is sampled by the near-infrared light.

3. Recent observational evidence

In the last few years it has become possible to measure the stellar velocity field down to scales as small as 5 light days and thus, to be able to place decisive constraints on the nature of the central mass concentration. This significant progress was possible on the one hand due to the discovery (by Forrest *et al.* 1987; Allen *et al.* 1990 and

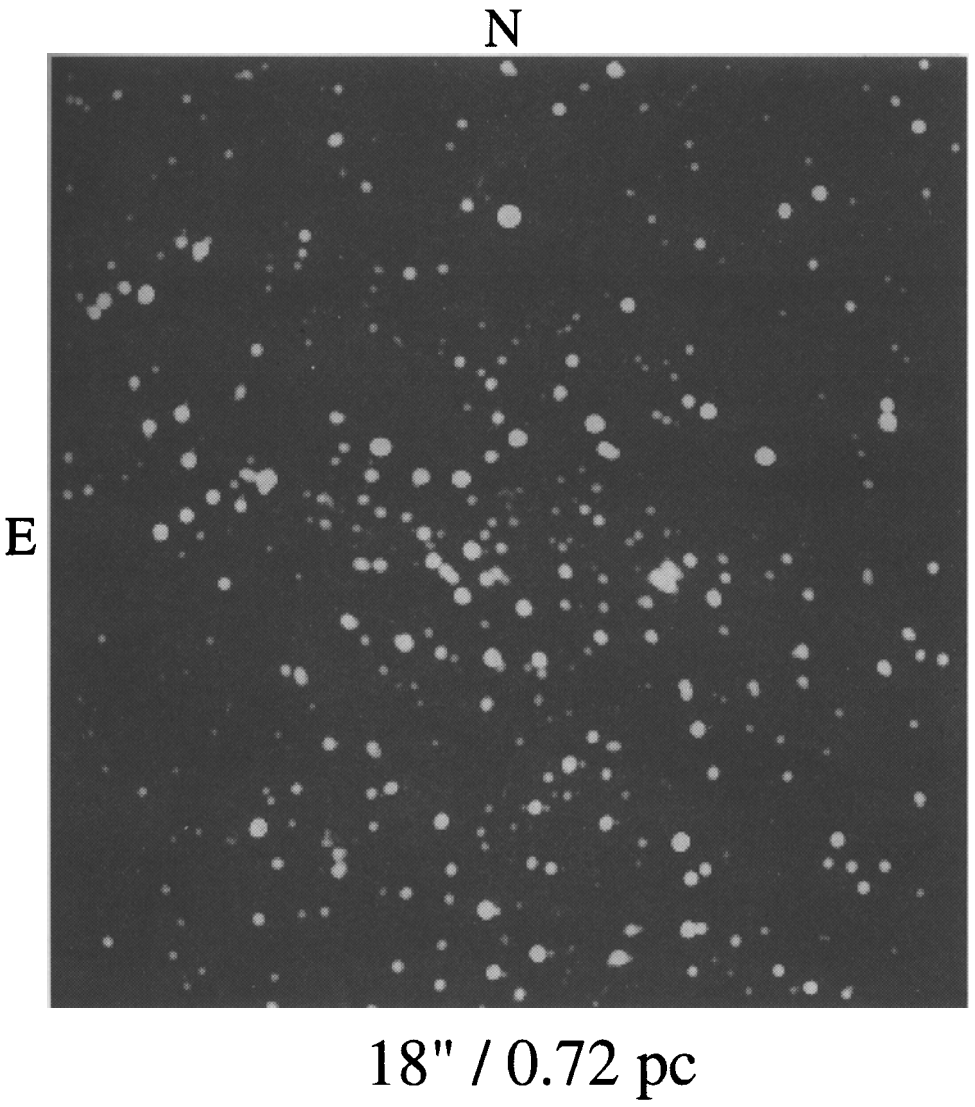

Plate 1.

Krabbe *et al.* 1991) of a compact cluster of hot, luminous emission line stars (the so called 'Ha stars'). Apart from being interesting in their own right (these stars must have fainted in the last few million years and now power the central parsec), they provide radial velocity measurements to a scale of ~ 1″ (0.04 pc). The other and – as it turns out - most important development has been the first measurement of stellar proper motions.

Plate 1 shows a 0.15″ resolution, 2μm image obtained with the MPE 'SHARP' camera on the 3.5 m New Technology Telescope (NTT) of the European Southern Observatory (ESO). The excellent resolution, image quality and high dynamic range of this image (the ratio between weakest and strongest sources is ~ 10^{-4}) are the result of combining 'speckle imaging' techniques (coadding many short exposure

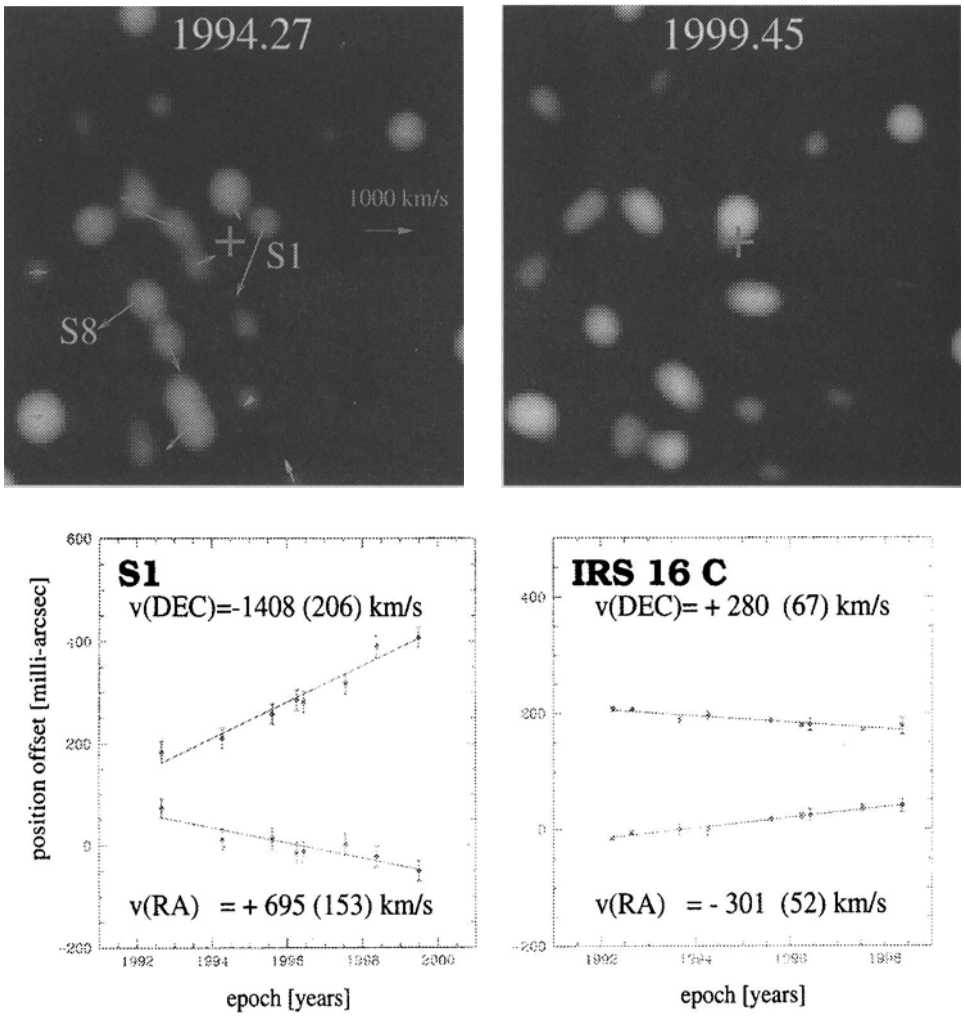

Plate 2.

images) with non-linear CLEANing techniques that remove the very substantial image artifacts in speckle imaging. These near-infrared images show close to 10^3 stars in the central parsec, concentrated and centered on or very near the compact radio source SgrA*. Using a novel field imaging spectrometer, 3D, Genzel *et al.* (1996) have been able to determine radial velocities for about 220 of these stars (see also Krabbe *et al.* 1995; Haller *et al.* 1996). Combining about 60 independent high resolution images between 1992 and 1997; Eckart & Genzel (1996, 1997 and unpublished) and Genzel *et al.* (1997) have derived (relative) proper motions for about 70 stars. Plate 2 shows two examples of the data obtained. In the upper section of Plate 2 the derived proper motion vectors (without error bars) are plotted for a number of stars on the high resolution image, assuming a Sun-Galactic Center distance of 8 kpc.

Of special interest is naturally the immediate vicinity of SgrA* (top part in Plate 2) where one finds a ~ 1″ diameter concentration of faint stars. Several of these stars in this so-called SgrA* cluster show proper motions in excess of 1000 km/s (stars

further out have velocities of only a few hundred km/s), the fastest one (S1: v ~1400 km/s) also being the closest one (~ 0.13″) to SgrA*. This finding is exactly what one would expect if SgrA* were coincident with a large compact mass. Because of the obvious importance of these large motions and the substantial technical difficulties in deriving believable proper motions of faint stars in such a crowded environment, a confirmation of these results was critical. This has now happened. A group from the University of California, Los Angeles has used the 10 m Keck telescope on Mauna Kea, Hawaii, and carried out yet higher resolution 2 μm imagery of the central few arcseconds. Combining data from 3 epochs, 1995, 1996 and 1997, this group fully confirms the very high velocities of the SgrA* cluster stars (Ghez *et al.* 1997). It is also important to ascertain that the observed positional changes indeed present orbital motions in the central gravitational field and that the stars are actually located in the Galactic Center. One can easily show that the positional changes cannot be caused, for instance, as a result of variability in a double or multiple stellar system, or due to a central gravitational lens (Eckart & Genzel 1996, 1997). More random variability of unrelated stars can also cause apparent motions but the consistency of the data within a given epoch and the continuous and steady positional changes over up to 8 epochs exclude such 'Christmas tree effects'. Because of the rapid increase of stellar density toward the Center, contamination by stars in the foreground or background are not very significant when one considers stars outside the central parsec. Projection effects within the central stellar cluster, however, have to be explicitly considered in the analysis (Genzel *et al.* 1997).

3·1 *Speckle spectroscopy*

In order to investigate the nature of the fast moving stars in more detail we obtained spectroscopic data on them. The stars in the central arcsecond close to the position of the radio point source SgrA* (in the following referred to as the SgrA* stellar cluster) are essentially all located within one single seeing disk. In addition they are very close to the bright members of the IRS 16 complex 1.5″ to 2″ east. The members of this complex are 100 times brighter than the stars close to SgrA*. Therefore there is currently no other way to obtain spectral information on these objects but speckle spectroscopy. In order to conduct these observations the SHARP speckle camera was equipped with an objective prism. The additional device consists of two optically contacted prisms. One is made from BaF2 and one made from Schott IRG 9 glass Mounted in front of the SHARP1 camera the two prisms have an on axis net-dispersion that gives full K-band spectra of 1.0" length in the image plane. This dispersion was chosen on purpose in order to minimize confusion problems in the highly crowded area surrounding the compact radio source SgrA*. The spectral range is limited by the K-band filter (1.95–2.40 μm) inside the camera. In this configuration low resolution ($\lambda/\Delta\lambda = 35$) K-band spectra of 4 objects within the central SgrA* stellar cluster could be obtained. These data were taken during NTT observing runs in April and June, 1996. The images were focused by optimizing the short exposure image contrast. Spectrally dispersed speckle interferograms were then recorded with exposure times of 0.5 seconds. We observed the region described above alternately through the full K-band and an He I 2.09 μm narrow band (R = 150) filter. In addition sky data through both filters near the flux calibrator source 9 Sgr were taken. During the April and June 1996 observing runs we took two data sets each with the dispersion

in north-south and east-west direction. For each data set the total number of recorded dispersed speckle images was between 1500 and 2000. The short exposure seeing was of the order of 0.6″ for the April 1996 and between 0.6″ and 1.0″ for the June 1996 run. The raw images were sky subtracted, a flat field was applied and dead pixels were corrected for by interpolation. The spectrally dispersed data can be flat-fielded since the relative spectral response of individual pixels at the chosen spectral resolution is the same to within ≤ 1%. This was checked independently by comparing spectra across the final objective prism images and by comparing flat field data taken through 2.058 μm He I, 2.165 μm Brγ, and 2.29 μm CO narrow band filters with a spectral resolution of R = 150 in 1994 (Eckart *et al.* 1995).

3.2 *Spectra in the central SgrA* cluster*

Since the field around SgrA* is very crowded most of the spectra are contaminated by other nearby stars. The central SgrA* cluster, however, is located west of the IRS 16

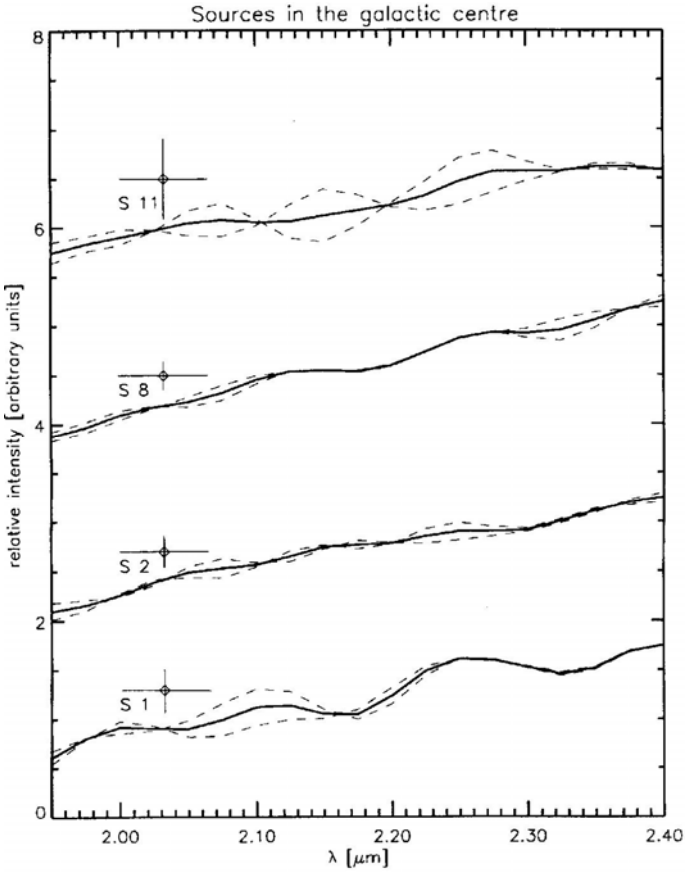

Figure 1. Extracted and calibrated speckle spectra of the sources S1,S2,S8 and S11 obtained from our speckle spectroscopy data as described in the text.The thick black lines represent the mean of the two methods described in the text. The spectral resolution of $\lambda/\Delta\lambda = 35$ as well as the 1σ error are indicated by the cross.

complex in a void of bright stars. The separation to any bright star is larger than 1″ and the contaminating effects of neighboring sources is comparatively small. About half of the SgrA* cluster sources are distributed in north-south direction. This coincidence could be exploited to extract low resolution spectra of individual stars in this region. In Fig. 1 we show the spectra of the sources S1, S2, S8 and S1l obtained with a combination of the two extraction techniques discussed above. The resulting spectra are fairly flat and increase slowly in flux toward longer wavelength. CO band-head absorption is absent so S1, S2, S8 and S1l cannot be late type giant stars. While somewhat redder than the most prominent HeI emission line stars in the IRS 16 complex their spectral characteristics are consistent with early type stars located in the central cluster (i.e. $A_K \sim 3$). There is also no indication for very strong line emission (or absorption) at the wavelength of the He I and/or Brg emission lines. This is not surprising, however, as the narrow HeI/Brγ emission lines cannot be seen even in the classical He I emission line stars at the low spectral resolution. These results are fully consistent with the broad band JHK colors and narrow band filter data presented in Eckart *et al.* (1995). From this data one can conclude that the $m_K \sim 14.5$ sources in the central SgrA* cluster are most likely moderately luminous ($L \sim$ 5,000 to 10,000) early type stars. If they are on the main sequence they would have to be O9–B0.5 stars with masses of 15 to 20.

4. Possible identification of SgrA* in the infrared

A comparison of all available data taken under the best seeing conditions allowed me to produce very deep K-band images of the central 3″ × 3" of the Galactic Center at a very high angular resolution (Eckart *et al.* 1997). This is especially true for the April, June 1996 and in July 1997 observations that were carried out with a 50 mas and 25 mas pixel scale. The high signal to noise in the combined data sets resulted in reproducible images below the diffraction limit of the NTT at an angular resolution of 70 mas (FWHM). A comparison of the images reveals the existence of a $m_K \sim 15$ source (S12) between S1, S2 and S3 in the June 1996 and July 1997 epochs. This source was not detected in August 1992 and April 1994. For these two epochs any source at that position would have been fainter than $m_K \sim 16.3$. Most importantly the location of S12 is coincident with the compact radio source SgrA*, to within the ±30 mas uncertainties of the radioinfrared reference frames (Menten *et al.* 1997). Despite this fact the proper motions of S12 are small when compared with the motions of the surrounding sources in the SgrA* cluster. In addition S12 appears to be a time variable source. The difference of 1.3 magnitudes corresponds to a minimum flux density variability of a factor of 3 to 4. The $m_k = 15.0$ brightness of the newly found source translates into a dereddened flux density of 13 mJy. For the dereddening we assumed that the visual extinction toward the center is $A_v = 30^m$ and the standard reddening law in which $A_K = 0.11 \times A_v$ not including possible local extinction.The upper limit to the flux density in April 1996 then corresponds to about 4mJy.

There are at least two plausible possible interpretations of the data that could explain the nature of the newly found source S 12:

- S12 is the infrared counterpart of the compact radio source SgrA*. Its luminosity and variability could be explained by the theoretical models of advection dominated accretion (Rees *et al* 1982; Melia 1992; Narayan *et al.* 1995, 1997).

- Alternatively the variability could be due to gravitational lensing. In this case the source would be a more distant central cluster member not intimately associated with SgrA* itself. At a distance to the Galactic Center of 8 kpc and an enclosed mass of 2.6×10^6, the size of the Einstein ring within which significant flux density changes and apparent motions due to lensing occur is then of the order of only 10 mas. Depending on the transverse velocity of the source the duration of the corresponding lense effect will only be about one year. Monitoring the flux density and comparing the light curve to theoretical light curves for lensing effects will help to determine the nature of the newly found source.

5. Derived mass distribution

A first rough analysis shows that the projected velocity dispersions of a number of stars in a given annulus of projected radius p increase with $p^{-1/2}$, between $p \sim 1$ pc and $p \sim 0.01$ pc, as expected in the potential of a central point mass (a 'Kepler law'). The location of the largest stellar velocities (the dynamic center), the stellar density maximum and the position of SgrA* (now determined relative to the stars to '30 milli-arcseconds, Menten et al. 1997) all agree to within ± 0.004 pc (0.1″, Ghez et al. 1997). Between $5'' \geq p \geq 1''$ – where both radial and proper motions, sometimes from the same stars, are available—the mean velocities in all three directions agree within the error bars. This means that anisotropy of the stellar orbits – caused, for instance by predominantly very elliptical orbits — does not play a significant role in the Galactic Center.

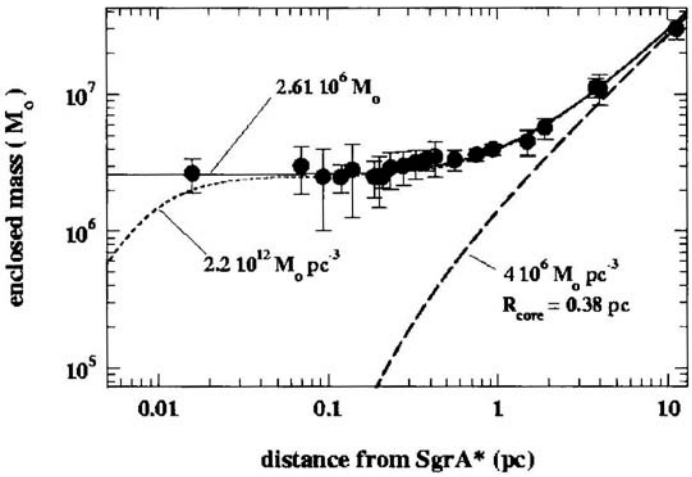

Figure 2. Mass modelling of the stellar proper and radial motions, with the addition of two points from gas kinematics at R = 1.5 and 4pc. Shown as filled circles with 1σ error bars are the various mass estimates listed in Table 2 of Genzel et al. (1997) and discussed in the text, assuming a Sun-Galactic center distance of 8 kpc. The thick dashed curve represents the mass model for the (visible) stellar cluster (M/L(2 μm) = 2, $R_{cor.}$, = 0.38 pc, ρ (R = 0) =4 × 10^6 pc $^{-}$, Genzel et al. 1996). The thin continuous curve is the sum of this stellar cluster, plus a point mass of 2.61×10^6. The thin dotted curve is the sum of the visible stellar cluster, plus a a = 5 Plummer model of a dark cluster of central density 2.2×10^1, pc^{-3}and R_o = 0.0065 pc. It provides a χ^2 fit 1σ worse than the best fit with central density > 7.5×10^1, pc^{-3}.

The final distribution of enclosed mass as a function of true radius (from SgrA*) is shown in Fig. 2 and is the result of applying the so called 'Jeans' equation as well as projected mass estimators to all available stellar radial and proper motion data (Eckart & Genzel 1997; Genzel et al. 1997). The data are fitted extremely well by the combination of a central point mass ($2.61\ [\pm 0.15_{sta.}, \pm 0.35_{stat\ +\ sys}] \times 10^6$,) and a nearly isothermal stellar cluster of core radius ~ 0.38pc and core density 4×10^6, pc^{-3}. The latter is a good fit to the stellar light distribution with a mass to 2 μm-band luminosity ratio of 2 (indicated as a fat dashed line in Fig. 2). The central mass is 'dark', as it has to have a mass to luminosity ratio of 100 or greater. If the central point mass is replaced by a dark cluster its central density has to be in excess of $2 \times 10^{1\cdot}$, pc^{-3} to still be consistent with the data, about 500,000 times greater than that of the visible cluster.

6. Nature of the central mass

Basic considerations on the stability of dark clusters composed of white dwarfs, neutron stars, stellar black holes or sub-stellar entities show that a dark cluster of mass 2.6×10^6, and density 2×10^1, pc^{-3} or greater can not be stable for more than about 10 million years (Maoz 1995; Genzel *et al.* 1997). The majority of the Galactic Center stars, however, are older than 10^8 or 10^9 years. It is also not possible that the dark mass concentration is the core-collapsed state of a dynamically evolving cluster. In that case the distribution -while very dense in a tiny core -would have a soft, quasi-isothermal envelope, unlike what is observed in the Galactic Center. Finally, if the dark mass were conjectured to consist of a degenerate gas of fermions, the m^{-2}, dependence of the Chandrasekhar mass on the mass m of the constituent particles requires that the mass of the fermions cannot be much larger than the electron mass. The only realistic configuration without net electric charge would then be a positron-electron plasma which would, however, rapidly decay through annihilation line radiation. Two further arguments strengthen the conclusion that the dark mass in the Galactic Center in fact must be a black hole. The first comes from the fact that SgrA* itself is known from VLBI measurements to have a proper motion less than about 20 km/s (Backer 1996). In the very dense Galactic Center core the fast moving stars near SgrA* and SgrA* should have approximately the same kinetic energy. The large (factor 100) difference in observed motions combined with the mass estimate of the fast stars from speckle spectroscopy means that SgrA* must be at least 10^4 times more massive than those stars, or 10^5, unless its motion is exactly along the line of sight. If one further assumes that the mass of SgrA* is at least as concentrated as its radio emission (radius $\leq 1.5 \times 10^{13}$ cm, Backer 1996), the inferred density of SgrA* is at least $10^{20\cdot5}$. This lower limit is only five orders of magnitude smaller than the equivalent density of a 2.6×10^6, black hole within its Schwarzschild radius of ~ 10^{12} cm. The second argument is an inversion of the well known dilemma that if SgrA* is a million solar mass black hole it is presently radiating at a rest mass energy to radiation, conversion efficiency of 10^{-5}, to 10^{-6}, considering the accretion of stellar wind gas from its environment (Melia 1992). The only possible way out – apart from very large amplitude variability in the accretion – is the argument that in pure radial (Bondi-Hoyle) or in low density, non-radial flows most of the rest mass energy of the accretion flow can be advected into the hole, rather than radiated away (Rees

et al. 1982; Melia 1992; Narayan et al. 1995, 1997). This explanation then requires the existence of an event horizon and does not work with any configuration other than a black hole (Narayan et al. 1997). Taking all these arguments together it is hard to escape the conclusion that the core of the Milky Way in fact harbors a massive, but presently inactive central black hole.

References

Allen, D. A., Hyland, A. R., Hillier, D. J. 1990, *Mon. Not. R. Astr. Soc,* **244**, 706.

Backer, D. C. 1996, in *Unsolved Problems of the Milky Way*, (eds.) L. Blitz and P. Teuben (Kluwer:Dordrecht), 193.

Balick, B., Brown, R. L. 1974, *Astrophys. J.*, **194**, 265.

Eckart, A., Genzel, R., Hofmann, R., Sams, B., TacconiGarman, L. E. 1995, *Astrophys. J.*, **445**, L26

Eckart, A., Genzel, R. 1996, *Nature,* **383**, 415.

Eckart, A., Genzel, R. 1997, *Mon. Not. R. Astr. Soc*, **284**, 576

Forrest, W. J., Shure, M. A., Pipher, J. L., Woodward, C. A. 1987, in 'The Galactic Center', (ed.) D. Backer, AIP Conf. Proc **155**, 153

Genzel, R., Townes, C. H. 1987, *Ann. Rev. Astr. Ap*. **25**, 377.

Genzel, R., Thatte, N., Krabbe, A., Eckart, A., Kroker, H., TacconiGarman, L. E. 1996, *Astrophys. J.*, **472**, 153

Genzel, R., Eckart, A., Ott, T, Eisenhauer, F. 1997, *Mon. Not. R. Astr. Soc*, **291**, 219.

Ghez, A., Klein, B., Morris, M., Becklin, E. 1997, (in prep).

Haller, J. W., Rieke, M. J., Rieke, G. H., Tamblyn, P., Close, L., Melia, F. 1996, *Astrophys. J.*, 456, 194.

Kormendy, J., Richstone, D. 1995, *Ann. Rev. Astr. Ap*. **33**, 581.

Krabbe, A., Genzel, R., Drapatz, S., Rotaciuc, V. 1991, *Astrophys. J.*, **382**, L19.

Krabbe, A., Genzel, R., Eckart, A., Najarro, F., Lutz, D. et al. 1995, *Astrophys. J. Lett*., **447**, L95.

Lacy, J. H., Baas, F., Townes, C. H., Geballe, T. R. 1979, *Astrophys. J.*, **227**, L17.

Lacy, J. H., Townes, C. H., Hollenbach, D. J. 1982, *Astrophys. J.*, **262,** 120.

LyndenBell, D., Rees, M. 1971, *Mon. Not. R. Astr. Soc*, **152**, 461.

Maoz, E. 1995, *Astrophys. J.*, **447**, L91.

Melia, F. 1992, *Astrophys. J.*, **387**, L25.

Menten, K. M., Reid, M., Eckart, A., Genzel, R. 1997, *Astrophys. J.*, **475**, L 111.

Myoshi, M, Moran, J. M., Hernstein, J., Greenhill, L., Nakai, N., Diamond, R, Inoue, M. 1995, *Nature*, **373**, 127.

Narayan, R., Yi, I, Mahadevan, R. 1995, *Nature*, **374**, 623

Narayan, R., Mahadevan, R., Grindlay, J., Popham, R. G., Gammie, C. 1997 (preprint).

Rees, M., Phinney, E. S., Begelman, M. C, Blandford, R. D. 1982, *Nature*, **295**, 17.

Rieke, G. H., Rieke, M. J. 1988, *Astrophys. J*, **330**, L33 .

Sellgren, K., McGinn, M. T, Becklin, E., Hall, D. N. B. 1990, *Astrophys. J.*, **359**, 112.

Serabyn, E., Lacy, J. 1985, *Astrophys. J.*, **293**, 445.

Wollman, E. R., Geballe, T. R., Lacy, J. H., Townes, C. H., Rank, D. M. 1977, *Astrophys. J.*, **218**, L103.

J. Astrophys. Astr. (1999) **20**, 197–210

Observational Evidence for Stellar Mass Black Holes

Tariq Shahbaz, *University of Oxford*, *Department of Astrophysics*, *Nuclear Physics Building*, *Keble Road*, *Oxford*, *0X1 3RH*, *England*.
e-mail:tsh@astro.ox.ac.uk

Abstract. I review the evidence for stellar mass black holes in the Galaxy. The unique properties of the soft X-ray transient (SXTs) have provided the first opportunity for detailed studies of the mass-losing star in low-mass X-ray binaries. The large mass functions of these systems imply that the compact object has a mass greater than the maximum mass of a neutron star, strengthening the case that they contain black holes. The results and techniques used are discussed. I also review the recent study of a comparison of the luminosities of black hole and neutron star systems which has yielded compelling evidence for the existence of event horizons.

Key words. X-ray sources—black holes.

1. Black holes and the soft X-ray transients

It is widely believed that black holes power AGN and some X-ray binaries, and that they inhabit the nuclei of many normal galaxies. The compulsion to believe that black holes are physical objects is driven by our faith in general relativity and the idea that a black hole is a logical end point of stellar and galactic evolution. The proof for the existence of stellar black holes has been a subject of considerable effort for the last three decades, since the first dynamical mass estimate for the compact object in Cyg X-l (Webster & Murdin 1972; Bolton 1972). More recently observational studies have concentrated on the identification of X-ray features unique to black holes and the direct measurement of the dynamical mass of the compact object.

After the discovery of the first two black holes (Cyg X-l and LMC X-3) involving High Mass X-ray Binaries (HMXBs), almost all the new black hole systems added to the list (see table 1) belong to the class of low mass X-ray binaries (LMXBs)[1]. A large fraction of these are part of the subgroup called the soft X-ray transients (SXTs), sometimes also referred to as *X-ray novae.* These are characterised by massive X-ray outbursts, usually lasting for several months, during which they are discovered by X-ray or γ-ray all-sky monitors such as those on *Ginga*, *Granat*, *GRO* and *RXTE* (see Fig. 1). The high fraction of black holes among SXTs is a natural consequence of the low accretion rates implied by their *evolved* secondaries (King, Kolb & Burderi 1996; see section 3.1).

[1]*High* and *Low* mass refer to the mass of the optical donor star.

Table 1. Black Hole candidates (adapted from White 1994).

Source	Alter. Name	Type	V	X-ray Activity
X1956 + 350*	Cyg X-1	HMXB	9	persistent
X0538 – 641*	LMC X-3	,,	17	persistent
X0540 – 697	LMC X-1	,,	14	persistent
X1659 – 487	GX339 – 4	LMXB	15–21	highly variable
A0620-00*	V616 Mon	,,	18	transient (1917,75)
GS2023 + 338*	V404 Cyg	,,	18	,, (1938, 56, 89)
4U1543 – 47		,,	17	,, (1971, 83, 92)
4U1630 – 47		,,		,, (every $\simeq$ 600 d)
A1524 – 62	KY Tra	,,	>21	,, (1974, 90)
H1743 – 32		,,		,, (1977)
H1705 – 25*	V2107 Oph	,,	21	,, (1977)
EXO 1846 – 031		,,		,, (1985)
GS1354 – 645	BW Cir	,,	22	,, (1987)
GS2000 + 25*	QZ Vul	,,	22	,, (1988)
GS1826 – 24		,,		,, (1988)
GRS1124 – 68*	GU Mus	,,	20	,, (1991)
GRO J0422 + 32*	V518 Per	,,	21	,, (1992)
GRS1009 – 45	N Vel 1993	,,	22	,, (1993)
GRS1716 – 249	V2293 Oph	,,	>22	,, (1993)
GRO J1655 – 40*	N Sco 1994	,,	17	,, (1994)

* Firm candidates supported by dynamical evidence.

The outbursts are due to a thermal instability in the outer regions of the accretion disc (van Paradijs 1996) which occur at mass transfer rates lower than $10^{-8}\ M_{\odot}\ \mathrm{yr}^{-1}$ These outbursts recur on a timescale of decades, but in the interim the SXTs are in a state of *quiescence*, the most unique feature of this class. It is during this period that it becomes possible to observe the companion star directly and perform both optical and IR photometry and spectroscopy. This is not possible in LMXBs in general, as the light is almost totally dominated by the X-ray irradiated accretion disc.

During outburst the X-ray spectra exhibit a hard power-law (with associated flickering) extending to several hundred keV; on which is superposed a softer X-ray component[2] with the exception of a few remarkable cases (e.g. GS2023 + 338; see Fig. 2). Thanks to the recent increase in sensitivity of X-ray detectors it has also become possible to study, in detail the X-ray spectra of several SXTs during quiescence. They are found to be very soft with a blackbody temperature of $T_{bb} \sim$ 0.2 keV (Tanaka & Shibazaki 1996). SXTs are being discovered at a rate of 1–2 per year and, although the number of systems is still small, it has been noted that their galactic distribution (unlike neutron star LMXBs) is consistent with that of Population I objects; they do not cluster towards the galactic center (White 1994), but along the plane of the galaxy. This suggests that the progenitors of the galactic black holes may have been massive Population I OB stars.

In the next section I shall review the high energy spectral characteristics of the black hole systems.

[2] Hence their name, coined so as to distinguish them from the harder Be X-ray transients, which are a completely different class of objects (van Paradijs & McClintock 1995).

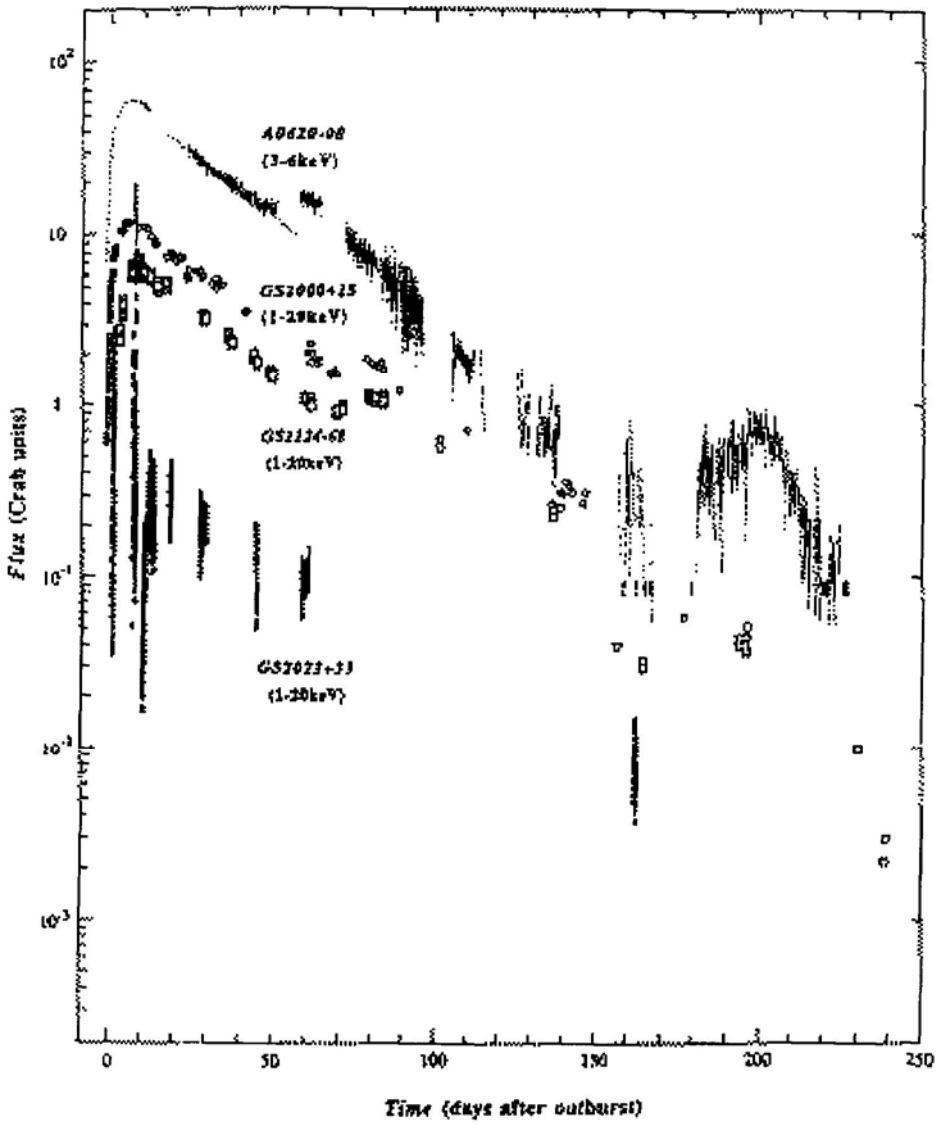

Figure 1. X-ray light curves of four bright soft X-ray transients (from Tanaka & Shibazaki 1996).

2. X-ray observations

The X-ray spectra of the black hole SXTs usually show two components (see Fig. 2). One is a power law with a photon index in the range 1.5–2.5 which dominates at high-energies (> 10 keV) and is occasionally detected up to energies of hundreds of keV. The other component is limited to photon energies below 10 keV, and is often called the "ultra-soft" component which is interpreted as the emission from an optically thick, geometrically thin accretion disc (Tanaka & Shibazaki 1996). It is roughly described by a Planck function with a blackbody temperature of kT_{bb} < 1 keV. It

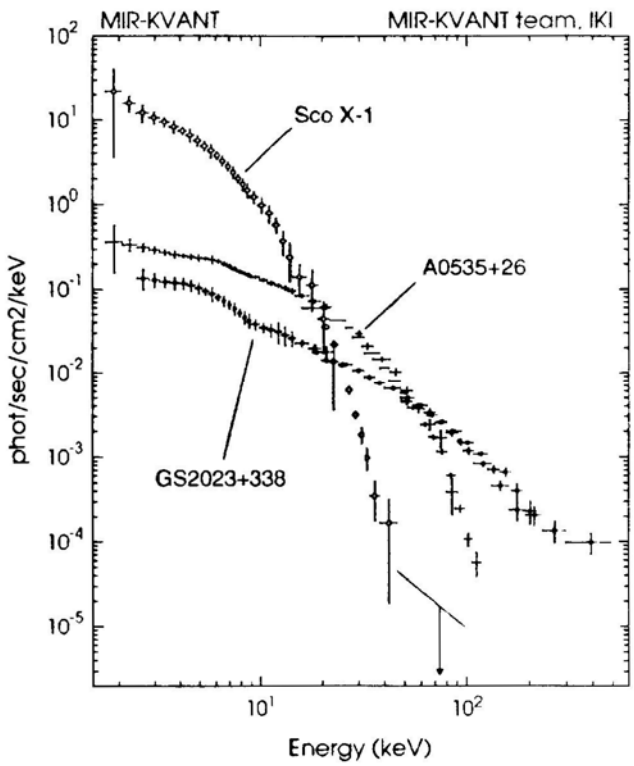

Figure 2. X-ray spectra of several black hole X-ray binaries, showing various combinations of ultra-soft and power-law components (from Gilfanov *et al.* 1995).

should be noted that several important black hole systems, such as GS2023+338 or GRO J0422 + 32, did not exhibit such an ultra-soft component (Tanaka 1989; Sunyaev 1992; see Fig. 2). Also it should be noted that neutron stars can also produce ultra-hard power-law tails when the luminosity drops below a certain limit (Barret & Vedrenne 1994). However, it seems that the high energy tails of the black hole systems might be harder (photon indices $\alpha > 2$) than for neutron star systems (van der Klis 1994).

The expected spectrum i.e. the 'multi-colour' disc spectrum from the accretion disc has been determined (Mitsuda *et al.* 1984). Spectral fits with this model have been made to X-ray spectra obtained with *GINGA* throughout the outbursts of several black hole SXTs (e.g. GS2000+25 and GS1124 – 68). An important feature of these fits was that $R_{\rm in}$ (the inner edge of the disc) remained constant, while the disc luminosity changed by more than an order of magnitude (Tanaka & Shibazaki 1996). The values of $R_{\rm in}$ derived from classical multi-colour blackbody fits are consistent with the innermost stable orbit (3 r_g) around the compact object. For neutron stars $R_{\rm in}$ was found to be systematically lower, by a factor of 3–4 (Tanaka 1992) compared to the black holes. However, the above model was simple, (1) it ignored relativistic effects; gravitational redshift effects near the black hole horizon which depend on the inclination angle of the disc and black hole spin and (2) it assumed that the local disc emission was a Planck function; in the hot inner disc where most of the X-ray are emitted, electron scattering may dominate over free-free absorption and so cause the

colour temperature to be greater than the effective temperature (Ebisawa 1994). A more robust model for the expected luminosity from the accretion disc of black hole SXTs has allowed the actual inner edge of the disc and black hole spin to be determined (Zhang, Cui & Chen 1997).

At higher energies, Granat detected the presence of a transient emission feature in GS112468 at ~ 480 keV (see Fig. 2) which has been interpreted as the 511 keV redshifted annihilation line (Sunyaev *et al.* 1992; Goldwurm *et al.* 1992). Similar features were also observed in Cyg Xl (Ling & Wheaton 1989) and 1E1740.7 – 2942 (Mandrou 1990), a fact which led some authors to propose this as a new black hole signature.

One cannot use any of the above X-ray features as an unambiguous signature of accreting black holes. However, they make a very useful tool in the search for new potential black holes. The most reliable test is still based upon dynamical grounds: the measurement of a mass function exceeding the maximum mass(~ $3M_{\odot}$) beyond which neutron stars become unstable (Friedman, Ipser & Parker 1986).

3. Optical and infrared observations

During the long periods of quiescence when the accretion disc is faint the SXTs are ideal systems to study at optical/IR wavelengths because the low-mass companion star features are strong and observable. This allows firm lower limits to be set on the mass of the compact object by simply obtaining a radial velocity study of the companion star. The dynamical measurement of the mass of the compact object in the black hole SXTs requires both a photometric and spectroscopic study of the companion star.

Using Kepler's law one can derive the binary mass function, which is given by

$$f(M) = \frac{M_1^3 \sin^3 i}{(M_1 + M_2)^2} = \frac{PK_2^3}{2\pi G} \tag{1}$$

where P is the orbital period, M_1 and M_2 are the masses of the compact object and companion star respectively, i is the binary inclination and K_2 is the radial velocity semi-amplitude of the companion. The efficient discovery of these objects by *Ginga*, *GRO* and *RXTE* means that more than a dozen are now known (see Table 1). The observational details for SXTs which have been most intensively studied are summarised in Table 2. GS2023+338 (=V404 Cyg) has the highest mass function (Casares, Charles & Naylor 1992; Casares & Charles 1994) at 6.1 $M_{\odot}$ and so is the best candidate for a system with a black hole.

High spectral resolution observations were first performed on A062000 (=V616 Mon), the prototypical member of this class, which was shown to have a K5V companion star in a 7.8 hr binary orbit; its radial velocity curve is shown in Fig. 3). With a mass function of $f(M)$ = 2.9 $M_{\odot}$ it was immediately recognised as a strong black-hole candidate since, for any reasonable values for the secondary star M_1 is significantly above the maximum mass of a neutron star (McClintock & Remillard 1986, 1990). Attempts have been made to further constrain the system parameters from the gas dynamics around the compact object, but this has led to inconsistent results Johnston, Kulkarni & Oke 1989; Haswell & Shafter 1990). The emission

Table 2. Dynamical mass measurements of SXTs.

Source	P (hrs)	$f(M)$ ($M_\odot$)	$v_{\rm rot}\sin i$ (km s^{-1})	q^*	i (°)	M_X ($M_\odot$)	M_2 ($M_\odot$)	Sp. type
J0422 + 32	5.1	1.21±0.06	≤80	>12	20–40	10±5	0.3	M2v
G1009 – 45	6.9	–	–	–	44±15	–	–	–
A0620 – 00	7.8	2.91±0.08	83	15±1	37±5	10±5	0.6	K3IV
G2000 + 25	8.3	5.01±0.12	86	24±10	56±15	10±5	0.7	K5V
N Mus 91	10.4	3.01±0.15	106	7±2	54^{+20}_{-15}	6^{+5}_{-2}	0.8	K3–4V
N Oph 77	12.5	4.65±0.21	–	–	70±10	7±2	0.4	–
Cen X-4	15.1	0.21±0.08	45	5±1	43±11	1.3±0.6	0.4	K5–7IV
Aql X-1	18.9	–	62^{+30}_{-20}	–	20–30	–	–	K0IV
J1655 – 40	62.6	3.24±0.09	–	3.0±0.1	69±2	5.5±1	1.2	F3–6IV
V404 Cyg	155.4	6.08±0.06	39	17±1	55±4	12±2	0.6	K0IV

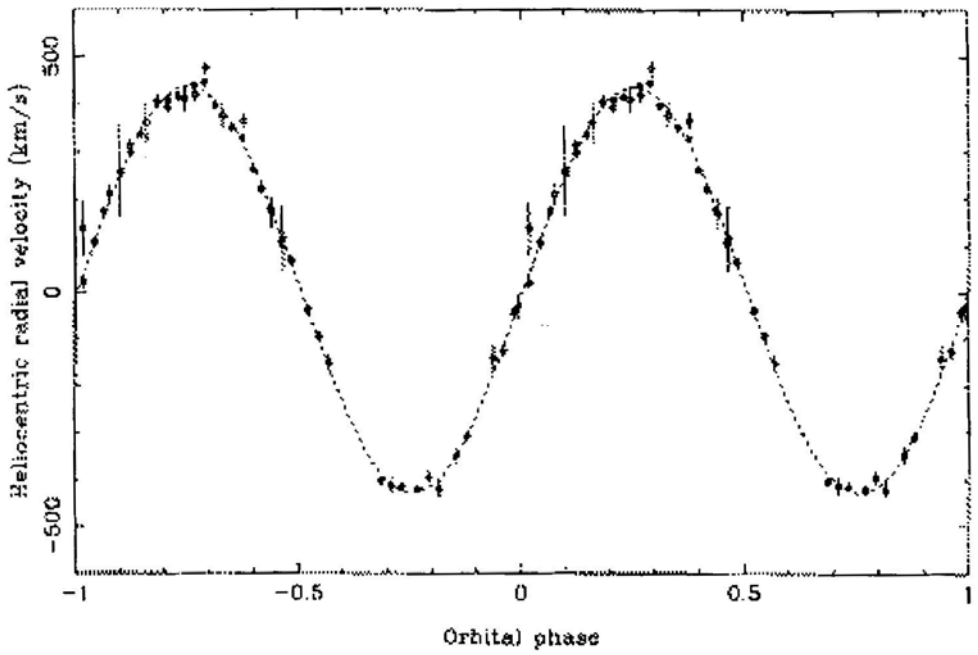

Figure 3. The radial velocity curve of A0620-00 (from Marsh, Robinson & Wood 1994).

profiles are broad and complex, with the presence of multiple components and non-axisymmetric effects which produce a phase misalignment with the companion star even for the extreme line wings (Orosz *et al.* 1994), which ought to arise in material close to the compact object). However, there are two additional observational signatures:

- the rotational broadening of the companion star's absorption spectrum and
- the ellipsoidal modulation of the companion star's light curve,

which can be combined with $f(M)$ [equation (1)] to yield a complete solution of the binary system parameters. With late-type, low-mass secondaries, the SXTs therefore provide the only method for obtaining accurate masses of systems in which the compact objects are suspected of being black holes. Determining these masses

provides a crucial challenge for understanding the late evolution of massive stars and the formation of supernova remnants (Verbunt & van den Heuvel 1995).

In the next section I will present the methods used to determine the mass of the compact object from studies at optical and IR wavelengths (for a detailed review on particular systems see van Paradijs & McClintock 1995). I will use the optical designations of the stars, as given in Table 2.

3.1 *The nature of the companion star*

Since the companion stars in the SXTs must be filling their Roche lobes in order to be transferring matter onto the compact object, we can assume that their size is given by (Paczynski 1971)

$$\frac{R_2}{a} = 0.46(1+q)^{-1/3} \tag{2}$$

where the mass ratio $q = M_1/M_2$. Combining this with Kepler's 3rd Law leads to the well-known result that the companion star's mean density $\overline{\rho} = 110/P^2_{\rm hr}$ gcm^{-3}. As the mean density of a K-type main sequence star is ~5 gcm^{-3}, only V616 Mon and GU Mus can be (relatively) unevolved stars. It is on this basis that the luminosity classes are given in Table 2, since there are no suitable luminosity discriminant within existing spectra. Computations of the evolution of such a star in a close binary lead to the concept of a "stripped-giant" (King 1993) and a constraint on the mass of the companion star to lie within the range $0.2 < M_2 < 1.3\ M_\odot$

3.2 *The rotational broadening of the companion star*

If the size of the companion star is constrained in the size of its Roche lobe, then the requirement of co-rotation in these close binaries leads to the result that its rotational velocity is given by Wade & Home (1988)

$$v_{\rm rot} \sin i = 0.46 K_2 \frac{(1+q)^{2/3}}{q}. \tag{3}$$

Hence q can be obtained directly from the radial velocity curve and the observed rotational broadening of the companion star's absorption lines. This is technically challenging as typical values of $v_{\rm rot}$ sini are 40–100 km s^{-1}, so high spectral resolution (≤ 1 Å) is needed. Given the faintness of these objects (V ≤18–20), large telescopes are required even for the brightest SXTs (Filippenko, Matheson & Barth 1995; Filippenko *et al.* 1995).

Using the 4.2 m *William Herschel Telescope* and a spectral resolution of less than 1Å covering the range λλ6400–6600, the rotational broadening of the companion star can be determined (see Fig. 4). The rotational velocity is determined by subtracting different broadened versions of the template star and performing a χ^2 test on the residuals (the broadening includes the effects of rotation and limb darkening). This analysis also determines the amount of continuum excess present due to the accretion disc (the *veiling factor*). The results of this kind of spectroscopic study are given in Table 2.

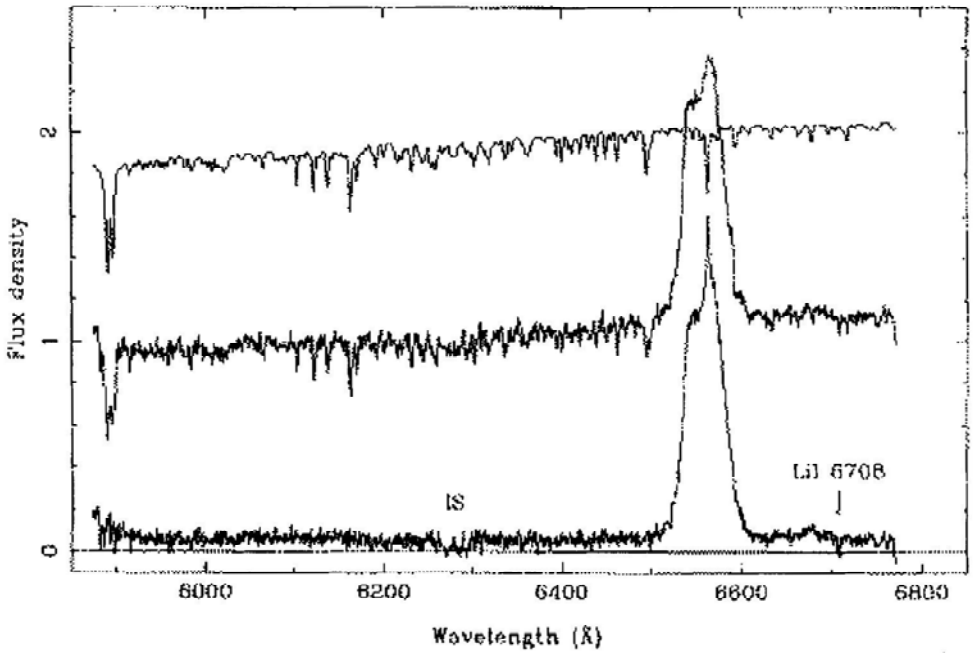

Figure 4. Determining the rotational broadening in A0620-00. From bottom to top: the K3V template star; the same spectrum broadened by 83kms^{-1}; Doppler corrected sum of A0620-00 (dominated by intense *Hα* emission from the disc); residual spectrum after subtraction of the broadened template (from Marsh, Robinson & Wood 1994).

From equation (1) one can see that one needs to determine the binary inclination if one wants to obtain the mass of the compact object. This can be done by exploiting the ellipsoidal modulation of the companion star.

3.3 *The ellipsoidal modulation of the companion star*

The characteristic double-humped modulation on the orbital period, is seen at optical/IR wavelengths in many SXTs (see Table 2). The variations arise because the observer sees differing aspects of the gravitationally distorted star as it orbits the compact object.

Interpreting the optical light curves is difficult as there is complicated structure in the optical light curves (Haswell *et al.* 1993; Wagner *et al.* 1992; Remillard, McClintock & Bailyn 1992). It is likely that there is some contamination by a combination of the accretion disc (e.g. the disc itself and the stream impact region), X-ray heating (although this should be small in quiescence) and possible starspots on the surface of the companion star (cf. RS CVn-type activity). The accretion disc contamination in the optical has been measured; e.g. in A0620-00 it is 30–50% at 5500 Å (McClintock & Remillard 1986) and 6% at 6600 Å (Marsh, Robinson & Wood 1994). Therefore, one expects the contamination to be little in the IR. This is what is observed; in the case V404 Cyg the accretion disc contamination is very small (Shahbaz *et al.* 1996).

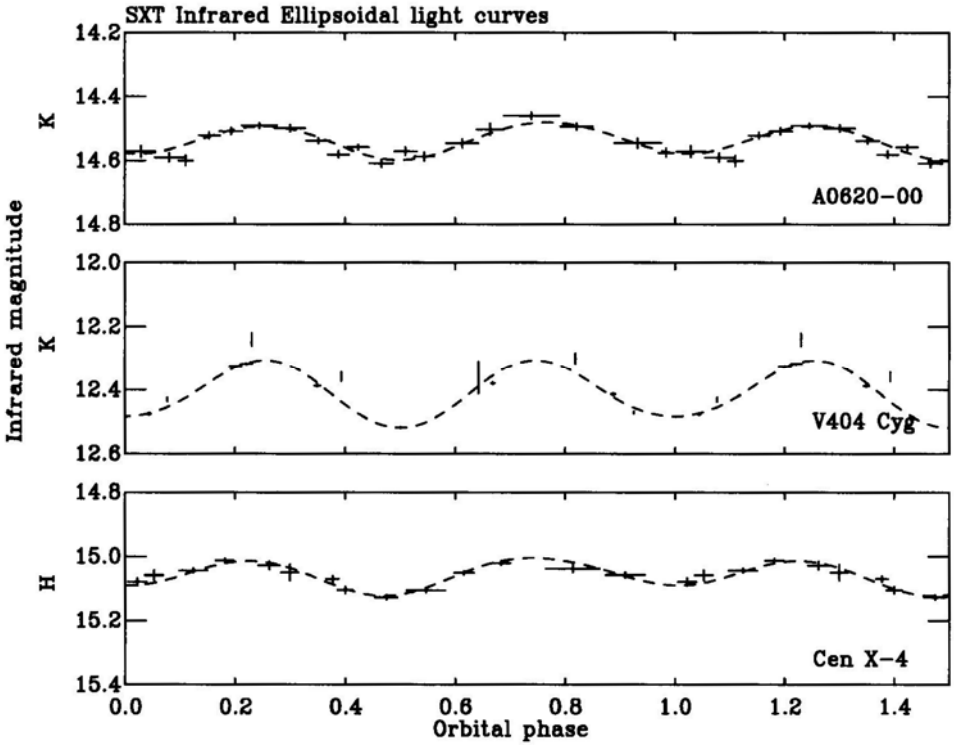

Figure 5. IR light curves of three soft X-ray transients (Shahbaz, Naylor & Charles 1993, 1994; Shahbaz *et al.* 1994).

A campaign of photometry was therefore undertaken in the IR; where the disc contamination is less and where the limb and gravity-darkening are less affected by uncertainties in $T_{\rm eff}$ JHK photometry from UKIRT (with IRCAM) and the AAT (with IRIS) yielded the first IR light curves for the SXTs (Shahbaz, Naylor & Charles 1993, 1997; Shahbaz *et al* 1994; see Fig. 5).

Model calculations show that the shape and amplitude of the modulation are a function of q and i. The difference between the two minima is a result of gravity darkening; the companion star's inner face is heavily gravity darkened and so emits less flux. The amplitude of the modulation is a strong (increasing) function of i, but it is relatively insensitive to q (for $q \geq 5$). Thus by fitting the ellipsoidal IR light curves of the SXTs (where the light arises from the companion star) with the model, the binary inclination can be determined (see Table 2).

The power of this technique is most clearly demonstrated for those systems (V404 Cyg, V616 Mon and GU Mus) in which the rotational broadening has been detected. This constrains q very tightly, precisely in the range where the ellipsoidal modulation is less sensitive. The latter, however, is able to tightly constrain i, and hence we are able to obtain the first direct mass measurements for black holes in our galaxy. Note that for Cen X-4, the compact object mass in the range $0.5–2M_{\odot}$. This is in excellent accordance with the expected mass of a neutron star, given the observation of an X-ray burst from Cen X-4 during its 1979 X-ray outburst (Matsuoka *et al.* 1980), and provides a useful confirmation of this basic approach for mass determination.

An independent method of estimating i and, therefore, testing these results would consist of modelling the shape of the absorption line profiles (Shahbaz 1998), which also provides a direct confirmation of the geometrical distortion of the companion star. This technique has the advantage over the IR ellipsoidal modulation of not being affected by any residual veiling (Sanwal *et al.* 1996; Shahbaz *et al.* 1996). However a new generation of large telescopes (8–10m) is needed before the method can be applied to the quiescent SXTs.

4. GRO J165540: The first eclipsing system

Nova Sco 1994 (=GRO Jl655–40) was discovered by *GRO* in July 1994 with a peak intensity of 0.7 Crab. After its initial outburst event, this system showed repeated (smaller) outbursts separated by $\simeq$ 120 days (Zhang *et al.* 1995). Radio outbursts were also detected, with associated superluminal radio-jets almost perpendicular to the line of sight (85 deg; Hjellming & Ruppen 1995). Nova Sco 1994 is thus one of the two galactic analogies of radio-quasars, the other being the extremely reddened X-ray transient GRS1915+105 (Mirabel & Rodriguez 1994). [With A_v = 26.5 the optical counterpart of GRS1915+105 is undetectable. Based on recent IR spectroscopy, it has been proposed that this system is a HMXB containing a late Oe or early Be mass donor. The nature of the compact object remains uncertain; Mirabel *et al.* 1997.]

With a quiescent magnitude of V = 17.3, the brightest quiescent magnitude of any SXT, the optical counterpart of Jl655–40 was promptly studied intensively. This lead to the discovery of a high mass function, $f(M) = 3.24 \pm 0.09\ M_{\odot}$. (Bailyn *et al.* 1995; Orosz & Bailyn 1997) which implies the presence of a black hole primary. The companion star has a spectral type of F36 Iv, the earliest known amongst the SXT mass donors, and orbits the compact object with a period of 2.6 days.

Photometry performed during outburst revealed the existence of optical eclipses (Bailyn *et al.* 1995), although these were not detected in hard X-rays (Harmon *et al.* 1995). From the absence of X-ray eclipses an upper limit to the inclination of $i < 76$ deg can be set. The optical light curves, obtained during the decline phase, were modelled including X-ray heating effects and mutual eclipses of the accretion disc and the companion star. The solutions support a binary inclination in the range 65–76 deg (van der Hooft et al. 1997). The quiescent light curves shows the classical ellipsoidal modulation of the companion, model fits to which (Orosz & Bailyn 1997) yield an inclination of $i = 69.5 \pm 0.1$ deg and a mass ratio of $q = 3.0 \pm 0.1$. These are combined with the mass function to give a black hole mass $Mx = 7.0 \pm 0.2\ M_{\odot}$, the most accurately determined yet.

It should also be noted that Nova Sco 1994 is also unique because of its high systemic velocity (–155 kms^{-1}) which has important implications for possible formation scenarios of the black hole (Brandt, Podsiadlowski & Sigurdsson 1994).

5. The ADAF model and evidence for event horizons

In the standard picture of an X-ray binary, we observe the emission from a geometrically-thin, optically thick accretion disc extending down to the last stable orbit around the compact object. However, in the black hole SXTs the standard

picture disc to too cool to radiate at the temperature of 0.2 keV as is observed in the black hole SXTs; it is impossible to fit the X-ray luminosity and spectra with any Standard disc model. The data suggests that the gas near the centre must be very hot (the mass accretion rate determined in the outer parts of the disc by the optical luminosity is much higher than the rate determined in the inner parts of the disc from the X-ray luminosity). Narayan (1996; see also Narayan, Barret & McClintock 1997) proposed a two zone model. The outer regions of the disc follows the standard picture; it radiates efficiently and produces the optical/UV flux. The inner disc at 10^4 R_s (R_s is the Schwarzschild radius) however, is hot and optically thin and the flow is advection dominated. In an advection dominated-accretion flow (ADAF) a large fraction of the viscosity generates heat is advected with the accreting gas and only a small fraction of energy is radiated. The ADAF model has been applied very successfully in fitting the broadband spectra (optical to X-ray of the SXTs in quiescence. The bulk of the X-ray and optical emission is produced within $10R_s \sim 300$ km of the black hole.

When a hot ADAF flow with its extremely low radiative efficiency encounters a quiescent black hole, then the enormous thermal energy stored in the gas simply disappears through the event horizon. What about a neutron star accretor? It is reasonable to expect that a quiescent SXT with a neutron star primary (e.g. Cen X-4) will also accrete via an ADAF flow with low radiative efficiency(~0.05%). The energy of the superheated gas cannot disappear, the gas falls on the neutron star and heats its surface. Once a steady sate is reached, the stars luminosity will be the same as for a thin disc and the efficiency will be ~10% higher, than for the black hole. Thus for the same mass accretion rate in quiescence we expect a black hole to be substantially less luminous than a neutron star, precisely as is observed (see Fig. 6).

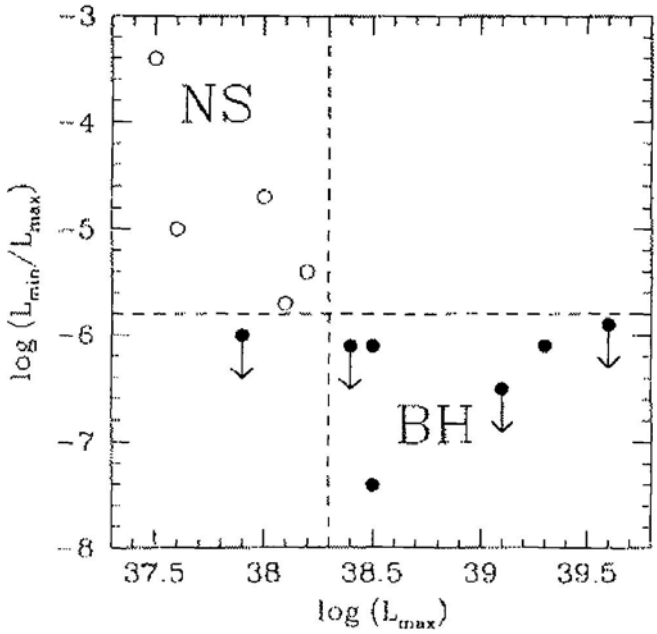

Figure 6. Evidence for ADAF flows in black hole SXTs. The figure is taken from Narayan *et al.* (1997).

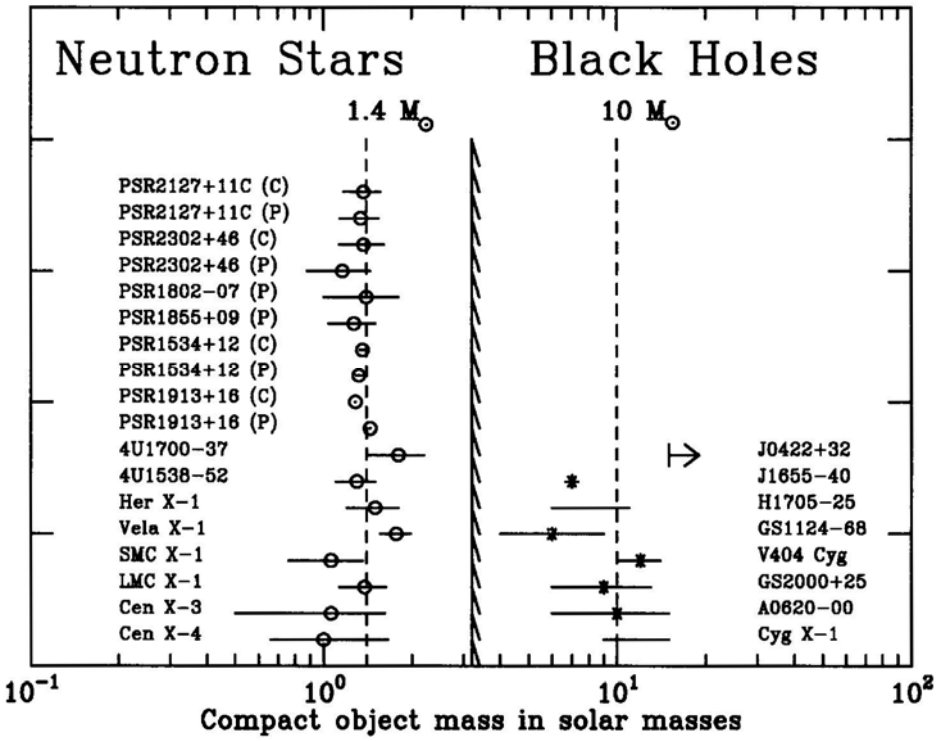

Figure 7. Mass distribution of neutron stars and black holes for which masses have been directly measured. Also shown are the X-ray pulsars (Thorsett et al, 1993). The neutron star systems lie at ~1.4 $M_{\odot}$. , whereas the black-hole candidates seem to cluster around ~10 $M_{\odot}$. The strongest black hole candidates are V404 Cyg and J1655–40. The vertical hashed line represents the Rhoades/Ruffini limit of 3.2*M*. (Rhoades & Ruffini 1974). The figure is taken from Miller, Shahbaz & Nolan (1998).

6. Summary

Observations of the masses of stellar remnants after supernova explosions are essential for an understanding of the core collapse of massive stars. Such masses have been known accurately only for X-ray and radio pulsars (all of which are < $2M_{\odot}$). However, there are now several stellar-mass black hole candidates whose minimum masses exceed the canonical neutron star maximum of $3M_{\odot}$. These are shown in Fig. 7.

Although there are a variety of X-ray signatures e.g. soft spectrum, hard power-law tail, X-ray variability, that suggest an SXT may harbour a black hole, it is clear that the only secure way to determine the mass of the compact object is through the study of its gravitational influence on its low-mass companion star. The radial velocity curves of the low-mass companion star in the SXTs gives a firm lower limit to the mass of the compact object. IR photometry of the ellipsoidal variation of the companion star gives the binary inclination, which when combined with the rotational broadening of the companion star, allows one to determine the actual mass of the binary components. By obtaining more compact object masses in X-ray binaries we will be able to put constraints on both the equation of state of nuclear

matter (Cook, Shapiro & Teukolsky 1994) and the formation theories of black holes (Brown & Bethe 1994; Timmes, Woosley & Weaver 1996). This will start to become possible in the next few years, with the advent of a new generation of 10 m class telescopes.

Finally, the dramatic difference in luminosity between a hot ADAF flow which falls down a black hole and one that strikes the surface of a neutron star provides the first evidence for an event horizon.

References

Bailyn, C. D., Orosz, J. A., McClintock, J. E., Remillard, R. A. 1995, *Nature*, **378**, 157.
Barret, D., Vedrenne, G. 1994, *Astrophys. J. Supp. Series*, **92**, 505.
Bolton, C. T. 1972, *Nature*, **235**, 271.
Brandt, W. N., Podsiadlowski, Ph., Sigurdsson, S., 1994, *Mon. Not. R. Astr. Soc*, **277**, L35.
Brown, G. E., Bethe, H. A. 1994, *Astrophys. J.*, **423**, 659.
Casares, J., Charles, P. A., Naylor, T. 1992, *Nature*, **355**, 614.
Casares, J., Charles, P. A. 1994, *Mon. Not. R. Astr. Soc*, **271**, L5.
Cook, G. B., Shapiro, S. L., Teukolsky, S. A. 1994, *Astrophys. J.*, **424**, 823.
Ebisawa, K. 1994, in *The Evolution of X-ray Binaries*, (ed.) S. S. Holt & C. S. Day (AIP), **308,** pl43.
Filippenko, A. V., Matheson, T., Barth, A. J. 1995, *Astrophys. J.*, **455**, L139.
Filippenko, A. V., Matheson, T, Ho, L. C. 1995, *Astrophys. J.*, **455**, L614.
Friedman, J. L., Ipser, J. R., Parker, L. 1986, *Astrophys. J.*, **304**, 115.
Gilfanov, M., *et al.* 1995, in Alpar *et al.* (eds), *p* 331.
Goldwurm, A., et al. 1992, *Astrophys. J.*, **389**, L79.
Harmon, B. A., *et al.* 1995, *Nature*, **374**, 703.
Haswell, C. A. Shafter, A. W. 1990, *Astrophys. J.*, **359**, L47.
Haswell, C. A., *et al.* 1993, *Astrophys. J.*, **411**, 802.
Hjellming, R. M., Ruppen, M. P. 1995, *Nature*, **375**, 464.
Johnston, H. M., Kulkarni, S. R., Oke, J. B. 1989, *Astrophys. J.*, **345**, 492.
King, A. R., 1993, *Mon. Not. R. Astr. Soc*, **260**, L5.
King, A. R., Kolb, U., Burderi, L. 1996, *Astrophys. J.*, **464**, L127.
Ling, J. C, Wheaton, W. A. 1989, *Astrophys. J.*, **343**, L57.
McClintock, J. E., Remillard, R. A. 1986, *Astrophys. J*, **308**, 110.
McClintock, J. E., Remillard, R. A. 1990, *Astrophys. J.*, **350**, 386.
Marsh, T. R., Robinson, E. L., Wood, J. H. 1994, *Mon. Not. R. Astr. Soc*, **266**, 137.
Mandrou, P, 1990, *IAU Circ.* 5032.
Matsuoka, M., *et al* 1980, *Astrophys. J.*, **240**, L137.
Miller, J. C, Shahbaz, T., Nolan, L. A. 1998, *Mon. Not. R. Astr. Soc*, **294**, L25.
Mirabel, I. F., Rodriguez, L. F. 1994, *Nature*, **371**, 46.
Mirabel, I. F., Bandyopadhyay, R., Charles, P. A., Shahbaz, T, Rodriguez, L. F. 1997, *Astrophys. J.*, **477**, L45.
Mitsuda, K., *et al.* 1984, *Pub. Astro. Soc. Japan*, **36**, 741.
Narayan, R. 1996, *Astrophys. J.*, **462**, 136.
Narayan, R., Barret, D., McClintock, J. E. 1997, *Astrophys. J.*, **482**, 448.
Orosz, J. A., Bailyn, C. D., Remillard, R. A., McClintock, J. E., Foltz, C. B. 1994, *Astrophys. J.*, **436**, 848.
Orosz, J. A., Bailyn, C. D. 1997, *Astrophys. J.*, **477**, 876.
Paczynski, B.1971, *Ann. Rev. Astron. Astrophys*., **9**, 183.
Remillard, R. A., McClintock, J. E., Bailyn, C. D. 1992, *Astrophys. J.*, **399**, L145.
Rhoades, C. E., Ruffini, R. 1974, *Phys. Rev. Lett*., **32**, 324.
Sanwal, D., *et al.* 1996, *Astrophys. J.*, **460**, 437.
Shahbaz, T, Naylor, T, Charles, P. A. 1993, *Mon. Not. R. Astr. Soc*, **265**, 655.
Shahbaz, T., Naylor, T., Charles, P. A. 1994, *Mon. Not. R. Astr. Soc*, **268**, 756.

Shahbaz, T., *et al.* 1994, *Mon. Not. R. Astr. Soc*, **271**, L10.
Shahbaz, T., Bandyopadhyay, R., Charles, P. A., Naylor, T. 1996, *Mon. Not. R. Astr. Soc*, **282**, 977.
Shahbaz, T., Naylor, T., Charles, P. A. 1997, *Mon. Not. R. Astr. Soc.*, **285**, 607.
Shahbaz, T. 1998, *Mon. Not. R. Astr. Soc*, **298**, 153.
Sunyaev, R. A., *et al.* 1992, *Astrophys. J.*, **389**, L75.
Tanaka, Y. 1989, in *Two Topics in X-ray Astronomy*, (ESA SP296, Paris), p 3.
Tanaka, Y. 1992, in *Ginga Memorial Symp.*, (eds) F. Makino & F. Nagase, p 19.
Tanaka, Y, Shibazaki, N. 1996, *Ann. Rev. Astron. Astrophys.*, **34**, 607.
Thorsett, S. E., Arzoumanian, Z., McKinnon, M. M., Taylor, J. H. 1993, *Astrophys. J.*, **405**, L29.
Timmes, F. X., Woosle, Y. S. E., Weaver, T. A. 1996, *Astrophys. J.*, **457**, 834.
van der Hooft F., *et al.* 1997, *Mon. Not. R. Astr. Soc*, **286**, L43.
van der Klis, M. 1994, *Astrophys. J. Supp. Series*, **92**, 511.
van Paradijs, J., McClintock, J. E. 1995, in *X-Ray Binaries* (eds.) W. H. G. Lewin, J. van Paradijs & E. P. J. van den Heuvel (CUP 26, Cambridge), p. 58.
van Paradijs, J. 1996, *Astrophys. J.*, **464**, L139.
Verbunt, F., van den Heuvel, E. P. J. 1995, in *X-Ray Binaries* (eds.) W. H. G. Lewin, J. van Paradijs & E. P. J. van den Heuvel (CUP 26, Cambridge), ρ 457.
Wade, R. A., Horne, K. 1988, *Astrophys. J.*, **324**, 411.
Wagner, M. R., Kreidl, T. J., Howell, S. B., Starrfield, S. G. 1992, *Astrophys. J.*, **401**, L97.
Webster, B. L., Murdin, P. 1972, *Nature*, **235**, 37.
White, N. E. 1994, in *The Evolution of X-ray Binaries*, (ed.) S. S. Holt & C. S. Day (AIP) **308**, p 53.
Zhang, S. N., Harmon, B. A., Paciesas, W. S., Fishman, G. J. 1995, *IAU Circ*, **6209**.
Zhang, N. S., Cui, W., Chen, W. 1997, *Astrophys. J.*, **482**, L155.

J. Astrophys. Astr. (1999) **20**, 211–220

Gravitational Waves: The Future of Black Hole Physics

B. S. Sathyaprakash, *Cardiff University*, *5*, *The Parade, Cardiff, CF2 3YB*, *U.K.*
email: B. Sathyaprakash @ astro. cf. ac. uk

Abstract. The new millennium will witness the operation of several long-baseline ground-based interferometric detectors, possibly a space-based detector too, which will make it possible to directly observe black holes by catching gravitational waves emitted by them during their formation or when they are perturbed or when a binary consisting of black holes in-spirals due to radiation reaction. Such observations will help us not only to test some of the fundamental predictions of Einstein's general relativity but will also give us the unique opportunity to map black hole spacetimes, to measure the masses and spins of black holes and their population, etc.

Key words. Black holes—gravitational waves—testing general relativity.

1. Introduction

Recent observations of nuclei of galaxies have revealed large mass concentrations over relatively small volumes (Valtoja & Valtonen 1989; Roos, Kastra & Hummel 1993; Kormendy & Richstone 1995; Bender, Kormendy & Dehnen 1996; Eckart & Genzel 1996; Gaskell 1996; Van der Marel *et al.* 1996; Rees 1997) The only plausible explanation for high compactness of the mass distribution is that galactic nuclei, including the Milky Way, contain a massive or a super-massive (10^6–$10^9 M_\odot$) black hole. There are also astronomical observations of stellar mass black holes in our own Galaxy (see talks by Moran, Eckart and Shahbaz in this volume). However, these are not direct observations of black holes. Classical black holes do not emit electro-magnetic waves and cannot, therefore, be observed directly in the electro-magnetic window. Their presence is inferred by observing physical processes occurring in their vicinity. Perturbed black holes do emit gravitational waves with a characteristic spectrum thus making it possible to observe them directly.

The rest of this article is organised as follows. We shall begin in section 2 with a short discussion of interferometric detectors of gravitational waves that are either under construction or are being planned. In section 3 we shall briefly review the black hole sources of gravitational waves which include in-spiral waves from binaries and quasi-normal mode ringing of black holes. In section 4 we shall discuss how observation of gravitational radiation from these sources helps us in testing general relativity, making accurate measurements of sources' parameters and measuring cosmological parameters. We use a system of units in which $G = c = 1$.

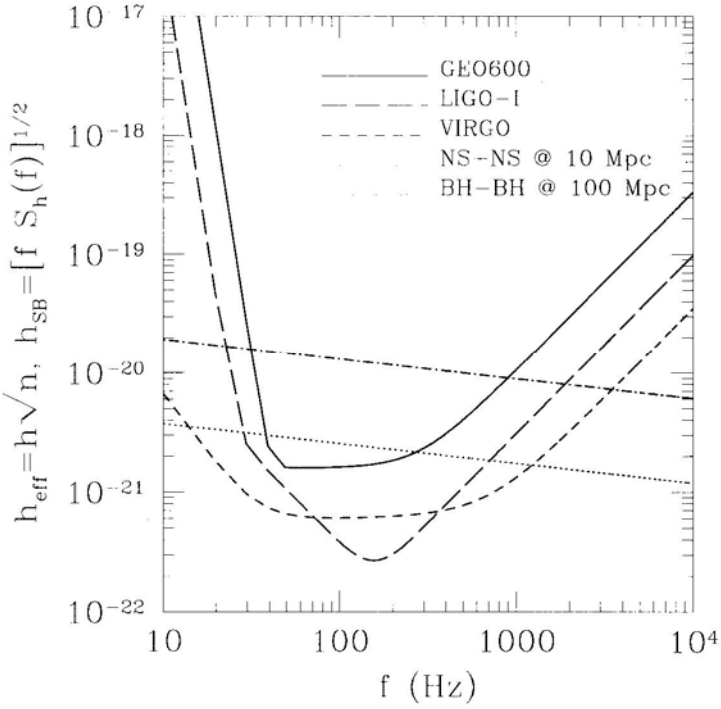

Figure 1. Sensitivity to bursts (SB) of ground-based interferometers and effective signal strengths of in-spiral signals.

2. Interferometric gravitational wave detectors

Several ground-based interferometric detectors of gravitational waves (Abramovici *et al.* 1992; Tsubono 1995; Caron *et al.* 1997; Lück *et al.* 1997) will come on-line in a few years from now and the laser interferometer space antenna (Bender *et al.* 1996) (LISA) -a joint ESA-NASA venture -might fly as early as 2009 (Bender *et al.* 1998). While the ground-based interferometers will have a good sensitivity in the frequency range of 1-1000 Hz, space-based detectors will cover the range 10^{-4}–10^{-1} Hz. These antennas are essentially omni-directional with their response better than 50% of the average over 75% of the sky.

Figures 1 and 2 show the instrumental noise expected in ground-based antennas and LISA, respectively, together with some of the sources they may see. The curves compare the amplitude noise spectral density per logarithmic bin with the effective signal strength[1]. For both ground- and space-based antennas black holes are one of the most prominent sources. In-spiral and merger of binaries consisting of stellar mass (0.5–100 $M_\odot$) compact objects (neutron stars or black holes) are candidate

[1]Nonmonochromatic signals whose time evolution is accurately known, e.g. binary in-spiral waves, can be integrated over time and their signal-to-noise can be improved, as compared to a simple Fourier transform, by the square root of the number of cycles observed. For such signals the effective strength is the modulus of the Fourier amplitude multiplied by the frequency.

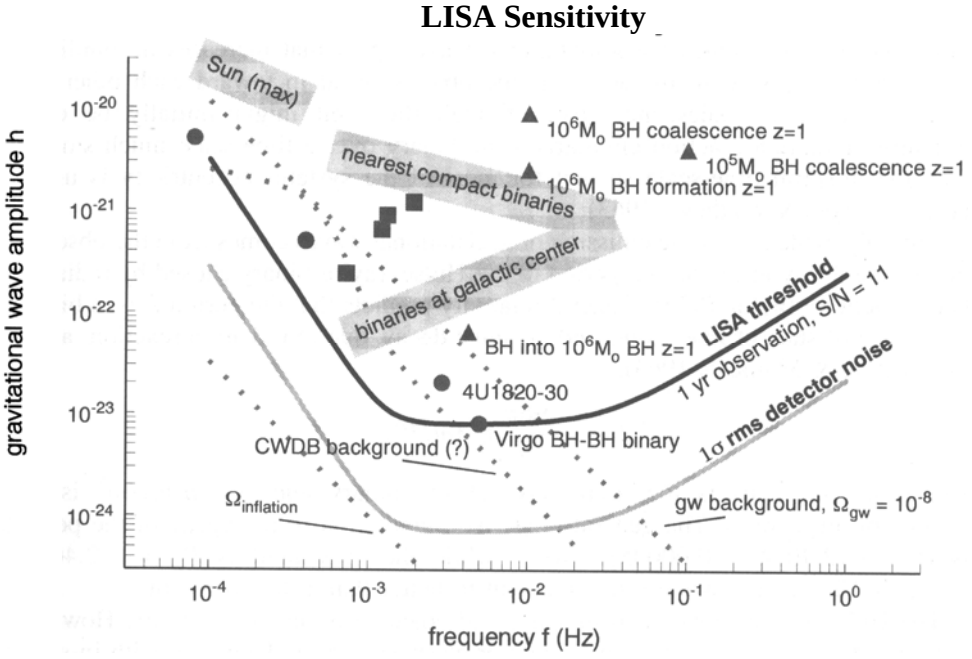

Figure 2. Sensitivity to bursts (SB) of LISA and effective signal strengths of several sources.

sources for the ground-based interferometers. Binaries consisting of objects both of which are super-massive black holes (SMBH) or one that is a SMBH and the other a stellar mass compact object are promising sources (Thorne & Braginsky 1976) for the LISA.

The net-work of ground-based interferometers will initially be able to survey a volume of $10^4 Mpc^3$ for in-spiralling compact binary stars and will enhance that volume by 3 orders of magnitude in about five years after they are built, thus increasing the number of potential events a thousand-fold. A space interferometer will be able to detect in-spiral and coalescence of super-massive black hole binaries wherever they occur in the Universe. Observation of gravitational radiation from such sources will make it possible to turn the esoteric theoretical studies of black holes into an observational science.

3. Gravitational waves from black holes

In this section we shall discuss astrophysical black hole sources of gravitational waves that are likely to be observed by interferometric detectors. These are in-spiral and merger waves from binaries and quasi-normal mode ringing produced when a black hole forms or is subject to external perturbation.

3.1 In-spiralling and merging compact binaries

Close binary systems composed of compact objects (such as black holes and neutron stars) will be an important source of gravitational waves for laser interferometric

detectors. The orbit of such a binary decays under the influence of gravitational radiation reaction, emitting a gravitational wave signal that increases in amplitude and "chirps" upward in frequency as the objects spiral in toward each other just before their final coalescence. Even though the orbit might initially be quite eccentric, radiation reaction circularises the binary over a timescale much smaller than the in-spiral timescale so that for most of the systems eccentricity is unimportant (Peters & Mathews 1963).

Our only evidence for the emission of gravitational waves comes from the observation of the decay in the orbital period of the Hulse-Taylor binary caused by radiation back-reaction (Taylor 1994). General relativity predicts that the period P of a binary consisting of stars of masses m_1 and m_2 must decay, due to radiation reaction, at the rate (Peters & Mathews 1963).

$$\dot{P} = -\frac{96\pi}{5}\,\eta(\pi m/P)^{5/3}, \tag{1}$$

where $m = m_1 + m_2$ is the total mass of the binary and $\eta = m_1 m_2/m^2$ is the symmetric mass ratio. The general relativistic prediction of the decay of the period is $\dot{P}_{\rm T} = -(2.40243\pm0.00005)\times10^{-12}$ while that measured is $\dot{P}_{\rm O} = -(2.408\pm 0.011)\times10^{-12}$. The two are in agreement to better than a few per cent.

The Hulse-Taylor binary will in-spiral and coalesce in about 10^8 years. However, there must be others, in our own Galaxy or elsewhere in the Universe, with in-spiral time-scale much smaller than that. Estimates based on the observed compact binary population in the Galaxy and theoretical modelling of the evolution of such systems give the rate of in-spiral to be about three per year within 200 Mpc of the Earth for neutron-star-neutron-star binaries, and within 1 Gpc for black-hole-black-hole binaries (Narayan, Piran & Shemi 1991; Phinney 1991; Zwart & Spreeuw 1996; Van den Heuvel & Lorimer 1996; Stairs *et al* 1997; Lipunov, Postnov & Prokhorov 1997). In-spiral signals occurring at these distances are strong enough to be detected by the net-work of ground-based interferometers in their advanced stage of operation.

Galaxy collisions, observed in high red-shift optical surveys, will necessarily involve the in-spiral and merger of the black holes at their centres. Order of magnitude estimates suggest that LISA should see SMBH-SMBH binary in-spiral at roughly once a year (Hils & Bender 1995; Sigurdsson & Rees 1996). The capture and in-spiral of compact bodies by a SMBH might be much more frequent, making observation of such events possibly quite routine (Haehnelt 1994; Rajagopal & Romani 1995; Vecchio 1997).

The dominant channel in which gravitational wave production takes place is at a frequency equal to twice the orbital frequency. Equation (1) can be integrated to get the evolution of the gravitational wave frequency $f(t)$:

$$f(t) = f_0[1 - (t - t_0)/\tau_0]^{-3/8}. \tag{2}$$

The quantity $\tau0$, called the *chirp time,* is the time left for the two bodies to in-spiral and coalesce starting from a time when the frequency was f_0:

$$\tau_0 = \frac{5}{256}\eta^{-1}m^{-5/3}(\pi f_0)^{-8/3}. \tag{3}$$

A binary source will become visible in a detector when the frequency of radiation it is emitting enters the detector band. Ground-based detectors will be able to track the

source from then on until the two stars merge; sources which ground-based interferometers will observe would merge within seconds or minutes after entering the detector band. However, in a space interferometer a binary might be observed when the time to coalesce (several to millions of years) is much larger than the duration of observation (one to several years).

The in-spiral phase terminates when the binary reaches its last stable orbit (LSO) at which time the two stars are roughly at a distance $R = 6m$ from each other. In terms of the frequency of gravitational radiation this is given by

$$f_{\rm lso} = (6^{3/2}\pi m)^{-1}, \tag{4}$$

which works out to be roughly 4300 Hz for $m = 1M_{\odot}$and 4.3 mHz for $m = 10^6\ M_{\odot}$. Binaries that reach LSO before reaching a detector's sensitivity band will be unobservable by that detector. For most ground-based interferometers the frequency where most of the signal power is extracted is around 50 Hz, which means the heaviest binaries they will observe is ~ 100 $M_{\odot}$. The LISA detector has good sensitivity down to about 10^4 Hz (see Fig. 2) and hence it will be able to detect super-massive black holes of total mass up to about $10^7\ M_{\odot}$

Gravitational waves carry energy and momentum from the system. By demanding that the energy lost in the form of gravitational radiation is precisely balanced by the decrease of energy in the system, one can derive a simple expression for the apparent luminosity of radiation $\mathcal{F}$, at great distances from the source, in terms of the variation of the gravitational wave amplitude (Schutz 1985):

$$\mathcal{F} = \frac{|\dot{h}|^2}{16\pi}. \tag{5}$$

The above relation can be used to make an order-of-magnitude estimate of the gravitational wave amplitude from a knowledge of the rate at which energy is emitted by a source in the form of gravitational waves. If a source at a distance r radiates away energy E, in a time T, predominantly at a frequency f, then writing $h = 2\pi fh$ and noting that $\mathcal{F} = E/(4\pi r^2 T)$, the amplitude of gravitational waves is

$$h \sim \sqrt{E/T}(\pi f r)^{-1}. \tag{6}$$

When the time development of a signal is known one can filter the detector output through a copy of the signal. This leads to an enhancement in the SNR, as compared to its narrowband value, by the square-root of the number of cycles the signal spends in the detector band. A signal lasting for a time T around a frequency f would produce $\eta \simeq fT$ cycles. Using this we can eliminate T from equation (6). Defining an effective amplitude $h_{\rm eff} \equiv \sqrt{n} \times h$ we have

$$h_{\rm eff} = \sqrt{E/f}/(\pi r). \tag{7}$$

Let us now apply this to the in-spiral case. The energy emitted in the process of in-spiral is equal to the binding energy: $|E| = \eta m^2/2R$ The distance between the two stars is related to gravitational wave frequency via $R = (M/\pi^2 f^2)^{1/3}$ which gives $|E| = \eta m^{5/3}(\pi f)^{2/3}/2.$ sing this in the expression for the effective amplitude we get,

$$h_{\rm eff} \simeq \frac{\eta^{1/2} m^{5/6} f^{-1/6}}{\pi^{2/3} r}, \tag{8}$$

which gives,

$$h_{\rm eff} \simeq 3\times10^{-21}\left(\frac{30\,{\rm Mpc}}{r}\right)\left(\frac{\eta}{0.25}\right)^{1/2}\left(\frac{m}{2.8\,M_\odot}\right)^{5/6}\left(\frac{1\,{\rm kHz}}{f}\right)^{1/6}$$
$$\simeq 4\times10^{-17}\left(\frac{1\,{\rm Gpc}}{r}\right)\left(\frac{\eta}{0.25}\right)^{1/2}\left(\frac{m}{10^6\,M_\odot}\right)^{5/6}\left(\frac{1\,{\rm mHz}}{f}\right)^{1/6}. \qquad (9)$$

The first of the equalities shows the amplitude of gravitational waves from stellar mass compact binaries at a distance of 30Mpc. Such sources fall within the sensitivity and frequency range of ground-based antennas. The second equality shows the amplitude of waves from the in-spiral of binaries consisting of super-massive black holes or a super-massive black hole and a stellar mass compact object. These will be observable by the LISA detector.

3.2 *Quasi normal modes of a black hole*

Vishveshwara (1970) demonstrated for the first time that gravitational waves scattered off a black hole will have a characteristic wave form when the incident wave has frequencies beyond a certain value depending on the size of the black hole. It was soon realised that perturbed black holes have quasi-normal modes of vibration and in the process emit gravitational radiation whose amplitude, frequency and damping time are characteristic of the hole's mass and angular momentum (Press 1971).

We can estimate the amplitude of gravitational waves emitted when a black hole forms at a distance r as a result of the coalescence of compact objects in a binary, in the following way. The effective amplitude is given by the formula (8) which involves the energy E put into gravitational waves and the frequency f at which the waves come off. By dimensional arguments E is proportional to the total mass m of the resulting black hole. The efficiency at which the energy is converted into radiation depends on the symmetric mass ratio η of the merging objects. One does not know what is the fraction of the total mass emitted nor the exact dependence on η. Flanagan & Hughes (1998) argue that $E \sim 0.03(4\eta)^2 m$. The frequency f is inversely proportional to m; indeed, for Schwarzschild black holes $f = (2\pi m)^{-1}$. Thus, the formula for the effective amplitude takes the form

$$h_{\rm eff} \sim \frac{4\alpha\eta m}{\pi r}, \qquad (10)$$

where α is a number that depends on the (dimensionless) angular momentum a of the hole and takes values between 0.7 (for $a = 0$, Schwarzschild black hole) and 0.4 (for a=1, maximally spinning Kerr black hole). Thus,

$$h_{\rm eff} \sim 10^{-21}\left(\frac{\eta}{0.25}\right)\left(\frac{m}{20\,M_\odot}\right)\left(\frac{r}{200\,{\rm Mpc}}\right)^{-1},$$
$$\sim 10^{-21}\left(\frac{\eta}{10^{-4}}\right)\left(\frac{m}{10^6\,M_\odot}\right)\left(\frac{r}{4\,{\rm Gpc}}\right)^{-1},$$
$$\sim 10^{-16}\left(\frac{\eta}{0.25}\right)\left(\frac{m}{2\times10^6\,M_\odot}\right)\left(\frac{r}{2\,{\rm Gpc}}\right)^{-1}. \qquad (11)$$

Here we have quoted three cases:

- Two $10M_{\odot}$black holes in-spiralling and merging to form another black hole. In this case the waves come off at a frequency of around 1.5 kHz. The initial ground-based network of detectors might be able to pick these waves up by matched filtering, especially when an in-spiral event precedes the ring-down.
- A $100M_{\odot}$ black hole plunging into a 10^6 $M_{\odot}$ hole at a distance of 4 Gpc giving out radiation at a frequency of about 30 mHz. Such an event produces amplitude just about enough to be detected by LISA.
- Two super-massive black holes spiral-in and merge to produce a fantastic amplitude of 10^{-16}, way above the LISA background noise, even at a distance of 2 Gpc. In this case the waves would come off at about 15 mHz. These events will be loud and clear in the LISA detector. It will not only be possible to detect these events but also accurately measure the black hole mass and angular momentum and map the spacetime around the hole.

4. Gravitational astronomy

Black hole sources of gravitational waves have a lot of very interesting physics and offer an opportunity to test some of the predictions of general relativity in the strongly nonlinear regime, such as the tails of gravitational waves, spin-orbit coupling induced precession, nonlinear tails, hereditary effects, etc (Blanchet & Sathyaprakash 1994; Blanchet & Sathyaprakash 1995; Thorne 1995; Schutz 1997; Sathyaprakash & Schutz 1998). They are also good test beds to constrain other theories of gravity. Gravitational waves emitted either during the in-spiral and merger of rotating super-massive black holes or when a stellar mass compact object falls into a SMBH, can be used to map the structure of spacetime and test uniqueness theorems on rotating black holes (Thorne 1995; Schutz 1997). LISA will be able to see the formation of massive black holes at cosmological distances by detecting the waves emitted in the process (Thorne 1995; Schutz 1997). These are but a very modest list of physics that will be borne out of observation of gravitational waves from these sources.

Association of an in-spiral event with an electromagnetic event, such as the observation of a gamma or X-ray burst, would help to deduce the speed of gravitational waves to a phenomenal accuracy. (Even a day's delay in the arrival times of gravitational and electromagnetic radiation from a source at a distance of a million light years would determine the relative speeds to better than one part in 10^8). This will require a good timing accuracy to determine the direction to the source and to signal astronomers to guide their telescopes in that direction for electromagnetic observation.

A network of detectors will be able to determine the polarisation of the waves. While Einstein's general relativity predicts only two independent polarisations, there are other theories of gravitation in which more states of polarisation, and even dipolar waves, are predicted. Therefore, an unambiguous determination of the polarisation of the waves will be of fundamental importance. Moreover, astronomical observations of binaries cannot yield the total mass but only the combination m sin i, where i is the inclination of the binary's orbit to the line of sight. However, measurement of polarisation can determine the angle i since the polarisation state depends on the binary's inclination with the line of sight.

The post-Newtonian expansion of the in-spiral wave form has unfolded a rich cache of nonlinear physics that governs the dynamics of a binary. There are many interesting effects such as the tails of gravitational waves, spin-spin interaction, precessional effects, nonlinear tails, tails-of-tails, etc., which can all be deciphered from a strong in-spiral signal. For instance, the effect of nonlinear tails on the phasing of gravitational waves can be detected in the in-spiral waves emitted by black hole black hole binaries (Blanchet & Sathyaprakash 1994; Blanchet & Sathyaprakash 1995) One can also set a limit on the mass of the graviton by observing the in-spiral waves. If graviton has a mass then it would alter the phasing of the waves and hence by tracking the in-spiral waves one can bound gravtion's mass to 2.5×10^{-22}eV using ground-based detectors and 2.5×10^{-26}eV using space-based detectors, in future observations of in-spiral events (Will 1998).

In-spiral waves are standard candles in the sense that by measuring their luminosity we can infer the distance to source (Schutz 1986). In the quadrupole approximation, the wave's amplitude does not depend separately on the individual masses of the component stars but only on a certain combination of the masses called the *chirp mass*, $\mathcal{M} = \eta^{3/5} m$, and the distance r to the source. If we can measure the amplitude and the chirp mass independently, then we can infer the distance to the source directly. In-spiral waves are detected by using matched filtering the detector output through a large number of copies of the expected wave form corresponding to different values of the source's parameters. That copy which matches closest to the signal (when a signal is present) gives the largest signal-to-noise ratio. This enables an accurate estimation of the chirp mass. The value of the signal-to-noise ratio enables us to infer the signal's amplitude as well. Thus, in-spiral waves serve as standard candles. This will allow us to determine the Hubble constant very accurately by identifying the host galaxy and measuring its red-shift. It should be noted that the in-spiral wave's phase is quite sensitive to post-Newtonian corrections and when they are included the mass-degeneracy is lifted. It is, therefore, possible to determine both masses of the binary by using post-Newtonian search templates.

The origin of super-massive black holes is not clear but they must be accreting ordinary stars, compact stars and black holes, in their vicinity, thereby increasing their mass. Accretion of ordinary stars is not likely to produce coherent emission of gravitational waves as the star would fragment much before reaching the horizon. However, spiralling-in of compact stars or black holes would definitely be producing coherent radiation. These events will be more frequent than merger of massive black holes and will be very interesting phenomena from fundamental physics point of view. We do not yet fully understand trajectories of test masses around Kerr black holes; observation of such events will greatly help in testing strong field predictions of general relativity and understanding black hole spacetimes including the uniqueness of the Kerr solution.

High red-shift galaxy catalogues show interactions between galaxies which is clearly indicative of collisions, mergers, etc. Galaxy collisions must be accompanied by the in-spiral and merger of the associated black holes. Space-based detectors will be sensitive to gravitational waves from such systems even at a red-shift $z = 1$ (roughly 3 Gpc) and would obtain a SNR $\sim 10^4$.

Since in-spiral signals are standard candles, observations of massive black hole coalescences at cosmological distances $z \sim 1$ by space-based detectors can facilitate an accurate determination of the distance to the source. Space-based detectors observe

a massive binary in-spiral for a whole year and they have, thus, the baseline of the Earth's orbit around the Sun to triangulate the source on the sky. They can do this to an accuracy of a degree at high SNR. At a distance of 3 Gpc this is about l0Mpc, a scale over which no more than one virialised galaxy cluster can be found. Thus, an optical identification of the host galaxy cluster and its red-shift, would enable the measurement of the deceleration parameter qo and hence the density parameter Ω of the Universe. A single source at $z \gtrsim 1$ is enough to measure both H_0 and Ω to an accuracy of better than 1%. Thus, space-based detectors can potentially contribute quite a lot to further our understanding of fundamental science and cosmology.

5. Conclusions and future directions

Doing all this exciting physics is possible if only we can measure the source parameters accurately and without any systematic bias. Gravitational wave measurements will be made by matched filtering the detector output in a multi-dimensional parameter space of signals, using search templates that are essentially copies of the expected signals. However, presently we only know the in-spiral wave form approximately (computed using post-Newtonian expansion of Einstein's equations (Damour & Deruelle 1981; Damour 1982; Damour 1983; Blanchet *et al.* 1995; Blanchet, Damour & Iyer 1995; Will & Wiseman 1996; Blanchet *et al.* 1996; Blanchet 1996), and this situation is not likely to change significantly in the near future. Moreover, the equations are solved in the quasi-static or adiabatic approximation. This approximation assumes that the radiation reaction time-scale is much larger than the orbital time-scale. The reaction forces on the bodies due to emission of gravitational waves are computed by averaging the luminosity over an orbital period. The adiabatic approximation will breakdown when the two bodies are very close ($r \sim 6$m) and understanding the dynamics of orbits close to merger and of general eccentric orbits will require a knowledge of the local radiation reaction force (instead of the averaging mentioned above).

Damour, Iyer and Sathyaprakash (1998) have proposed an improved version of the post-Newtonian wave forms, called *P*-approximants. These improved models help not only in the detection of these signals but also facilitate in an accurate determination of the signal parameters. However, detection and measurement of radiation from realistic black hole binaries with large spins and high eccentricity will require very accurate template wave forms. One of the missing links in this regard is the unavailability of a local expression for radiation reaction force. There is now a world-wide effort to solve some of these problems.

References

Abramovici, A. *et al.* 1992, *Science*, **256,** 325.

Bender, P. *et al.* LISA: *Pre-Phase A Report,* MPQ 208 (Max-Planck-Institut für Quantenoptik, Garching, Germany). (Also see the Second Edition, July 1998).

Bender, R., Kormendy, J., Dehnen, W. 1996, *Astrophys. J. Lett.*, **464,** L123.

Blanchet, L., Iyer, B. R., Will, C. M., Wiseman, A. G. 1996, *Class. Quantum. Gr.*, **13**, 575.

Blanchet, L., 1996, *Phys. Rev.*, **D54,** 1417.

Blanchet, L., Sathyaprakash, B. S. 1995, *Phys. Rev. Lett.,* **74**, 1067.

Blanchet, L., Sathyaprakash, B. S. 1994, *Class. Quant. Grav.*, **11**, 2807.

Blanchet, L., Damour, T., Iyer, B. R., Will, C. M., Wiseman, A. G. 1995, *Phys. Rev. Lett*., **74**, 3515.
Blanchet, L., Damour, T., Iyer, B. R. 1995, *Phys. Rev*., **D51**, 5360.
Caron, B. *et al.* 1997, *Class. Quantum Grav*., **14**, 1461.
Damour, T., Iyer, B. R., Sathyaprakash, B. S. 1998, *Phys. Rev*., **D57**, 885.
Damour, T. 1983, in *Gravitational Radiation*, (ed.) N. Deruelle and T. Piran, pp 59–144 (North-Holland, Amsterdam).
Damour, T. 1982, *CR. Acad. Sci*., *Paris*, **294**, (II) 1355.
Damour, T., Deruelle, N. 1981, *Phys. Lett*., **87A**, 81.
Eckart, A., Genzel, R. 1996, *Nature*, **383**, 415.
Flanagan, É. É., Hughes, S. 1998, *Phys. Rev. D*, **57**, 4535.
Fukushige, T., Ebisuzaki, T., Makino, J. *Astrophys. J.*, **396**, L61.
Gaskell, C. M. 1996, *Astrophys. J.*, **646**, 107.
Haehnelt, M. G. 1994, *Mon. Not. R. Astron. Soc.*, **269**, 199.
Hils, D., Bender, P. L. 1995, *Astrophys. J. Lett*., **445**, L7.
Kormendy, J., Richstone, D. 1995, *Ann. Rev. Astron. & Astrophys.*, **33**, 581.
Lipunov, V. M., Postnov, K. A., Prokhorov, M. E. 1997, *Mon. Not. R. Astron. Soc*., **288**, 245.
Lück, H. *et al.,* 1997, *Class. Quantum Grav*., **14,** 1471.
Narayan, R., Piran, T., Shemi, A. 1991, *Astrophys. J*., **379,** L17.
Peters, P. C, Mathews, J. 1963, *Phys. Rev.,* **131,** 435.
Phinney, E. S. 1991, *Astrophys. J*., **380,** L17.
Press, W. H. 1971, *Astrophys. J. Lett.,* **170,** L105.
Rajagopal, M., Romani, R. W. 1995, *Astrophys. J.*, **446,** 543.
Rees, M. J. 1997, *Class. Quantum Grav*., **14,** 1411.
Roos, N., Kastra, J. S., Hummel, C. A. 1993, *Astrophys. J.*, **409,** 130.
Sathyaprakash, B. S., Schutz, B. F. 1998, submitted to Living Reviews in Relativity.
Schutz, B. F. 1997, gr-qc/9710080, gr-qc/9710079.
Schutz, B. F. 1985, *A First course in general relativity*, (Cambridge Univ. Press, Cambridge.)
Schutz, B. F., 1986, *Nature*, **323**, 310.
Sigurdsson, S., Rees, M. J. 1996, *Mon. Not. R. Astron. Soc*., **284**, 318.
Sillanpää, A., Haarala, S., Valtonen, M. J., Sundelius, B., and Byrd, G. G. 1998, *Astrophys. J*., **325**, 628.
Stairs, I. H. *et al.,* 1997, astro-ph/9712296.
Taylor, J. H. 1994, *Rev. Mod. Phys*., **66**, 711.
Thorne, K. S. 1995, in *Proceedings of Snowmass 1994 Summer Study on Particle and Nuclear Astrophysics and Cosmology,* (eds) E. W. Kolb and R. D. Peccei, (World Scientific, Singapore) pp. 160–184.
Thorne, K. S., Braginsky, V. B. 1976, *Astrophys. J*., **204**, L1.
Tsubono, K. 1995, in *Gravitational Wave Experiments* (World Scientific, Singapore), p. 112.
Valtoja, L., Valtonen, M. J., Byrd, G. G. 1989, *Astrophys. J*., **343**, 47.
Van den Heuvel, E. P. J., Lorimer, D. R. 1996, *Mon. Not. R. Astron. Soc.*, **283**, L67.
Van der Marel, R. P., de Zeeuw, T., Rix, H. W., Quinlan, G. D. 1996, *Nature*, **385**, 610.
Vecchio, A. 1997, *Class. Quantum. Gr.*, **14**, 1431.
Vishveshwara, C. V. 1970, *Nature*, **227**, 936.
Will, C. M. 1998, *Phys. Rev. D*, **57**, 2061.
Will, C. M., Wiseman, A. G., 1996, *Phys. Rev*., **D54**, 4813.
Zwart, S. F. P., Spreeuw, H. N. 1996, *Astron. Astrophys*., **312**, 670.

J. Astrophys. Astr. (1999) **20**, 221–232

Gravitational Collapse, Black Holes and Naked Singularities

T. P. Singh, *Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005, India*
email: tpsingh@tifr.res.in

Abstract. This article gives an elementary review of gravitational collapse and the cosmic censorship hypothesis. Known models of collapse resulting in the formation of black holes and naked singularities are summarized. These models, when taken together, suggest that the censorship hypothesis may not hold in classical general relativity. The nature of the quantum processes that take place near a naked singularity, and their possible implication for observations, is briefly discussed.

Key words. Cosmic censorship—black holes—naked singularities.

1. Introduction

After a star has exhausted its nuclear fuel, it can no longer remain in equilibrium and must ultimately undergo gravitational collapse. The star will end as a white dwarf if the mass of the collapsing core is less than the famous Chandrashekhar limit of 1.4 solar masses. It will end as a neutron star if the core has a mass greater than the Chandrashekhar limit and less than about 35 times the mass of the sun. It is often believed that a core heavier than about 5 solar masses will end, not as a white dwarf or as a neutron star, but as a black hole. However, this belief that a black hole will necessarily form is not based on any firm theoretical evidence. An alternate possibility allowed by the theory is that a naked singularity can form, and the purpose of the present article is to review our current understanding of gravitational collapse and the formation of black holes and naked singularities.

A black hole has been appropriately described by Chandrashekhar as the most beautiful macroscopic object known to man. Only a few parameters suffice to describe the most general black hole solution, and these objects have remarkable thermodynamic properties. Further, excellent observational evidence for their existence has developed over the years (Rees 1998). Thus, there can be no doubt about the reality of black holes, and the gravitational collapse of very many sufficiently massive stars must end in the formation of a black hole.

However, the following question is still very much open. If the collapsing core is heavy enough to not end as a neutron star, does this guarantee that a black hole will necessarily form? The answer to this question has to come from the general theory of relativity, and unfortunately this remains an unsolved problem.

What we do know from general relativity about gravitational collapse is broadly contained in the celebrated singularity theorems of Geroch, Hawking and Penrose. It has been shown that under fairly general conditions, a sufficiently massive collapsing object will undergo continual gravitational collapse, resulting in the formation of a

gravitational singularity. The energy density of the collapsing matter, as well as the curvature of spacetime, are expected to diverge at this singularity.

Is such a singularity necessarily surrounded by an invisible region of spacetime, i.e. has a black hole formed? The singularity theorems do not imply so. The singularity may or may not be visible to a far away observer. If the singularity is invisible to a far away observer, we say the star has ended as a black hole. If it is visible, we say the star has ended as a naked singularity. We need to have a better understanding of general relativity in order to decide whether collapse always ends in a black hole or whether naked singularities can sometimes form.

Given this situation, Penrose was led to ask (Penrose 1969) whether there might exist a cosmic censor who forbids the existence of naked singularities, 'clothing each one of them with a horizon'? Later, this led to the cosmic censorship hypothesis, which in broad physical terms states that the generic singularities arising in the gravitational collapse of physically reasonable matter are not naked. Till today, this hypothesis remains unproven in general relativity, neither is it clear that the hypothesis holds true in the theory. What is of course true is that the hypothesis forms the working basis for all of black hole physics and astrophysics. If cosmic censorship were to not hold, then some of the very massive stars will end as black holes, while others could end as naked singularities. As we will argue in section 3, these two kinds of objects have very different observational properties.

There are various very important reasons for investigating whether or not cosmic censorship holds in classical general relativity. As we have mentioned above, the hypothesis is vital for black hole astrophysics. Unfortunately this fact is rarely appreciated by the astrophysics community. The hypothesis is also necessary for the proof of the black hole area theorem. It is not clear what the status of this theorem will be if the hypothesis were to not hold. If naked singularities do occur in classical relativity, they represent a breakdown of predictability, because one could not predict the evolution of spacetime beyond a naked singularity. Such singularities would then provide pointers towards a modification of classical general relativity, so that a suitable form of predictability is restored in the modified theory. Further, naked singularities might be observable in nature, if they are allowed by general relativity. Undoubtedly then, it is important to find out if the censorship hypothesis is valid.

We wish to make two further remarks. Firstly, while a great deal is known about the properties of stationary black holes, we know very little about the process of black hole *formation.* In fact we know as little about the formation of black holes as we do about the formation of naked singularities. Secondly, it has sometimes been remarked that a theory of quantum gravity is likely to get rid of the singularities of classical general relativity, irrespective of whether these singularities are naked or covered. Why then does it eventually matter whether or not cosmic censorship holds? The answer to this legitimate objection is the following. A quantum gravity theory is expected to smear out a classical singularity and replace it by a region of very high, albeit finite, curvature. If the classical singularity is hidden behind a horizon (i.e. is a black hole), this quantum smeared region remains invisible to an external observer. However, if the classical singularity is naked, the smeared region of very high curvature will be visible to far away observers, and the physical processes taking place near this smeared region will be significantly different from those taking place outside the horizon of ordinary astrophysical black holes. Hence, from such an experimental standpoint, quantum gravity has little bearing on the question of cosmic censorship.

To put it differently, quantum gravity is not expected to restore the event horizon, if the horizon is absent in the classical theory.

Since a theorem proving or disproving the hypothesis has not been found, attention has shifted to studying model examples of gravitational collapse, to find out whether the collapse ends in a black hole or a naked singularity. While specialised examples such as have been studied are nowhere near a general proof, they are really all that we have to go by, as of now. However, there does seem to be an underlying pattern in the results that have been found in these examples, which gives some indication of the general picture. It is interesting that all models studied to date admit both black hole and naked singularity solutions, depending on the choice of initial data. In the next section, we give a summary of what has been learnt from these examples and what they probably tell us about cosmic censorship. In the third section we will address the question of whether naked singularities might occur in nature, and if so, what they would look like to an observer.

The reader is also invited to study a few other excellent reviews (Clarke 1993; Joshi 1993; Joshi 1997; Wald 1997; Penrose 1998) on the subject of cosmic censorship, which have appeared in recent years. Some of these reviews emphasize aspects other than those presented here, and some arrive at conclusions on cosmic censorship other than those given here. I would also like to draw attention to an earlier review of mine on this topic, which is somewhat more detailed, though a little dated (Singh 1996). Also, for a detailed bibliography the reader is requested is look up this earlier review; the references in the present article are not exhaustive, and largely confined to the more recent papers.

2. Theoretical evidence for the formation of black holes and naked singularities

We consider the gravitational collapse of physically reasonable classical matter, where by 'physically reasonable' is meant that the matter satisfies one or more of the energy conditions (weak energy condition, strong energy condition and the dominant energy condition). Also, most of the collapse studies that have been carried out so far deal with spherical collapse—even this simplest of systems is poorly understood, in so far as cosmic censorship is concerned. It is of course true that spherical collapse, if allowed to proceed to completion, results in a Schwarzschild black hole. However, a spherical collapsing system can also admit timelike or null singularities which can be naked. From the point of view of an observer falling with the star this can happen if a singularity forms inside the star (say at its center) before the boundary of the star enters its Schwarzschild radius. Such a singularity can be naked.

Since the exact solution of Einstein equations for spherical collapse with a general form of matter is not known, collapse of matter with various equations of state has been studied. In the following pages, we review some of these results.

2.1 *Spherical dust collapse*

Historically, the earliest model of gravitational collapse is due to Oppenheimer and Snyder. They showed that the collapse of a homogeneous dust sphere results in the formation of a black hole (by dust is meant an idealized perfect fluid for which the

pressure is zero). It was thought that the formation of the black hole will not be affected even if the specialized assumptions of this model (homogeneity, sphericity, dust equation of state) are relaxed. However, we now know that this is not so.

When the assumption of homogeneity is relaxed, there is an exact solution of Einstein equations-the Datt-Tolman-Bondi solution, which describes collapse of a dust sphere with non-uniform initial density. Two kinds of singularities can result–shell crossing and shell focusing. While the former have a Newtonian analog, at least some of the latter appear to be of purely relativistic origin. It has been shown by various authors (for detailed references see Singh (1996); also see Dwivedi & Joshi 1997; Herrera *et al.* 1997) that the shell-focusing singularities can be of both the black hole and naked type, depending on the initial conditions.

As an illustration, we mention the following interesting case. Consider the collapse of a dust sphere, starting from rest, and having an initial density profile near the center given by

$$\rho(R) = \rho_0 + \rho_1 R + \tfrac{1}{2}\rho_2 R^2 + \tfrac{1}{6}\rho_3 R^3 + \cdots .$$

It turns out that the singularity is naked if $\rho_1$is less than zero, and also if $\rho_1$is equal to zero and ρ_2 is less than zero. If both $\rho_1$and ρ_2 are zero and ρ_3 is negative, then we define a dimensionless quantity $\xi = |\rho_3|/\rho_0^{5/2}$. The singularity is naked if $\xi \geq 25.48$ and covered if is less than this number. If ρ_1, ρ_2 and ρ_3 are all zero, the singularity is covered, the Oppenheimer-Snyder collapse being a special case of this.

It is known that the collapse of null dust (directed radiation), described by the Vaidya spacetime, also gives rise to both black hole and naked singularity solutions, depending on the rate of infall (for references see Singh (1996)).

2.2 *Spherical collapse of fluids with pressure*

Exact solutions of Einstein equations describing collapse of fluids are rare. Hence little is known about the end state of collapse in these systems. To some degree, numerical methods have been used to integrate Einstein equations and study light propagation. In comoving coordinates, the energy-momentum tensor is diagonal and its components are the energy density, the radial pressure and the tangential pressure. For a perfect fluid, the two pressures are identical.

A significant development was the work of Ori & Piran (1990) who investigated the self-similar gravitational collapse of a perfect fluid with an equation of state $p = k\rho$. It is readily shown that the collapse leads to the formation of a curvature singularity. The assumption of self-similarity reduces Einstein equations to ordinary differential equations which are solved numerically, along with the equations for radial and non-radial null geodesics. It is then shown that for every value of k (in the range investigated: $0 \leq k \leq 0.4$) there are solutions with a naked singularity, as well as black hole solutions.

An analytical treatment for this problem was developed by Joshi & Dwivedi (1992). After deriving the Einstein equations for the collapsing self-similar perfect fluid they reduce the geodesic equation, in the neighborhood of the singularity, to an algebraic equation. The free parameters in this algebraic equation are in principle determined by the initial data. The singularity will be naked for those values of the parameters for which this equation admits positive real roots. Since this is an

algebraic equation, it will necessarily have positive roots for some of the values of the parameters, and for the initial data corresponding to such values of the parameters the singularity is naked.

Lifshitz & Khalatnikov (1961) (and Podurets (1966)) worked out the form of the solution near the singularity for the equation of state of radiation. This work is a precursor to the Belinskii-Lifshitz-Khalatnikov (BLK) series solutions near singularities. Following the method of Podurets, we investigated the nature of the non-central shell-focusing singularity which can form during the collapse of a fluid (Cooperstock *et al.* 1997). It is easily shown that such a singularity is covered so long as the radial pressure is positive. By considering the case of a perfect fluid, we showed that negative pressure allows for a naked singularity if the ratio of the pressure to the density is $\leq -1/3$; and the singularity is covered if this ratio exceeds $-1/3$. We note that the weak energy condition allows for pressure to be negative, although it is questionable whether negative pressures could develop during the final stages of realistic stellar collapse. The BLK series solutions offer a promising avenue for investigating cosmic censorship, which deserves to be pursued further.

On physical grounds, imperfect fluids are more realistic than perfect ones; very little is known about their collapse properties though. An interesting paper is the one by Szekeres & Iyer (1993), who do not start by assuming an equation of state. Instead they assume the metric components to have a certain power-law form, and also assume that collapse of physically reasonable fluids can be described by such metrics. The singularities resulting in the evolution are analysed, with attention being concentrated on shell-focusing singularities at $r > 0$. They find that although naked singularities can occur, this requires that the radial or tangential pressure must either be negative or equal in magnitude to the density.

Another model which has recently attracted some attention is the collapse of a fluid having only tangential pressure (Magli 1997; Singh & Witten 1997; Barve, Singh & Witten 1999). The analysis of the Einstein equations is considerably simpler than the case in which radial pressure is also present. Hence this is a useful system for studying the stability of dust naked singularities against the introduction of pressure. It has been found that while certain equations of state admit only black hole type singularities, other state equations admit naked singularities as well.

We conclude this brief discussion of fluids by commenting on the issue of whether fluids are a reasonable form of matter in so far as cosmic censorship studies are concerned. An objection sometimes raised against fluids is that they form singularities even in Minkowski spacetime, and hence the naked singularities that have been found do not have anything to do with general relativity. However, while some of the singularities, like the shell-crossings, and possibly the weak shell-focusings, have Minkowskian analogs, it is by no means clear that all the singularities (for instance the strong curvature shell-focusings) have counterparts in flat spacetime evolution. At the very least, this has to be investigated further, so as to get a precise distinction of the singularities that have Minkowskian analogs, from those that are of purely relativistic origin.

It is however more useful to examine cosmic censorship keeping the astrophysical context in mind, and here we know that a fluid description of stellar matter is physically quite appropriate. Thus, if a naked singularity were to result in the collapse of a real star made of fluid matter, we would be compelled to seriously pursue its observational consequences.

2.3 *Collapse of a massless scalar field*

In a series of papers, Christodoulou (1986; 1987a; 1987b; 1993; 1994) has pioneered analytical studies of the spherical collapse of a self-gravitating massless scalar field. He established the global existence and uniqueness of solutions for the collapsing field, and also gave sufficient conditions for the formation of a trapped surface. For a self-similar scalar collapse he showed that there are initial conditions which result in the formation of naked singularities.

Christodoulou was also interested in the question of the mass of the black hole which might form during the collapse of the scalar wave-packet. Given a one parameter family $S[p]$ of solutions labeled by the parameter p which controls the strength of interaction, it was expected that as p is varied, there would be solutions with $p \rightarrow P_{weak}$ in which the collapsing wave-packet disperses again, and solutions with $P \rightarrow P_{\rm strong}$ which have black hole formation. For a given family there was expected to be a critical value $P = P_*$ for which the first black hole appears as p varies from the weak to the strong range. Do the smallest mass black holes have finite or infinitesimal mass? This issue would be of interest for censorship, since an infinitesimal mass would mean one could probe arbitrarily close to the singularity.

This problem was studied by Choptuik (1993) numerically and some remarkable results were found. He confirmed that the family $S[p]$ has dispersive solutions as well as those forming black holes, and a transition takes place from one class to the other at a critical $P = P_*$. The smallest black holes have infinitesimal mass. Near the critical region, the mass M_{bh} of the black hole scales as $M_{bh} \approx (P–P_*)^{\gamma}$ where γ is a universal constant (i.e. same for all families) having a value of about 0.37. The near critical evolution can be described by a universal solution of the field equations which also has a periodicity property called echoing, or discrete self-similarity. That is, it remains unchanged under a rescaling $(r,\ t) \rightarrow (e^{-n\Delta}r,\ e^{-n\Delta}t)$ of spacetime coordinates. n is an integer, and Δ is about 3.4. Subsequently, these results have been confirmed by others.

At the critical solution, the mass of the forming black hole goes to zero, *as* $p \rightarrow p_*$ from the right. This critical solution is a naked singularity. However, since the naked singularity is realised for a specific solution in the one parameter family, it is a subset of measure zero. As regards cosmic censorship, the more significant features are the black holes of arbitrarily small mass, and hence arbitrarily high curvature that is visible to a far away observer. It is more physical to think of censorship in terms of whether or not regions of unbounded high curvature are generically visible, and not just whether singularities are visible. Looked at in this way, scalar collapse provides a serious counterexample to censorship.

Similar critical behaviour has also been found in numerical studies of collapse with other forms of matter. Axisymmetric collapse of gravitational waves was shown to have a g of about 0.36, and $\Delta \approx 0.6$. For spherical collapse of radiation (perfect fluid with equation of state $p = \rho/3$) the critical solution has continuous self-similarity, and g of about 0.36. However it has become clear now that the critical exponent γ is not independent of the choice of matter. A study of collapse for a perfect fluid with an equation of state $p = k\rho$ shows that γ depends on k. For a given form of matter, there appears to be a unique g, but the value changes as the form of T_{ik} is changed.

An important issue regarding the solutions with $p > p_*$ is the following. These solutions are identified as black holes because of the presence of an apparent horizon.

However, current numerical studies do not probe the singularity itself, and one cannot for now rule out the possibility that the solutions with $p > p_*$ fall into two classes: (**a**) those which have a Cauchy horizon lying outside the apparent horizon, and hence are naked singularities, and (b) those which are black holes. There is actually some evidence for this in the work of Brady (1995), and this aspect needs to be investigated further.

A much more detailed discussion of scalar collapse and critical phenomena can be found in other recent reviews (Choptuik 1998; Gundlach 1997).

2.4 *Spherical collapse with general form of matter*

There is a certain degree of similarity in the collapse behaviour of dust, fluids and scalar fields—in all cases some of the initial data lead to black holes, while other data lead to naked singularities. This would suggest an underlying pattern which is probably characterized, not by the form of matter, but by some invariants of the gravitational field. Hence investigations of collapse which put no restriction on T_{ik} apart from an energy condition should prove useful.

An interesting attempt in this direction was made by Dwivedi & Joshi (1994). They assumed a general T^{i}_{k} obeying the weak energy condition, and also that the collapsing matter forms a curvature singularity. As we noted earlier, in the comoving coordinate system, matter is described by its energy density and the radial and tangential pressures. Along with these three functions, three functions describing the metric enter a set of five Einstein equations, which are coupled with an equation of state in order to close the system. The geodesic equation for radial null geodesies is written in the limit of approach to the singularity, and it is shown that the occurrence of a visible singularity is equivalent to the occurrence of a positive real root for the geodesic equation, suitably written. Since this equation depends on free initial data, it follows that for a subset of the initial data there will be positive real roots and the singularity will be visible.

2.5 *Null geodesic expansion and cosmic censorship*

A line of investigation which may prove useful for studying collapse of a general form of matter is to examine the evolution of the expansion θ of a congruence of outgoing null geodesics. Some preliminary work has recently been done (Singh, 1998). Consider first the case of the collapsing spherical dust cloud. If a point on the cloud ends up as a covered singularity, then q at this point evolves to a negative value, as expected, starting from its initial positive value. A naked singularity forms precisely in those cases for which the initially positive θ continues to remain positive all the way until singularity formation. We have given an argument suggesting that this property of θ (i.e. its remaining positive throughout the evolution for some initial data) is stable against small changes in the equation of state.

Hence, if dust admits a naked singularity for some initial data, a naked singularity will form also in the collapse of a fluid for which the ratio of pressure to density is small but nonzero, provided one starts from the same initial data. It may be possible to generalise these results, by using the Raychaudhri equation to predict which initial conditions lead to a black hole, and which ones to naked singularities. This is at present under investigation.

2.6 *Non-spherical gravitational collapse*

Amongst the very few studies of non-spherical collapse that have been carried out so far is the numerical work of Shapiro and Teukolsky on oblate and prolate collisionless spheroids. Since there really have been no recent developments in this direction, I refer the reader to the discussion of their work in section 3 of my earlier review (Singh 1996), and to the review by Wald (1998).

A recent work on non-spherical perturbations of spherical collapse deserves mention. Iguchi *et al.* have shown that the naked singularities arising in spherical dust collapse are marginally stable against odd-parity non-spherical gravitational wave perturbations (Iguchi, Nakao & Harada 1997).

Unlike the spherical case, very little is known about gravitational collapse and cosmic censorship for non-spherical systems.

2.7 *Properties of naked singularities*

There are now sufficiently many known examples of naked singularities for one to enquire about properties of such singularities. It may be that there are well-defined laws of 'naked singularity mechanics', just as there are the laws of black hole mechanics (though there is no indication at the moment that such a thing is true). At present there is only some scattered knowledge about properties like curvature strength, stability of the Cauchy horizon and redshift.

Examples of both weak curvature and strong curvature naked singularities have been found (Singh 1996). While spacetime cannot be extended through the latter kind of singularity, it may possibly be extendible through a weak singularity. For a discussion of extendibility see Clarke (1993).

If the Cauchy horizon accompanying a naked singularity were to be unstable, that could be evidence in favour of cosmic censorship. However, examples of stable as well as unstable Cauchy horizons are known in classical collapse. (See Penrose (1998) for some more discussion on Cauchy horizon stability).

The redshift of the null rays emanating from a singularity can be shown to be infinite, in the known examples, assuming that the standard redshift definition can be used all the way up to a singularity. In this sense, naked singularities are as black as black holes themselves. However, this does not appear to be a good way to preserve censorship because ultimately quantum effects near the naked singularity must be taken into account, and these will serve to distinguish a black hole from a naked singularity.

It can also be shown in a straightforward way that any shell-focusing naked singularities that might form in spherical collapse are necessarily massless (Lake 1992; Cooperstock *et al.* 1997). It can be said that if a naked singularity forms, its most significant property is that regions of extremely high curvature are exposed. This will have observable consequences which will be essentially unaffected by the other properties mentioned above. Hence these other properties can only have secondary importance.

2.8 *Status of the cosmic censorship hypothesis*

Until we learn something definite about non-spherical collapse, it is not possible to conclude about the validity of the hypothesis. However I would like to suggest, on the

basis of what is known, that the hypothesis is unlikely to be true in classical general relativity. We also note that we are regarding a visible region of unbounded high curvature as a violation of censorship, even if this visible region does not contain an actual singularity.

The examples of naked singularities known in spherical collapse arise for various forms of matter. This includes dust, perfect fluids, imperfect fluids and scalar fields. There are also some general arguments suggesting the occurrence of naked singularities for any form of matter satisfying the weak energy condition (e.g. an existence proof, and the behaviour of null geodesic expansion). None of this, taken by itself, constitutes a proof. But, taken together, these arguments strongly suggest that visible regions of unbounded curvature arise generically in spherical gravitational collapse. Also, black holes arise generically in spherical collapse.

Now, we know from the singularity theorems that the occurrence of singularities in spherical collapse is stable against the introduction of non-spherical perturbations. In view of this, it is very hard to see why the naked singularities arising in spherical collapse should be unstable against non-spherical perturbations, whereas the black holes forming in spherical collapse should be stable against such perturbations.

It is only fair to say that different people have drawn widely different conclusions about cosmic censorship from the currently known examples. Since our viewpoint is that censorship possibly does not hold in classical relativity, we would like to ask next if naked singularities could actually occur in nature, and if they do, what would they look like.

3. Are there naked singularities in nature?

3.1 *Maybe no ...*

Even if general relativity were to generically admit naked singularity solutions, it does not follow that these singularities actually occur in nature. It could be that stars simply do not possess the initial conditions necessary for formation of naked singularities.

Furthermore, there could actually be some principle, over and above general relativity, which forbids naked singularities. This would be in the same spirit in which the advanced wave solutions of electrodynamics are forbidden. We have recently pursued a line of thought wherein the second law of thermodynamics prohibits naked singularities (Barve & Singh 1997).

Our idea can be deduced from Penrose's work on the second law of thermodynamics (Penrose 1981). As explained by Penrose, a fundamental understanding of the second law can be had only if we understand why the initial entropy of the Universe is so low, compared to the maximal value it could have had. The matter, including radiation, was itself in a high entropy state because of the thermal equilibrium that prevailed soon after the Big Bang. Hence there must be an entropy associated with the gravitational field and this gravitational entropy must have been initially very low, so that the net entropy (matter plus gravity) becomes extremely small.

Such a gravitational entropy will have to be defined from the Riemann curvature. The Ricci part of the curvature diverges at a Friedmann Big Bang singularity. Since

we are interested in a low gravity entropy at the Big Bang, it is plausible that this entropy is related to the Weyl part of the curvature, which is zero at the Friedmann singularity. This has come to be known as the Weyl Curvature Hypothesis: in order to have an explanation of the second law, the Weyl curvature must be zero (or at least negligible compared to the Ricci curvature) at the initial cosmological singularity. It should be said though that a concrete mathematical relation between the Weyl curvature and gravitational entropy has not yet been found.

It is possible to regard a naked singularity forming in collapse as an 'initial' singularity, because geodesics terminate in the past at the singularity. Hence it is reasonable to require that only those naked singularities can occur which satisfy the Weyl hypothesis. That is, a suitable quantity constructed from the Weyl curvature must go to zero as the naked singularity is approached in the past along an outgoing geodesic. If a naked singularity solution occurs in general relativity but violates the Weyl hypothesis then its existence in nature is forbidden by the second law. Such a naked singularity is a singularity with very high initial gravitational entropy, contrary to what is expected for the second law to hold.

We tested the behaviour of the Weyl scalar in a few simple examples of spherical naked singularities and found it to diverge at the singularity, along outgoing geodesics. It diverges as fast as the Ricci part, and hence violates the Weyl hypothesis. Since strong inhomogeneity tends to favour a high Weyl, and since it also favours naked singularities, it is likely that this divergence behaviour is generic to naked singularities. Thus naked singularities may be anti-thermodynamic entities.

I do not know of any easy way out of this line of argument. The argument may fail only if it turns out that there is actually no connection between gravitational entropy and the Weyl curvature.

3.2 *Maybe yes . . .*

It would of course be a much more interesting state of affairs if no principle forbids naked singularities, and if they were to be found in nature. Hence we would like to enquire what the observational signatures of naked singularities will be. It is nearly certain that naked singularities will not emit significantly through classical processes, because of the extremely large redshifts. However, a quantum treatment of the matter and of gravity will be unavoidable near the singularity, and these quantum effects will result in an observable emission, in spite of the large classical redshift. In the absence of a quantum theory of gravity, the best one can do is compute the quantum particle creation in the classical gravitational field which becomes very strong as the singularity is reached. This semiclassical treatment will in fact suffice until the final Planck epoch prior to the singularity formation.

Thus, in effect one is asking what is the analog of Hawking radiation in the case when a star collapses to a naked singularity. Answering this is not as direct as the Hawking radiation calculation for a black hole, because of the presence of a Cauchy horizon. Part of the future null infinity is exposed to the naked singularity and hence one cannot perform the usual expansion of matter field modes in the future. This prevents the usual Bogoliubov transformation and the standard particle creation calculation from being carried out. This is one of the reasons why not much work has been done on this important problem and computation is still in its infancy.

One way out is to compute the quantum expectation value of the stress energy tensor—this can be calculated locally, and on future null infinity, in the approach to the Cauchy horizon. The outgoing flux of radiation is a measure of the emission from the naked singularity. In four dimensions an exact calculation is not possible, but the outgoing flux in the geometric optics approximation has been calculated (Ford & Parker 1978) for dust shell-crossing singularities. In this case the flux does not diverge in the approach to singularity formation. We (Barve, Singh, Vaz & Witten 1998) have recently used the method of Ford and Parker to compute the flux of a massless scalar field on the Cauchy horizon resulting from a shell-focusing dust naked singularity. This time the flux diverges, suggesting that the back-reaction will avoid formation of the naked singularity. In an interesting paper, Vaz and Witten (Vaz & Witten 1998) have calculated the spectrum of this radiation, and shown it to be very different from the black-body spectrum of Hawking radiation.

In two dimensions, as a result of the conformal anomaly, the outgoing flux can be calculated exactly, without having to resort to the geometric optics approximation. This was earlier done (Hiscock, Williams & Eardley 1982) for the null-dust (Vaidya) naked singularity and repeated by us (Barve *et al.* 1998) for the dust (Tolman-Bondi) naked singularity. In both cases the flux diverges on the Cauchy horizon. Similar features have been found in studies (Vaz & Witten 1994; 1996; 1997) of the quantum behaviour of naked singularities in some string-inspired gravity models. The back-reaction calculation in all the above models is extremely hard to perform, as can be expected. But it is plausible that essentially the back-reaction will remove the classical naked singularity, without significantly affecting the flux emitted to infinity.

The observable signature of a naked singularity appears to be the burst of radiation emitted as the Cauchy horizon is approached, and the characteristic non-thermal spectrum which this radiation possesses. This is to be contrasted with the slow evaporation of a quantum black hole via black-body radiation. It would be important to generalise the above results to find the typical signatures of quantum naked singularities and to explore if there are any astrophysical objects whose properties resemble those of a naked singularity.

Acknowledgments

I would like to thank the organizers for inviting me to give a talk at this meeting. It is a pleasure to thank Sukratu Barve, Srirang Deshingkar, I. H. Dwivedi, Sanjay Jhingan, Pankaj Joshi, Giulio Magli, Cenalo Vaz and Louis Witten for many useful discussions. I acknowledge partial support of the *Junta Nacional de Investigacão Cient'ifica e Tecnológica* (JNICT) Portugal, under contract number CERN/S/FAE/1172/97.

References

Barve, S., Singh, T. P. 1997, *Mod. Phys. Lett. A*, **12**, 2415; gr-qc/9705060.
Barve, S., Singh, T. P., Vaz, C, Witten, L. 1998, *Nucl. Phys*., **B532,** 361.
Barve, S., Singh, T. P., Vaz, C, Witten, L. 1998, *Phys. Rev*., **D58,** 104018.
Barve, S., Singh, T. P., Witten, L. 1999, gr-qc/9901080..
Brady, P. R. 1995, *Phys. Rev. D*, **51**, 4168.

Christodoulou, D. 1986, *Commun. Math. Phys.*, **105**, 337; 1986, *ibid.*, **106**, 587; 1987a, *ibid*., **109**, 591; 1987b, *ibid*, **109**, 613; 1993, *Commun. Pure Appl. Math*., **XLIV** 339; 1993, *ibid*, **XLVI**, 1131; 1994, *Ann. Math*., bf 140 607.
Clarke, C. J. S. 1993, *Analysis of spacetime singularities*, (Cambridge University Press, 1993) and in this volume.
Clarke, C. J. S. 1993, *Class. Quant. Grav.*, **10**, 1375.
Choptuik, M. W. 1993, *Phys. Rev. Lett.*, **70**, 9.
Choptuik, M. W. 1998, grqc/9803075.
Cooperstock, F. I., Jhingan, S., Joshi, P. S., Singh, T. P. 1997, *Class. Quant. Grav*., **14**, 2195.
Dwivedi, I. H., Joshi, P. S. 1994, *Commun. Math. Phys*., **166**, 117.
Dwivedi, I. H., Joshi, P. S. 1997, *Class. Quant. Grav.*, **14**.
Ford, L. H., Parker, L. 1978, *Phys. Rev. D*, **17**, 148.
Gundlach, C. 1997, grqc/9712084.
Hiscock, W. A., Williams, L. G., Eardley, D. M. 1982, *Phys. Rev. D*, **26**, 751.
Herrera, L., Prisco, A. Di, Hernandez-Pastora,J. L., Santos, N. O. 1997, grqc/9711002.
Iguchi, H., Nakao, K., Harada, T. 1997, Kyoto University preprint KUNS 1475, to appear in *Phys. Rev.* D.
Joshi, P. S., Dwivedi, I. H. 1992, *Commun. Math. Phys*., **146**, 333.
Joshi, P. S. 1993, *Global Aspects in Gravitation and Cosmology* (Oxford, 1993).
Joshi, P. S. 1997, in *Singularities*, *Black Holes and Cosmic Censorship* (ed.) P. S. Joshi (IUCAA, Pune, 1997), grqc/9702036.
Lake, K., 1992, *Phys. Rev. Lett*, **68**, 3129
Lifshitz, E. M., Khalatnikov, I. M. 1961, *Soviet Physics JETP*, **12**, 108.
Magli, G. 1997, *Class. Quant. Grav*., **14**, 1937; grqc/9711082.
Ori, A., Piran, T. 1990, *Phys. Rev. D*, **42**, 1068.
Penrose, R. 1969, *Rivista del Nuovo Cimento*, **1**, 252.
Penrose, R. 1981, in *Quantum Gravity 2* (ed.) C. J. Isham, R. Penrose & D. W. Sciama (Oxford).
Penrose, R. 1998, in *Black Holes and Relativistic Stars* (ed.) R. M. Wald (Chicago University Press, 1998), and article in this volume.
Podurets, M. A. 1966, *Soviet Physics-Doklady*, **11**, 275.
Rees, M. J. 1998, in *Black Holes and Relativistic Stars* (ed.) R. M. Wald (Chicago University Press, 1998), and article in this volume.
Singh, T. P. 1996, in *Classical and Quantum Aspects of Gravitation and Cosmology* (ed.) G. Date & B. R. Iyer (Inst, of Math. Sc, Madras), grqc/9606016.
Singh, T. P. 1998, *Phys. Rev*., **D58**, 024004.
Singh, T. P., Witten, L. 1997, *Class. Quant. Grav.*, **14**, 3489.
Szekeres, P., Iyer, V. 1993, *Phys. Rev. D*, **47**, 4362.
Vaz, C, Witten, L. 1994, *Phys. Lett. B*, **325**, 27; 1996, *Class. Quant. Grav*., **13**, L59; 1997, *Nucl. Phys. B*., **487**, 409.
Vaz, C, Witten, L. 1997, *Phys. Lett.*, **B442**, 90.
Wald, R. M. 1998, grqc/9710068.
Wald, R. M. 1998, in *Black Holes and Relativistic Stars* (ed.) R. M. Wald (Chicago University Press, 1998).

J. Astrophys. Astr. (1999) **20**, 233–248

The Question of Cosmic Censorship*

Roger Penrose, *Department of Mathematics*, *University of Oxford*, *24-29*, *St. Giles*, *Oxford 0X1 3BD*, *UK*
email: rouse@maths.ox.ac.uk

Abstract. Cosmic censorship is discussed in its various facets. It is concluded that rather little clear-cut progress has been made to date, and that the question is still very much open.

1. The role of cosmic censorship in gravitational collapse

Chandra's famous work on the maximum mass of white dwarf stars (Chandrasekhar 1931) pointed the way to our present-day picture (cf. Hawking & Penrose 1970) that for bodies of too large a mass, concentrated in too small a volume, unstoppable collapse will ensue, leading to a singularity in the very structure of space-time. The deduction of a strict occurrence of an actual singularity in physical space-time would, however, be based on an assumption that no quantum-mechanical principles intervene to change the nature of space-time from that which is classically described by Einstein's general relativity. Indeed, the term 'singularity', in this physical context, really refers to a region where the conventional classical picture of space-time breaks down, to be replaced, presumably, by whatever physics is to go under the name of 'quantum gravity'. It is the normal expectation that this occurs only when classical space- time curvatures diverge — quantum effects taking over when radii of space-time curvature of the order of the Planck length are attained.

In the standard picture of collapse to a black hole (of. Penrose 1978, for example), these singularities are not visible to observers at a large distance from the hole, being 'shielded' from view by an absolute event horizon. Thus, whatever unknown physics takes place at the singularity itself, its effects are not observable by such an observer. The assumption of cosmic censorship is that in a generic gravitational collapse the resulting space-time singularity will indeed be shielded from view in this way. Accordingly, it is taken that the alternative of a naked singularity – i.e. a visible singularity – would not occur (except, conceivably, for some very special initial collapse configurations which could not be expected to take place in an actual astrophysical circumstance).

It is not hard to conceive of physical situations in which one of the standard criteria for 'unstoppable gravitations collapse' is satisfied. All that is required is for sufficient mass to fall into a small enough region. For the central region of a large galaxy, for example, the required concentration could occur with the stars in the region still being separated from each other, so there is no reason to expect that there could be some overriding physical principle which conspires always to prevent such

* Reprinted with permission from the book 'Black Holes and Relativistic Stars', edited by R.M. Wald and published by the University of Chicago Press, Illinois, USA.

unstoppable collapse. However, we cannot simply deduce from this that a black hole will be the result. This deduction requires the crucial assumption that cosmic censorship, in some form, holds true.

Two familiar mathematical criteria for 'unstoppable collapse' are the existence of a trapped surface or of a point whose future light cone begins to reconverge in every direction along the cone. In either of these situations, in the presence of some other mild and physically reasonable assumptions, like the nonnegativity of energy (plus the sum of pressures), the nonexistence of closed timelike curves, and some condition of genericity (like the assumption that every causal geodesic contains at least one point at which the Riemann curvature is not lined up in a particular way with the geodesic), it follows (by results in Hawking & Penrose 1970) that a space-time singularity of some kind must occur. (Technically: the space-time manifold must be geodesically incomplete in some timelike direction.)

It appears to be a not uncommon impression among workers in the field that as soon as one of these conditions is satisfied – say the existence of a trapped surface- then a black hole will occur; and, conversely, that a naked singularity will be the result if not. However, it should be made clear that neither of these deductions is in fact valid. The deduction that a black hole comes about whenever a trapped surface is formed *requires* the assumption of cosmic censorship. Moreover, the deduction that some kind of space-time singularity comes about (in general situations), whether or not it is a naked one, requires some such assumption like that of the existence of a trapped surface (of. e.g. Hawking & Penrose 1970). Thus, the presence of a trapped surface does not imply the absence of naked singularities; still less does the absence of a trapped surface imply the presence of a naked singularity, These points have relevance to various investigations which attempts to address the issue of cosmic censorship by the study of specific examples. Frequently, the absence of a trapped surface appears to be regarded as a criterion for - or at least a strong indication of – cosmic censorship violation (cf. for example, Shaprio & Teukolsky 1991). It should be clear from the above remarks that the issue is by no means as simple as that.

While the question of cosmic censorship remains very much an open one at the present time – possibly the most important unsolved problem in classical general relativity – progress in certain areas has been made, and I shall attempt to address some of these in the following remarks. However, these should not be regarded as in any way a comprehensive survey of progress in these areas, but merely as a personal assessment of the present status of the subject.

2. Plausibility of criteria for unstoppable collapse

Before addressing the issue of cosmic censorship directly, it will be appropriate to make a few remarks concerning the question of whether the above criteria –namely, the existence of a trapped surface or of a reconverging light cone – actually do represent conditions that would be realized when too much mass is concentrated in too small a volume. Various researchers (most particularly Shoen & Yau 1993) have presented arguments to show that sufficiently large mass concentrations do indeed lead to the presence of trapped surfaces. I shall not attempt to summarize this area of work here, The arguments are mathematically quite difficult, but the physical implications of these arguments are still far from clear to me. On the other hand, the

reconverging light cone condition is easier to see as representing a plausible criterion. I shall have a few comments to make concerning this case. Basically, the argument for the physical realizability of the reconverging light cone condition is that given in Penrose (1969). Imagine a certain amount of massive material, say of total mass $\mathcal{M}$, and allow it to fall to within a roughly defined region whose diameter is of the general order of 4 GM. We consider a space-time point p somewhere in the middle of this region, and examine the future light cone $\mathcal{C}$ of p. Thus, $\mathcal{C}$ is swept out by the future-endless rays (null geodesies) with past endpoint p. The strict condition that $\mathcal{C}$ 'satisfies the reconverging light cone condition' would be that on every ray γ generating $\mathcal{C}$ there is a place where the divergence of the rays changes sign. If it is assumed that such a ray is geodesically complete in the future direction (and that the energy flux across the ray is nonnegative), then it follows that, to the future of p along the ray, there is a point *conjugate* to p (i.e. a point', distinct from p, with the property that there is a 'neighbouring ray to γ' which intersects γ to p and again at q; more precisely, there is a nontrivial Jacobi field along γ which vanishes both at p and at q). The idea is that as the material falls in across $\mathcal{C}$ it causes "focussing power" in the lensing effect of the Ricci tensor component along the ray (namely, $R_{ab}l^a l^b$ where l^a is a null tangent vector to γ) due to the energy density in the matter falling in across $\mathcal{C}$. There are simple integral expressions that can be written down (cf. Clarke 1993, in particular) which provide sufficient conditions for a conjugate point to arise, so it is merely an order-of-magnitude requirement that there is sufficient infall of material to ensure that the focussing condition will be satisfied. The situation could be made to be qualitatively similar to the original Oppenheimer-Snyder (1939) collapsing dust cloud (pressureless fluid), but where there is no symmetry assumed and no particular equation of stake employed (like that of Oppenheimer and Snyder's dust).

However, the strict form of the reconverging light cone condition is that *every* ray through p should encounter sufficient material for divergence reversal to occur. This condition might well be considered to be unreasonably strong. It might not be evidently satisfied if the collapsing material is concentrated in a number of separated bodies, say stars, rather that in a continuous medium. Many of the rays through p might then miss the actual collapsing material, so the focussing along such a ray could be much smaller than required (a point specifically raised with me by Robert Wald), being only a secondary effect due to the nonlocal overall focussing produced by Weyl curvature. In face, this does not make a serious difference, but to see that it does not is not entirely straight forward. The essential point is that for the purposes of the singularity theorem being appealed to here (Hawking & Penrose 1970), it is not necessary to assume that every future ray through ρ encounter divergence reversal. All that is required, roughly speaking, is that the elements of area of cross section of the intersection $\mathcal{C} \cap \partial I^+(p)$ of $\mathcal{C}$ with the boundary $\partial I^+(p)$ of the (chronological) future I^+ of p should eventually decrease in future directions, at every point of the cross section. At those places where a sufficient amount of the matter directly encounters 'crossing regions' of $\mathcal{C}$, where two parts of this null hypersurface encounter one another and both cross through into the interior of I^+ (so that they do not remain on the region of $\mathcal{C}$ that lies on $\partial I^+(p)$). To see that this must be the case when a sufficient concentration of material encounters $\mathcal{C}$, one may appeal to the qualitative similarity between the situations arising when the collapsing material consists of a continuous and fairly uniform medium, and when it consists of a number of discrete bodies (such as stars or, at a different level of description, the constituent

particles of the medium) which closely approximate that medium. This is sufficient for establishing that p constitutes a 'future-trapped set', in the sense that is required for Hawking & Penrose (1970), and the deduction of the presence of a singularity follows just as before. I shall not go into the details of this argument here, it being more appropriate to leave this for some later discussion. (It has a bearing on other matters that have been addressed in the literature, concerning the lensing effects of different kinds of mass distribution; c f. Holz & Wald 1997).

3. Causal definition of naked singularities

Let us take it, then, that there will be certain astrophysical situations in which such 'unstoppable' gravitational collapse actually takes place—'unstoppable' in the sense that it leads to some kind of space-time singularity (in accordance with the singularity theorems). When this occurs, and assuming the form of cosmic censorship which assets that no such singularity is naked in the sense of being visible to some observer at infinity, then the boundary of the past of the 'set of observers at infinity' constitutes the *absolute event horizon.* External to this horizon are the space-time events which can send signals to infinity; cosmic censorship asserts that no singularity can lie within this external region. This presents us with the classical situation of a black hole.

Normally, the cosmic censorship issue is phrased in some such way, i.e. in terms of observers at large distances from the collapse — and, in precise mathematical discussions, in terms of observers at future null infinity $\mathcal{I}^+$. The absolute event horizon is then the boundary $\partial I^-[\mathcal{I}^+]$ of the past of $\mathcal{I}^+$. However, it may be felt that this is not necessary the appropriate notion of cosmic censorship. For one could imagine a situation in which an observer, near to a gravitationally collapsing body, witnesses a naked singularity arising, this singularity being visible to that particular observer. We might envisage that, in this universe model, the observer and the observed collapse are both within a large region of mass concentration which ultimately – perhaps on a cosmological timescale collapses to trap both the observer and observed region, thereby preventing signals from reaching 'infinity'. In such a picture, the singularity would not be 'naked' in terms of the definition which uses $\mathcal{I}^+$ since signals reaching the observer are themselves the singularity, so the singularity would indeed be naked in a more local sense. There being no theory governing what happens as the result of the appearance of such a singularity, this particular observer would not be able to account, in scientific terms, for whatever physical behaviour is seen. This is the kind of 'unpredictability' that one wishes to avoid in a physical situation. One of the reasons for desiring a cosmic censorship principle, after all, is the elimination of such physical uncertainties. A definition of 'naked singularity' which depends upon what is accessible to observers at infinity ($\mathcal{I}^+$) does not achieve this satisfactorily.

Thus, it seems not reasonable to posit a somewhat stronger form of cosmic censorship than one phrased in terms of observers at infinity. In accordance with this, an appropriate notion of *strong cosmic censorship* has been formulated (Penrose 1978) which has the added advantage that it turns out to be symmetrical in time. Moreover, for a given space-time $\mathcal{M}$ this notion turns out to be equivalent to *global hyperbolicity* for $\mathcal{M}$.

The condition can be phrased (see Geroch, Kronheimer, and Penrose 1972) in terms of the notion of *indecomposable pasts* – IPs – and *indecomposable futures* - IFs. An IP is a past-set (i.e. a subset of $\mathcal{M}$ which is the same as its own chronological past) which, in addition, is not the proper union of two other past-sets. An IF is defined correspondingly, with 'future' replacing 'past'. It can be shown (Geroch, Kronheimer & Penrose 1972) that the IPs are precisely the pasts of timelike (or causal) curves in $\mathcal{M}$ and the IFs are the futures of such curves. Such a timelike (or causal) curve is said to generate the IP and IF in question. An IP is called *proper* – a PIP – if it consists of the set of points lying to the chronological past of some point of $\mathcal{M}$ (which would be the future endpoint of such a generating curve). A PIF is, correspondingly, the (chronological) future of some point of $\mathcal{M}$. A terminal IP – a TIP – is an IP which is not a PIP; a terminal IF – a TIF – is an IF which is not a PIF. Sometimes, the TIPS and TIFs are called ideal points for the space-time $\mathcal{M}$.

Let us assume that $\mathcal{M}$ is *strongly causal* (i.e. that every point of $\mathcal{M}$ has an arbitrarily small neighbourhood such that no causal curve leaves and then reenters it; (cf. Hawking & Ellis 1973; Penrose 1972). In face, it is sufficient, in what follows; to assume that $\mathcal{M}$ is both *future distinguishing* – i.e. that no two points of $\mathcal{M}$ have the same chronological future) – and *past distinguishing* – i.e. that no two points of $\mathcal{M}$ have the same chronological past). Then the points of $\mathcal{M}$ are canonically in one-to-one correspondence with the PIPs of $\mathcal{M}$ and also in one-to-one correspondence with the PIFs of $\mathcal{M}$. he TIPs and TIFs of $\mathcal{M}$ namely, $\mathcal{M}$'s ideal points, provide the points of what is called the *causal boundary* $\mathcal{M}$ of $\mathcal{M}$ The *future* causal boundary $\partial^+\mathcal{M}$ of $\mathcal{M}$ is the set of TIPs of $\mathcal{M}$ and the *past* causal boundary $\partial^-\mathcal{M}$ is the set of TIFs of $\mathcal{M}$. As thus defined, $\imath\mathcal{M}$ is a disjoint union of $\partial\mathcal{M}$ and $\partial\mathcal{M}$ There are circumstances under which it may be felt that certain of the points of $\partial^+\mathcal{M}$ should be identified with certain of the points of $\partial\mathcal{M}$ (see Geroch, Kronheimer & Penrose 1972; such situations could be considered to represent violations of cosmic censorship). However, for the purpose of this article, I shall prefer to regard the two sets $\partial^+\mathcal{M}$ and $\partial-\mathcal{M}$ as being actually disjoint. The entire union $\partial\mathcal{M} \cup \partial-\mathcal{M} \cup \mathcal{M} = \partial\mathcal{M} \cup \mathcal{M}$ is the causal closure of $\overline{\mathcal{M}}$ of $\mathcal{M}$.

It is convenient to divide these ideal points into two classes, according to whether they are to represent *singular* points of $\mathcal{M}$ or *points at infinity* for $\mathcal{M}$. The simplest way to make such a distinction if to say that a TIP represents a point at (future) *infinity* – an ∞ -TIP–if it is generated by a timelike curve that is of infinite length into the future, and a TIF represents a point at (past) infinity – an ∞-bf TIF – if it is generated by a timelike curve that is of infinite length into the past. The remaining TIPs and TIFs – the singular TIPs and TIFs then represent the singular points of $\mathcal{M}$. Although the distinction between points as infinity and singular points appears to be the simplest, it is not always regarded as the most appropriate (cf. Clarke 1993 for example). According to the definition given here (taken from Penrose 1978) one can have a 'point at infinity' for which the space-time curvature diverges as that ideal point is approached. It might be reasonable to call such an ideal point a 'singular point at infinity', but it would be given by an ∞-TIP or ∞-TIF as defined here, nevertheless, and not by what I am referring to as a 'singular TIP' or 'singular TIF'.

Now, a naked singularity may be described as a singularity which lies to the future of some point of space-time, but which can also be 'seen' by some observer. The reason for the first part of this description is that one would not want some cosmic

censorship principle to exclude the big bang; that is to say, the big bang should not count as a 'naked singularity'. Thus, a naked singularity lies both to the future of some point $p \in \mathcal{M}$ and to the past of some other point $q \in \mathcal{M}$. In terms of TIPs, a *naked singular* TIP would be a singular TIP R which contains the point p and which lies to the past of a point q, i.e.

$$p \in R \text{ and } R \subset I^{-}(q) \text{ for some } p, q \in \mathcal{M}. \tag{1}$$

Here the standard notation $I^{-}(q)$ is used for the chronological past of a point 'correspondingly, $I^{-}[Q]$ stands for the chronological past of a *set* Q, and $I^{+}(q)$ and $I^{+}[Q]$ stand for the chronological futures of a point p and a set', respectively (notations already employed above). A point at infinity might also be 'naked', in the same sense, so we can define a naked ∞ — TlPR in just the same way. We can similarly define a naked singular TIF or ∞ – TIP's. TIF as a TIF S for which:

$$S \subset I^{+}(p) \text{ and } q \in S \text{ for some } p, q \in \mathcal{M}. \tag{2}$$

4. Strong cosmic censorship

Many of the reasons for wishing to exclude naked singular apply also to naked points at infinity. It is not better, from the point of view of uniqueness of evaluation, that a singularity. (Anti-de Sitter space in an example of a space-time possessing naked points at infinity; there are indeed reasons of this kind for regarding this model as 'unphysical'.) In any case, as was remarked upon above some ∞ – TIPs or ∞ – TIFs might arise from regions of infinite curvature and could be thought of as, in some sense, 'singular', in any case. Thus, it seems to be a reasonable formulation of the requirement that a space-time $\mathcal{M}$ be in accordance with cosmic censorship that there should be no naked TIPs in the above sense, whether they be singular TIPs or ∞ - TIPs. As was shown in Penrose (1979), this condition is *equivalent* to the condition that $\mathcal{M}$ be free of naked TIPs–by virtue of the fact that it is equivalent, also, to the condition that $\mathcal{M}$ be *globally hyperbolic*

The assertion that $\mathcal{M}$ is free of naked singularities in this sense can also be phrased in other equivalent ways. It is convenient to introduce causal relations between TIPs as follows. We say that the TIP P *causally precedes* the TIP Q if $P \subseteq Q$, and that the TIP P *chronologically precedes* the TIP Q if there exists a point $q \in Q$, such that $p \subset I^{-}(q)$. Then, by the above definiton, a TIP is naked if there is another TIP to its chronological future. Defining the causal relations between TIFs correspondingly if R and S are TIFs, R causally precedes S if S $S \subseteq R$ and R chronologically precedes S if there exists $r \in R$ such that $S \in I^{+}(\mathrm{r})$–we see that the TIF S is naked if there is another TIF to its chronological past. In this sense, naked singularities (or points at infinity) are *timelike* entities.

We can now formulate our principle of *strong cosmic censorship* as the assertion that naked singularities or points at infinity (in the above sense) do not occur in generic space-times, where it is assumed that Einstein's equations hold with some reasonable equations of state for the matter. This is still a little vague because of occurrence of the words 'generic' and 'reasonable' in the definition. In my opinion, it is probably not particularly helpful to try to be more precise at this stage. There is a fairly well defined intuitive meaning for the word 'generic'; no doubt, when the

appropriate theorem comes along, then an appropriately relevant definition of 'generic' will become clearer (as was the case with some singularity theorems; cf. Hawking & Penrose 1970). Without some such restriction, however, counterexamples to cosmic censorship can occur, such as in certain very special situations of spherically symmetrical collapse (cf. Christodoulou 1994; Choptuik 1993). As regards 'reasonable equations of state' the essential points is that we should exclude equations that could lead to singularities in generic situations even in special relativity (e.g. leading to infinite density), such as occurs with 'dust' when caustics arise in the flow lines. The equations of state should also be such as to ensure energy positivity (and perhaps stronger restrictions such as the dominant energy condition).

As yet, there is no mathematical theorem asserting the truth of any appropriate form of cosmic censorship in general relativity; yet, as we have seen above, cosmic censorship is an essential ingredient of the standard picture of gravitational collapse to a lack hole. As was indicated in section 5.2 above, one can present convincing arguments to show that situations can occur – and, indeed, *will* occur in appropriate circumstances of gravitational collapse when sufficient matter is being concentrated in too small a region in which singularities will arise according to general relativity. But without a cosmic censorship assumption, there is no guarantee that these singularities will not be naked. If the singular region is not naked, even merely in the weaker sense, that observers at *infinity* are not able to 'see' the singular region (example so that no ∞-TIP representing a point of $\mathcal{I}^+$ contains a naked singular TIP), then there will be some region of the space-time including the singularity region, that cannot be seen from $\mathcal{I}^+$. In other words, the causal past of $\mathcal{I}^+$ (written $J^-[\mathcal{I}^+$]is not the whole of the space-time M so the chronological past $I^-[\mathcal{I}^+]$ cannot exhaust $\mathcal{M}$ either. As stated in section 5.3 above, the boundary ∂I defines the *horizon* of the resulting black hole.

There is another crucial role that is played by cosmic censorship in the theoretical discussions of gravitational collapse. Without such an assumption, one cannot deduce the well known *area increase* theorem (cf. Floyd & Penrose 1971; Hawking 1972). This theorem asserts that the areas of cross section of a black-hole horizon $\partial I^-[I^+]$ are nondecreasing into the future. There are various different versions of this theorem, depending upon which version of cosmic censorship is adopted (see Penrose 1978 for a discussion of several of these). The area-increase theorem is important, particularly because of its relation to black-hole entropy and thermodynamics (Bekenstein 1973; Hawking 1975).

5. Thunderbolts

One remark should be made here concerning censorship proposals of this nature. They do not, as they stand, eliminate the possibility of what Hawking (1993) refers to as *thunderbolts,* first considered in Penrose (1978). This is the hypothetical situation according to which a gravitational collapse results in a 'wave of singularity' coming out from the collapse region, which destroys the universe as it goes! On this picture, the entire space-time could remain globally hyperbolic since everything beyond the domain of dependence of some initial hypersurface is cut off ('destroyed') by the singular wave. An observer, whether at infinity or in some finite location in the space-time, is destroyed just at the moment that the singularity would have become visible,

so that observer cannot actually 'see' the singularity. One condition which excludes this particular possibility (Penrose 1978, conditions CC4) is

no ∞–TIP contains a singular TIP.

For an asymptotically flat space-time $\mathcal{M}$ we may expect that the future-null conformal boundary $\mathcal{I}^+$ of $\mathcal{M}$ can be identified with its set of ∞ –TIPs. In any situation where the above condition is violated, we have an ∞ – TIP which directly 'sees' the singularity (in the sense of being causally to its future). In the situation where a thunderbolt is present, the 'observer at infinity' represented by that ∞ – TIP would be destroyed by the wave of infinite curvative at that very moment, but we still have an ideal point there, represented by this ∞ – TIP. However, the conformal boundary $\mathcal{I}^+$ would cease to be smooth at that point.

In section 5.4, one formulation of the condition of strong cosmic censorship was given as an assertion that 'timelike' singularities (or points at infinity) are to be excluded. The above condition for ruling out thunderbolts can be phrased as the condition that a point at infinity cannot lie *causally* to the future of a singular point (defined in terms of TIPs). One could imagine formulating an 'extrastrong' version of cosmic censorship in which *all* causally separated (distinct) TIPs are excluded (in the sense that no TIP shall properly include another TIP; cf. section 5.4) and not merely the timelike separated ones that are excluded by ordinary strong cosmic censorship. However, this condition would be unreasonable because it would rule out all asymptotically flat space-times! Being a null hypersurface, the future of the conformal boundary $\mathcal{I}^+$ of an asymptotically flat $\mathcal{M}$ contains null generators, and any pair of TIPs representing two distinct points of the same generator would be causally separated in the above sense.

However, it might well be reasonable to expect that a slightly weaker extrastrong version of cosmic censorship might be appropriate, in which it is asserted that there is no pair of distinct TIPs P, Q such that $P \subset Q$, and where P is a singular TIP (and where the corresponding statement in terms of TIFs could also be appended, if desired). This would incorporate both strong cosmic censorship and the exclusion of thunderbolts, in the above sense. It remains to be seen whether such a formulation might still be too strong.

6. Some arguments against cosmic censorship

Most of the arguments presented to date which are aimed at disproving cosmic censorship have been concerned with the examination of specific examples. However, there is an inherent difficulty in using arguments of this kind, because any specific example that can be studied in detail is liable to be 'special' in some way or other, and unlikely to be considered to be 'generic' in some appropriate sense. At least, this applies to specific examples that can be studied analytically in detail. It may be that with the further development of numerical techniques and computer power, examples might eventually be considered which could indeed be argued to be appropriately 'generic'. On the other hand, there is the compensating difficulty that with numerical solutions there may be some doubt, in any particular case, whether a seeming singularity is actually a genuine singularity, or whether the singularity is indeed naked. As things stand, specific examples can only give *indications* as to whether cosmic censorship is likely to be true, not definitive answers.

The first example of a gravitational collapse leading to a naked singularity was that given by Yodzis, Seifert, & Muller zum Hagen (1973). They pointed out that even in exactly spherically symmetrical collapse, infinite-curvative naked singularities could arise with dust, owing to the presence of caustics in the family of dust world-lines (at which the density diverges), provided that these caustics occur before an absolute event horizon is reached. As suggested in section 5.4, such circumstances should not be considered as providing violations of cosmic censorship because the infinite densities that arise have nothing to do with general relativity (or, indeed with gravity at all) because they occur just as readily with the equations for dust in special relativity.

Such regions of infinite density are sometimes referred to as 'shell-crossing' singularities, since they are regions where the different shells of collapsing material begin to cross one another. However, this terminology is not altogether appropriate because the difficulties with infinite density do not occur in the regions where different dust flows actually cross each other, but at the boundary of such a region, where there is a caustic in the flow lines and one flow becomes three[1] Nevertheless, where there are indeed several superimposed flows, the energy momentum tensor of 'dust' cannot be used, but instead one has a sum of a number of different terms of this kind, i.e.

$$T_{ab} = \rho u_a u_b + \cdots + \tau w_a w_b, \tag{3}$$

where $\rho,\ldots,\tau$ are the respective densities of the different components of the dust (pressureless fluid) and where each of $u^a,\ldots,w^a$ is a unit future-timelike vector giving the direction its flow. For each component of the flow, the flow world-lines are geodesics, and each of ρu_a ,..., $\tau w a$ is divergence-free. Of course, regions of infinite density can still arise whenever one of the systems of flow lines encounters caustics.

We can generalize the above finite sum of terms to a situation in which there is a *continuum* of terms. This gives us an instance of the kind of system that is treated by the *Vlasov* equation. More generally, the Vlasov equation covers the cases when there is a continuous superposition of fluids which possess pressure – rather than being just 'dust' as in the cases considered above (the simpler case of a 'collisionless' fluid).

In the collapse situation studied by Shapiro & Teukolsky (1991), referred to in section 5.1, the Vlasov equation is used but (as was pointed out to me by Alan Rendall) the individual fluid components do not possess pressure (the 'collisionless' case), and it is not clear that the situation is free of the problems that occur with the Yodzis, Seifert & Muller zum Hagen (1973) example. Building upon earlier ideas of Thorne, who suggested that prolate spheriodal collapse might lead to naked singularities (because of a resemblance to *cylindrical* trapped surface-free collapse; cf. also Thorne 1972; Chrusciel 1990), Shapiro and Teukolsky consider the collapse of a prolate azisymmetrical body composed of collisionless material, and they argue that naked singularities can arise. They indicate the presence of regions at which the density diverges and argue from the absence of trapped surfaces that these singularities could well be naked. Moreover, they point out that their singular regions do not arise merely from infinite density, because they extend outside the matter region.

[1] It does not become merely *two* overlapping flows, for topological reasons. One of the three flows 'counts' as negative and the other two as positive, preserving the net count of overlapping flows, as one passes from one side of the caustic to the other.

However, as was pointed out in section 5.1 above, more needs to be established if we are to ascertain whether these singularities are indeed naked. In particular, we would need to examine the regions of the space-time lying to the future of the singularity, but this is not possible within the framework of the computer calculation that they carry out, since the calculation terminates as soon as the singularity is reached.

Moreover, as was pointed out by Iyer & Wald (1991), collapse situations that appear to resemble that of Shapiro and Teukolsky can be constructed where no trapped surfaces appear before *nonnaked* singularities arise. In both the Iyer-Wald example and the Shapiro-Teukolsky example, there is a reasonable-looking family of constant-time spacelike hypersurfaces according to which the time evolution is described. In neither example are there trapped surfaces before a singularity appears. However, the singularity is clearly *not* naked in the Iyer-Wald example, because their example is simply the ordinary extended Schwarzschild solution described according to a nonstandard time coordinate. This sheds considerable doubt on the Shapiro-Teukolsky suggestion that their singularity is actually naked.

A closely related situation was studied by Tod (1992). In this example, there is a collapsing shell of 'null matter' (a delta-function shell of massless dust) which fall into a region of Minkowski space that it surrounds. The mass density can vary arbitrarily with spatial direction, and the (convex, smooth) shape of the shell, at one particular time, can also be chosen arbitrarily. By choosing this shape to be a suitable prolate ellipsoid it is not hard to ensure that caustics in the collapsing shell – and hence singularities – arise before there are any trapped surfaces. The description is given in terms of standard t = const. hypersurfaces in the interior Minkowski space. Nevertheless, the situation is completely consistent with the conventional picture of gravitational collapse to a black hole. Trapped surfaces do in fact occur in the space-time, but not until after the t-value at which singularities arise. This is again similar to the Shapiro-Teukolsky situation, and there is no reason to expect a violation of cosmic censorship.

Other examples have been described (Choptuik 1993: Christodoulou 1994) in which the collapsing matter consists of a massless scalar field. In some of these, there are naked singularities. However, all these examples are extremely special, owing to the fact that spherical symmetry is assumed. Accordingly, it is hard to see that such examples can shed a great deal of light on the general issue of cosmic censorship. The condition of genericity is far from being satisfied. Moreover, according to a recent result of Christodoulou (1997), 'almost all' examples, even within *this* limited class, are free of naked singularities.

It would thus appear that there is, so far, no convincing evidence against cosmic censorship's being a principle with which classical general relativity accords. Quantum general relativity, on the other hand, does raise some serious problems in this regard. It is hard to avoid the conclusion that the endpoint of the Hawking evaporation of a black hole would be a naked singularity – or at least something that one a classical scale would closely resemble a naked singularity. Nevertheless these considerations are not directly relevant to what is normally referred to as 'cosmic censorship', which is intended to be a principle applying to *classical* general relativity only. When quantum effects are allowed, negative energy densities are possible – needed for the consistency of the Hawking effect, in which the area-increase property for a black-hole horizon is violated. In any case, unless there are mini-black holes in the universe (and the observational evidence seems to be against his), there would seem to be no

direct astrophysical or cosmological role for the 'objects' which represent the final stages of Hawking evaporation, owing to the absurdly long timescales needed for this process when it originates with an astrophysical black hole. (For theoretical considerations, on the other hand, these 'objects' could well be important – but that is another story!)

7. Some arguments in favour of cosmic censorship

There being no convincing evidence against cosmic censorship, we must ask whether, on the other hand, there is any convincing evidence in favour of it. Indeed, there are no results that I am aware of which give direct and convincing support to the view that there is a mathematical theorem asserting some form of cosmic censorship in classical general relativity. But are there any plausible general lines of argument aimed in this direction? I am not sure. In Penrose (1979, pp. 625–626), I put forward some rather vague suggestions of this nature, but, to my knowledge, these have not been followed up in a serious way. The idea was to try to show, roughly speaking, that Cauchy horizons are unstable, in some appropriate sense – at least for an initial Cauchy hypersurface Σ which is either compact or appropriately asymptotically flat. The idea would be that in the 'generic' case, the Cauchy horizon $H^+(\Sigma)$ would be replaced by a singularity, so that the maximal space-time consistent with evolution from Σ would in fact be the domain of dependence of Σ. This would have to be globally hyperbolic, i.e satisfy strong cosmic censorship. There is some evidence for such an instability (for asymptotically flat Σ) in work which shows that the 'inner horizon' (Cauchy horizon) of the Reissner-Nordström space-time (and of the Kerr space-time) is unstable (owing to the occurrence of infinitely blueshifted radiation); cf. Simpson & Penrose (1973), McNamara (1978a, 1978b), and Chandrasekhar & Hartle (1982). For further references concerning the issue of the (in)stability of black-hole Cauchy horizons in general relativity, see Ori (1997) and the article by Israel (chap. 7) in this volume.

On the other hand, there is some evidence that when there is a positive cosmological constant in Einstein's equations, stable Cauchy horizons may be possible. This situation comes about when the surface gravity of the cosmological horizon is greater than that of the Cauchy horizon which can occur with Reissner-Nordstrom-de Sitter and Kerr-de Sitter space-times (see Chambers & Moss 1994; cf. also Mellor & Moss 1990,1992, Brady & Poisson 1992). This situation is closely related to that considered below, in which inequalities arise from 'dropping particles into black holes'. As suggested below, it may well be that cosmic censorship requires a zero (or at least a nonpositive) cosmological constant.

Even if such a general result could be proved, it is not clear to me that this would really establish what is required for a suitable cosmic censorship theorem. It would not seem to rule out the thunderbolts discussed in section 5.5. This would really be necessary in order that the standard picture of gravitational collapse to a black hole can be obtained. What is the theoretical evidence that this picture is indeed likely to be always the correct one, according to classical general relativity? There seems to be little direct mathematical evidence. There are, however, certain rigorous mathematical results that give *indirect* support for cosmic censorship in this form. Ironically, these results have come about from a specific attempt to *disprove* cosmic censorship!

In Penrose (1973) I put forward a family of examples of gravitational collapse in which there is collapsing spherical shell of null dust, the density being an arbitrary function of direction out from the centre, the region inside the shell being Minkowski space. Shortly afterwards, Gibbons (1972) pointed out that there is no need for the shell to be spherical, and he considered this more general case of a smooth convex collapsing shell of null dust which surrounds a region of Minkowski space. (This is the generalization employed by Tod, 1992, referred to in section 5.6 above). As the shell collapses inwards, the matter density (the coefficient of a delta function) increases inversely as the area of cross section of the shell until it gets to a point where it can reverse the divergence of an intersecting outgoing light flash that originates in a region within the Minkowski space. If it does this all the way around, then the intersection S of that light flash with the infalling matter shell will be a trapped surface in the region just beyond the shell. All the geometry that needs to be considered for this takes place in Minkowski space. It depends only on the shapes of the collapsing shell and outgoing light flash, and on the matter density distribution on the shell.

Suppose that, in some particular shell geometry and matter distribution, it is possible to find an outgoing light flash for which S does provide us with a trapped surface. Then, according to the standard picture of collaspe to a black-hole - of which cosmic censorship is the most contentious part - the space-time will settle down to become a Kerr space-time in future asymptotic limit. If we assume this to be the case, we find that a certain geometrical inequality must hold true. Suppose the area of S is A_0, the area of the intersection of the absolute event horizon $\partial I^-[I^+]$ with the infalling matter shell is A_1, the future limit of the area of cross section of the (Kerr) horizon is A_2, and the area of the horizon of a Schwarzschild black hole of the same mass m is $A_3 = 16\pi m^2$ (units such that $G = c = 1$). We then have

$$A_0 \leq A_1 \leq A_2 \leq A_3 \leq 16\pi m_0^2, \tag{4}$$

where m_0 is the rest mass of the total energy-momentum of the incoming null dust shell[2].

The first inequality follows from the fact that the shell is infalling; the second follows from the area-increase theorem (which, as we recall, requires cosmic censorship); the third follows because the area of the Kerr horizon is smaller than that of Schwarzschild for the same mass; the fourth is a consequence of $m \leq m_0$, a relation which expresses the fact that, although there might be a loss of mass due to gravitational radiation, there will not be a gain (because of the Bondi-Schis mass-loss theorem and the asymptotic Minkowskian traingle inequality), the radiation being assumed to be entirely outgoing. Note that, in addition to cosmic censorship, there are (reasonable, but unproved) assumptions that the black hole actually *settles down* to become a Kerr space-time in the asymptotic limit (the known theorems simply assume-ing stationarity) and that the usual asymptotic assumptions for asymptotically flat space-times hold good (both at spacelike and null infinity).

The two quantities A_0 and m_0 depend only on the initial Minkowski space setup. If any such example could be found for which $A_0 > 16\pi m_0^2$, then this would provide a counterexample to the standard picture of gravitational collapse-essentially contra-

[2] That this is the rest mass of the incoming shell is a point that was glossed over in Penrose (1973).

dicting cosmic censorship. However, no example of this kind has ever been constructed. Moreover, various versions of the inequality $A_0 \leq 16\pi m_0^2$ have been proved by different authors (some of which refer to a somewhat different situation in which the geometry within a spacelike hypersurface is used); see Gibbons (1972, 1984, 1997) Jang and Wald (1977), Geroch (1973), and Huisken & Ilmanen (1997). Although none of these results directly established any form of cosmic censorship, they may be regarded as offering it some considerable support. Cosmic censorship could be said to supply a behind-the-scenes 'reason' why these inequalities are true!

There are also other types of inequalities which have been regarded as 'tests' of cosmic censorship. One may ask the question whether it is possible to 'spin up' a Kerr (or Kerr-Newman) black hole to a degree where its angular momentum exceeds the value for which a horizon is possible, by allowing particles to drop into it. The mass, angular momentum, and charge of the particles come into the calculation, and various inequalities relate these to the black hole's geometrical parameters, in order that the horizon be preserved. It appears that these inequalities are always satisfied (cf. for example, Wald 1974; Semiz 1990) — except, curiously, if there is a positive cosmological constant (a situation pointed out to me by Gary Horowitz; cf. Brill *et al* 1994) I am not sure of the significance of this final proviso. Of course, it might be the case that cosmic censorship requires a zero cosmological constant. We recall from section 5.4 that a negative cosmological constant (in anti-de Sitter space) leads to naked points at infinity. It is not at all inconceivable that a positive cosmological constant might correspondingly lead to naked singular points. This has relation to the issue of the instability of Cauchy horizons, as noted above.

The question of whether a black-hole horizon can be 'destroyed' by perturbing it with Mailing matter is really part of the general question of the stability of a black-hole horizon. An unstable horizon could be expected to lead to a naked singularity. As far as I am aware, the arguments that have been given for horizons stability are fairly firm, but not yet fully conclusive. It would be interesting to know whether the presence of a cosmological constant makes a significant difference. My own feelings are left somewhat uncertain by all these considerations.

8. Do we need new techniques?

As will be seen from the preceding remarks, we are still a long way from any definite conclusions concerning cosmic censorship. It is possible that radically new mathematical techniques will be required for any real progress to be made. My particular preferences would be for techniques related to developments in *twistor theory*. At the most immediate level, twistor theory is concerned with the geometry of the space $\mathbb{PT}$ of rays (null geodesics) in a space-time $\mathcal{M}$ The points of $\mathcal{M}$ would be regarded as secondary structures, and those of $\mathbb{PT}$ as being somewhat more fundamental. A point of $\mathcal{M}$ is interpreted, in $\mathbb{PT}$,in terms of the family of rays through that point. This family has the structure of a sphere in $\mathbb{PT}$ -in fact, a *Riemann* sphere, which is a 1-dimensional complex manifold. The central idea of twistor theory is to call upon the power of *complex analysis* (physically, because links with quantum mechanics). For this purpose, the 5-dimensional manifold $\mathbb{PT}$ must be thought of in terms of a larger complex manifold which, in the case when $\mathcal{M}$ is Minkowski space, turns out to be complex projective 3-space.

There are many problems, as yet unsolved, associated with how twistor theory is to be applied to general (vacuum) space-time, and it will be a long time before it has anything serious to say about cosmic censorship. Nevertheless, it has found a large number of applications (see, in particular, Bailey & Baston 1990; Mason & Woodhouse 1996). So far, it has not been significantly used to treat global questions in general relativity. The closest it has come to this is in the work of Low (1990, 1994), where some progress is made towards the understanding the causal structure of a space-time in terms of linking properties of spheres (or of loops, in the case of a space-time of 2 + 1 dimensions) in this space of rays.

In relation to this, it may be noted that there is a connection between cosmic censorship and the topology of $\mathbb{PT}$ if $\mathcal{M}$ satisfies strong cosmic censorship (i.e is globally hyperbolic), then the space $\mathbb{PT}$ is *Hausdorff,* whereas it is not Hausdorff in many cases where cosmic censorship fails. For any real progress to be made towards applying twistor theory to questions such as cosmic censorship, however, some major advances in understanding how the Einstein (vacuum) equations relate to twistor theory are needed. There does seem to be a deep connection between twistor theory and the Einstein equations, however – as yet elusive. This link is mediated through the equations for helicity 3/2 massless field (Penrose 1992). It has been known for some time that the consistency conditions for such fields (in potential form) are the Einstein vacuum equations (Buchdahl 1958; Deser and Zumino 1976 and Julia 1982). The other end of the link is the fact that the space of charges for such fields in *Minkowski* space is *twistor* space. Bringing together all the facets of this connection has proved to be a difficult problem (see Penrose 1996).

Although twistor theory remains a long way from addressing any significant issues of cosmic censorhip, it does have relevance to various issues connected with general relativity and space-time geometry (see Huggett & Tod 1985; Penrose & Ridler 1986; Penrose 1996). Perhaps it already has something to say about cosmology. The picture of a big bang leading to a Friedmann-type universe with negative spatial curvative and hyperbolic spatial geometry fits in well with the complex analytic (Riemann sphere) underlying philosophy, while the flat and closed spatial geometries do not do nearly so well (see Penrose 1997). Although negative spatial curvature cannot really be said to be a 'prediction' of the theory, it is perhaps the nearest to one, in general relativity and cosmology, that the theory has yet come up with.

Acknowledgements

I am particularly grateful to Robert Wald for various important remarks and for his help with the reference. I am also grateful to the NSF for support under contract PHY93-96246.

References

Bailey, T. N., Baston., eds 1990, Twistors in Mathematics and Physics, London Mathematical Society Lecture Notes Series, No. 156 (Cambridge University Press, Cambridge).

Bekeristein, J. 1973, *Phys. Rev.*, **D7**, 2333.

Brady, P. R, Poisson, E. 1992, *Class. Quant. Grav.*, **9**, 121.

Brill, D. R., Horowitz, G. T., Kastor, D, Traschen, J. 1994, *Phys. Rev.*, **D49**, 840.

Buchdahl, H. A. 1958, *Nuovo Cim*., **10**, 96.
Chambers, C. M., Moss, I. G. 1994, *Class. Quant. Grav.*, **11**, 1035.
Chandrasekhar, S. 1931, *Astrophys. J*., **74**, 81.
Chandrasekhar, S., Hartle, J. B. 1982, *Proc. R. Soc. London.*, **A384**, 301.
Christodoulou, D. 1994, *Ann. Math*., **140**, 607.
Christodoulou, D. 1997. The instabilities of naked singularities in the gravitational collapse of a scalar field (to appear).
Choptuik, M. 1993, *Phys. Rev. Lett*, **70**, 8.
Chrusciel, P. 1990, *Ann. Phys*., **202**, 100.
Clarke, C. J. S. 1993, The Analysis of space-time singularities, Cambridge Lecture Notes in Physics (Cambridge University Press, Cambridge).
Deser, S., Zumino, B. 1976, *Phys. Lett*., **B62**, 335.
Floyd, R. M., Penrose, R. 1971, *Nature*, *Phys. Sci*, **229**, 117.
Geroch, R. 1973, *Ann. N.Y. Acad. Sci*, **224**, 108.
Geroch, R, Kronheimer E. H., Penrose, R. 1972, *Proc. R. Soc. London.*, **A347**, 545.
Gibbons, G. W. 1972, *Commun. Math. Phys*., **27**, 87.
Gibbons, G. W. 1984, in Global Riemannian Geometry (eds) T. Willmore and N. J. Hitchin (Ellis Horwood, Chichester).
Gibbons, G. W. 1997, Collapsing shells and the isoperimetric inequality for black holes, hel-th/9701049.
Hawking, S. W. 1972, *Commun. Math. Phys*., **25**, 152.
Hawking, S. W. 1975, *Commun. Math. Phys*., **43**, 199.
Hawking, S. W. 1993, in The rennaissance of General relativity (in honour of D. W. Sciama), (eds) G. Ellis, A. Lanza and J. Miler (Cambridge University Press, Cambridge.)
Hawking, S. W., Ellis, G. F. R. 1973, The large scale structure of space-time (Cambridge University Press, Cambridge.)
Holz, D. E., Wald, R. M. 1997, A new method for determining cumulative gravitational lensing effects in inhomogenous universes, *Phys. Rev. D.* (Submitted).
Huggett, S. A., Tod, K. P. 1985, An introduction to Twistor Theory, London Mathematical Society Student Texts (L.M.S., London)
Huisken, G., Ilmanen, T. 1997, Proof of the Penrose inequality (to appear).
Iyer, V., Wald, R. M. 1991, *Phys. Rev*., **D44**, 3719.
Jang, P. S., Wald, R. M. 1977, *J. Math.*, *Phys*., **18**, 41.
Julia, B. 1982, *Comptes Rendus Acad. Sci.* Paris 295, Ser II, 113.
Low, R. 1990, *Class, Quant. Grav.*, **7**, 177.
Low, R.1994, *Class, Quant, Grav*., **11**, 453.
Mason, L. J., Woodhouse, N. M. J. 1996, Integrability, self-duality and twistor theory (Oxford University Press, Oxford).
McNamara, J. M. 1978a, *Proc. R. Soc. London*, **A358**, 499.
McNamara, J. M. 1978b, *Proc. R. Soc. London*, **A364**, 121
Mellor, M., Moss, I. G. 1990, *Phys. Rev.*, D41, 403.
Mellor, M., Moss, I. G. 1992, *Class. Quant. Grav.*, **9,** L43.
Oppenheimer, J. R., Snyder, H. 1939, *Phys. Rev.*, **56**, 455.
Ori, A. 1997, *Gen. Rel.Grav.*,**29**, 881.
Penrose, R. 1969, *Rivista del Nuovo Cim. Numero speciale* **1**, 252.
Penrose, R. 1972, Techniques of differential topology in relativity, CBMS Regional conf. ser. in Apl. Math., No. 7 (S.I.A.M., Philadelphia).
Penrose, R. 1973, *Ann. N. Y. Acad. Sci.*, **224**, 125.
Penrose, R. 1978, in Theoretical principles in astrophysics and relativity, eds. N. R. Liebowitz, W. H. Reid, P. O. Vandervoort (University of Chicago Press, Chicago)
Penrose, R. 1992, in General relativity: an Einstein centenary survey, (eds) S. W. Hawking and W. Isreal (Cambridge University Press, Cambridge)
Penrose, R. 1992, in Gravitation and Modern Cosmology (eds) A. Zichichi, N. de Sabbata and N. Sanchez (Plenum Press, New York).
Penrose, R.1996, in Quantum gravity: International school of cosmology and Gravitation, XIV Course, (eds) P. G. Bergmann, V. de Sabbata and H. J. Treder (World Scientific, Singapore).

Penrose, R., Rindler, W. 1996, Spinors and Space-time, vol. 2 Spinor and Twistor methods in space time geometry (Cambridge University Press, Cambridge)
Shapiro, S., Teukolsky, S. A. 1991, *Phys. Rev. Lett.*, **66***,* 994.
Schoen, R., Yau, S. T. 1993, *Commun. Math*, *Phys*., **90**, 575.
Semiz, I. 1990, *Class, Quant. Grav*., **7**,353.
Simpson, M., Penrose, R. 1973, *Int. J. Theor. Phys.*, **7**, 183.
Thorne, K. S. 1972, in Magic without Magic, ed. J. R. Klauder (Freeman, San Francisco).
Tod, K.P.1992, *Class.Quant.Grav.*, **9**, 1581.
Wald, R. M. 1974, *Ann. Phys.*, **82**, 548.
Yodzis, P., Seifert, H. J., Muller zum Hagen, H. 1973, *Commun. Math. Phys*., **34**, 135.

J. Astrophys. Astr. (1999) **20**, 249–257

New Mathematical Approaches to Classical Censorship Problems

C. J. S. Clarke, *Faculty of Mathematical Studies*, *University of Southampton*, *Southampton*, *SO17 IBJ*, *UK*

Abstract. An account is given of recent advances in mathematical techniques for extending space-times through weak singularities. This gives one more hope of proving a cosmic censorship theorem, and hence of understanding whether the final state of gravitational collapse will indeed be a black hole.

Key words. Singularities—censorship—black holes.

1. Introduction

Our current picture of black holes is governed by the implicit acceptance of the cosmic censorship hypothesis: that any singularities will be contained inside the black hole horizon (i.e. the *weak* version of the hypothesis). If this were false, and gravitational collapse could produce naked singularities, then the observational predictions could be significantly changed—although it should be noted that at present most of the observations that are considered refer to regions at some distance from the horizon, where the exact nature of the compact object at the centre is less crucial. Unfortunately, progress towards formulating a provable version of the hypothesis, and then proving it, appears to have reached an impasse. The aim of this talk is to outline methods that may resolve this problem.

The overall idea behind attempted proofs of the cosmic censorship hypothesis is usually as follows. The singularity theorems predict the occurrence of incomplete geodesics under specified physical conditions. Such incomplete geodesics should be classifiable into two groups: those ending at "weak" singularities, through which space-time can be extended, perhaps with lower differentiablity, and which therefore we need not worry about; and those ending at "strong" singularities, associated with curvature that is sufficiently great to allow one to prove the existence of a horizon. This general idea fails to be workable in practice, however, because all attempts to make precise the terms "strong" and "weak" do not produce an exhaustive classification of singularities. There is a large gap of intermediate types between those that are weak enough to allow an extension of space-time, and those which are so strong that one can prove some sort of censorship theorem. In order to overcome this, we need to enlarge the class of "weak" singularities; in other words, we need to be able to extend space-time, in a physically meaningful way, through a much wider class of singularities. Thus we need to be able to make physical sense of points in such extended space-times where the metric is quite badly behaved (so that they would be regarded as "singular" in relation, say, to the conditions of the Hawking-Penrose singularity theorems), but not so badly behaved that a horizon forms.

What then is the appropriate criterion for dividing endpoints of incomplete geodesics (putative singularities) into those which are physically meaningful as interior points in the space-time, and those which are so strong that this is not the case? I suggest that the criterion should be the breakdown of predictability. If it is possible to prescribe physical conditions to the past of the point and predict their continuation to the future, then the point does not disrupt predictability and should not be regarded as a true singularity—I shall call it *inessential*, rather than "weak". If, however, this is not so, then we do have a true singularity associated with the breakdown of classical physics.

2. General formulation

To make this clearer, let me set a context in which we can define this idea of predictability more precisely. We work entirely locally, so that the space-time can be identified with $\mathbb{R}^4$. (There is no reason why the ideas should not be extended to general manifolds equipped with a Riemannian background metric, but this seems unnecessary until we have more understanding of the local situation.) The following specification[1] will be illustrated later by reference to a simple linearised version. The metric g, with signature $(-+++)$ is not necessarily continuous (though for most of the actual results so far we cannot be that general), and has an inverse g^{ij} defined almost everywhere.

DEFINITION. A C^1 surface Σ is said to be *g-spacelike* if there is a C^1 function f and a positive constant K such that

$$\Sigma = \{x \mid f(x) = 0\} \text{ with } \nabla f(x) \neq 0 \text{ for } x \in \Sigma$$
$$\text{and } -g^{ij}f_{,i}(x)f_{,j}(x) > K \text{ for almost all } x \in \Sigma.$$

(By a surface I mean a submanifold, possibly with boundary, of codimension 1.)

There are many different types of dynamical equations used in general relativity (perfect fluid with a constitutive equation, field theories, Einstein-Vlasov and other statistical physics formulations etc). So as to be able to talk about the general structure rather than the particular details, we suppose we are dealing with a category $\mathcal{S}$ of local *solution sets*, each of whose objects $\mathcal{S}(U, \Sigma)$ is specified by an open set U in $\mathbb{R}^4$ and a surface $\Sigma \subset U$, and consists of a triple $(S_{U,\Sigma}, T_\Sigma, \mathcal{S}_{U,\Sigma})$ where $S_{U,\Sigma}$ is the set of *solutions* of some chosen differentiability of the field equations over U; T_Σ is the corresponding set of allowable (possibly constrained) *Cauchy data*; and $\mathcal{S}_{U,\Sigma} : S_{U,\Sigma} \to T_\Sigma$ assigns Cauchy data to solutions. Here $S_{U,\Sigma}$ is a subset of some Banach space B_U (or, for gauge theories, a product of a Banach and an affine-Banach space) and T_Σ is a subset of a Banach or affine-Banach space C_Σ. If $g(\phi)$ for $\phi \in B_U$ denotes the metric associated with the solution ϕ, then we require that Σ be $g(\phi)$-spacelike for all $\phi \in S_{U,\Sigma}$, and we also require there to exist a map $r_{U,\Sigma}$ taking each $\phi \in B_U$ for which Σ is $g(\phi)$-spacelike to an element of C_Σ, such that $\mathcal{S}_{U,\Sigma}$ is a restriction of $r_{U,\Sigma}$.

[1]Warning: this section contains some scenes of gratuitous abstraction which may be offensive to readers of a practical disposition.

Corresponding to this is a category $\mathcal{W}$ whose objects are $\mathcal{W}(U, \Sigma) = (B_U, C_\Sigma, r_{U,\Sigma})$. The morphisms of $\mathcal{W}$ are triples $\psi = (\psi_1 : U \to U', \psi_2 : \Sigma \to \Sigma', \psi_1^* : B_U \to B_{U'}, \psi_2^* : C_U \to C_{U'})$ with ψ_i diffeomorphisms and satisfying the obvious consistency conditions.

$$\phi_2 = \phi_1|_\Sigma, \qquad r_{U',\Sigma'} \circ \phi_1^* = \phi_2^* \circ r_{U,\Sigma}$$

(whenever either side of the second equation is defined), together with the category axioms under composition. The morphisms of $\mathcal{S}$ are then required to be the restrictions of those of $\mathcal{W}$. We also suppose that these categories are closed under finite unions and intersections of the Us and Σs in the obvious way. Finally, we denote by S_U the set of fields ϕ such that $\phi \in S_{U,\Sigma}$ for some Σ.

Suppose D is open in $\mathbb{R}^4$, Σ is a C^1 surface in D and $\phi \in S_{D,\Sigma}$. I will define an extension of the usual idea of D being the domain of dependence of Σ, by requiring continuous dependence in D on the data in Σ (including uniqueness) modulo diffeomorphisms.

DEFINITION. The triple (D, Σ, ϕ) is a *domain of continuous dependence* if, given $\epsilon > 0$ and an open D' with $\bar{D}'$ compact in U and $D' \cap \Sigma \neq \emptyset$, there exists $\delta > 0$ such that the following holds:

if $\phi' \in S_{U',\Sigma}$ has $\|s_{U,\Sigma}(\phi) - s_{U',\Sigma}(\phi')\|_{C_\Sigma} < \delta$, then there are ψ, U'', V, $\bar{\phi} \in S_{V,V\cap\Sigma}$ such that

1. $\psi : \mathcal{W}(U', \Sigma) \to \mathcal{W}(U'', \Sigma)$ has $\psi_2 =$ identity, $\psi_2^* =$ identity;
2. $D' \subset V$
3. $\bar{\phi}|_{V\cap U''} = \psi_1^* \phi'|_{V\cap U''}$
4. $\|\bar{\phi}|_{D'} - \phi|_{D'}\|_{B_{D'}} < \epsilon$.

This enables me to give the required definition of when a point is inessential (i.e. not a true singularity) as follows:

DEFINITION. A point $p \in D$ $U \subset \mathbb{R}^4$ is $\mathcal{S}$-inessential with respect to a space-time $\phi \in S_U$ if there is a pair (D, Σ) with $p \in D \subset U$ such that $(D, \Sigma, \phi|_D)$ is a domain of continuous dependence.

3. Low differentiability solutions of Einstein's equations

Under the above definition, a point is inessential if, in some neighbourhood of it, the dynamical equations have a solution which is stable and unique under perturbations of Cauchy data. We require this concept to extend to situations where the metric is not smooth. As a programme for proving cosmic censorship in full generality, this is probably idealistic. We can, however, examine particular cases of possible counter examples to cosmic censorship, in the hope of verifying that they are in fact inessential, and hence build up further evidence for the hypothesis and increase our understanding of what it involves. I shall separate the problem into the two parts of (**a**) showing that a given metric is in fact a solution of the Einstein equations in a distributional sense in a neighbourhood that includes a possible "singularity"; (**b**) investigating the stability of the solution by examining linearised perturbations. I start in this section with (a).

The obstacle to applying Einstein's equations to a low differentiability metric is that, while the metric can be differentiated in the sense of distribution theory, possibly producing δ-functions and similar distributions, in order to form the Einstein tensor we must multiply together these derivatives, and the multiplication of distributions is not a well-defined operation unless various restrictions are imposed. Without the adoption of some scheme for multiplying distributions, one can only consider metrics of the class defined by Geroch & Traschen (1987) in which metric is continuous and its distributional derivative is equivalent to a square integrable function.

I will describe here a systematic way of carrying out multiplication, due to Colombeau, that enables us to study a wider class of metrics, including those associated with cosmic strings. (Full details in a relativistic setting, together with references to work by other authors, are described in Clarke *et al.* (1996).) The essential idea is that one first smoothes the functions under consideration (in our case, the components of the metric tensor), then multiplies the smoothed functions, and finally extracts an equivalent final distribution (where this can be done in a consistent manner).

Suppose Φ is a member of the space $D(\mathbb{R}^n)$ of test functions: smooth (i.e. $C^{\infty)}$ $\mathbb{R}$-valued functions on $\mathbb{R}^n$ with compact support; and that

$$\int \Phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1.$$

Given $\epsilon > 0$, we define

$$\Phi^{\epsilon}(\boldsymbol{x}) = \frac{1}{\epsilon^n}\Phi\left(\frac{\boldsymbol{x}}{\epsilon}\right)$$

so that Φ^{ϵ} has a support scaled by ϵ and an amplitude adjusted so that its integral is still unity. If $f : \mathbb{R}^n \to \mathbb{R}$ is a function, not necessarily continuous, then by a smoothing of f we mean the convolution

$$\tilde{f}_{\epsilon}(\boldsymbol{x}) := \int f(\boldsymbol{y}+\boldsymbol{x})\Phi^{\epsilon}(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \int f(\boldsymbol{z})\Phi^{\epsilon}(\boldsymbol{z}-\boldsymbol{x})\mathrm{d}\boldsymbol{z}.$$

(Smoothed functions thus depend implicitly on Φ.)

Smoothing is defined in the same way for distributions, but with some notational changes, A distribution R is regarded as a $\mathbb{R}$-valued functional.

$$\mathcal{D}(\mathbb{R}^n) \ni \phi \mapsto (R, \phi) \in \mathbb{R}$$

On the space $\mathcal{D}(\mathbb{R}^n)$ of test functions, and the convolution is defined by

$$\tilde{R}_{\epsilon}(\boldsymbol{x}) = (R, \Phi^{\epsilon}(. - \boldsymbol{x})).$$

An intuitively plausible procedure for defining the product RS of two distributions R and S would then be to define the action of the product on a test function ψ by first defining the corresponding action of the product of the smoothed quantities $\tilde{R}_{\varepsilon}$ and $\tilde{S}_{\varepsilon}$, and then taking the limit as the smoothing is made progressively finer, with $\varepsilon \to 0$:

$$(RS, \psi) = \lim_{\epsilon \to 0} \int \tilde{R}_{\epsilon}(\boldsymbol{x})\tilde{S}_{\epsilon}(\boldsymbol{x})\psi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

For example, if $R = \delta$, the Dirac δ-function, and S is the Heaviside function, then this prescription yields the attractive solution $\delta S = \delta/2$, *provided* that $\Phi(-x) = \Phi(x)$.

If one considers products involving more complicated distributions, then the dependence of the answer on the nature of Φ becomes more detailed. For example, writing x^{-1} for the distribution defined by taking the Cauchy principal value in integrals, $x^{-1}\delta = k\,\delta'$ where k is a Φ-dependent constant.

Colombeau's definition of generalised functions does not get rid of the ambiguity arising from this sort of Φ-dependence, but it enables one to keep track of possible ambiguities in a systematic and consistent context. It starts from a space $\varepsilon_M(\mathbb{R}^n)$of functions depending on both the position x and a smoothing kernel Φ, on which are imposed conditions that reflect the special case of the smoothing of a distribution, but which are more general. Multiplication is defined on these pointwise. This produces a space with some of the required properties, but in which multiplication does not coincide with the ordinary multiplication even for C^∞ functions. This is rectified by defining an equivalence relation on $\varepsilon_M(\mathbb{R}^n)$ and passing to a space $\mathcal{G}$ of equivalence classes (see Clarke *et al*. (1996) for details of this rather involved construction).

Once one has a way of multiplying generalised functions, there is no problem in defining the Einstein tensor; but for this to be physically meaningful we need to be able to verify that the final result can be regarded as a distribution, since the Colombeau algebra contains many elements that do not have this property.

To this end, G will be called *weakly equivalent* to a distribution K (written $G \approx K$) if, for each test function ϕ, we have

$$\lim_{\epsilon\to 0}\int_{\mathbb{R}^n} R(\Phi_\epsilon, \boldsymbol{x})\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^n} K(\boldsymbol{x})\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

for some (and hence for any) representative R of G and for any Φ for which the moments of order 2,..., N are zero for some sufficiently large N (depending possibly on the choice of R).

A metric is only regarded as physically interpretable if the components of the Einstein tensor are weakly equivalent to distributions.

This method has been applied in Clarke *et al*. (1996) to the static thin cosmic string, with metric

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}r^2 + A^2r^2\mathrm{d}\phi^2 + \mathrm{d}z^2$$

and then generalised in Wilson (1997) to a time-varying string, loosing energy through radiation having the energy momentum tensor for a null fluid, with metric

$$\mathrm{d}s^2 = e^{2\gamma}(-\mathrm{d}t^2 + \mathrm{d}r^2) + r^2\mathrm{d}\phi^2 + \mathrm{d}z^2$$

where γ is an arbitrary function of $u = t - r$. In this case the energy momentum tensor is weakly equivalent to a distribution, as follows:

$$\begin{aligned}
[\tilde{T}^u{}_u\sqrt{-\tilde{g}}] &\approx T^u{}_u\sqrt{-g} - 2\pi\mathrm{e}^{\gamma(u)}(1-\mathrm{e}^{-\gamma(u)})\delta^{(2)}(x,y),\\
[\tilde{T}^z{}_z\sqrt{-\tilde{g}}] &\approx T^z{}_z\sqrt{-g} - 2\pi\mathrm{e}^{\gamma(u)}(1-\mathrm{e}^{-\gamma(u)})\delta^{(2)}(x,y),\\
[\tilde{T}^a{}_b\sqrt{-\tilde{g}}] &\approx T^a{}_b\sqrt{-g} \quad \text{(other components)}
\end{aligned}$$

where $T^\mu{}_\nu$ denotes the non-distributional energy momentum tensor of the radiation-fluid away from $r = 0$. (Moreover, it can be verified that the rate of decrease of mass by the string is properly matched by the flux of energy at future null infinity.

4. Stability of a class of singularities

I shall now describe the second way of investigating the possibility of singularities being inessential, namely by showing that at the linearised level the Cauchy problem is well defined in their neighbourhood.

Linearising about a background solution go and imposing gauge conditions produces, in general, a system of linear hyperbolic equations of the general form

$$L_m(g_0)\phi_m = 0, \qquad L_g(g_0)h = T(\phi_m)$$

where the ϕ_m are matter fields, h is the linearised metric and the Ls are first or second order differential operators. At this linearised level, a singularity p is regarded as inessential if these equations have a well-defined Cauchy problem in a neighbourhood of p. It seems likely that this depends only on the principal part of the system, and so the basic features of the approach can be presented by simplifying the system to the wave equation

$$\Box_{g_0}\phi = 0. \tag{1}$$

We are examining, therefore, the condition that a singularity must satisfy for it not to interupt the predictable evolution of a solution to this equation; we regard this as a *necessary* condition for the singularity to be inessential.

Working locally in a region U which I identify with a subset of $\mathbb{R}^4$, we take a surface S to the past of the singularity p (regarded as a point in $U \subset \mathbb{R}^4$ at which the metric is not smooth) on which Cauchy data are to be posed, and we examine the existence and uniqueness of solutions to (1) in U to the future of S. An immediate problem is, how is (1), which involves the derivatives of the metric g, to be interpreted in the presence of points where these derivatives do not exist? The standard approach to this is to interpret (1) "distributionally" (or "weakly"), by multiplying by an arbitrary test function $\psi : U \to \mathbb{R}$ with compact support in U and integrating by parts over the region U^+ to the future of S, to obtain

$$\int_{U^+} \phi_{,i}(\psi_{,j}\sqrt{-g}g^{ij})d^4x = -\int_{U^+} \psi f\sqrt{-g}d^4x - \int_S \psi\phi_1 dS \tag{2}$$

where dS is the 3-volume element defined by g on S and ϕ_1 is the value of the normal derivative of ϕ on S which is prescribed by the Cauchy data. For this to make sense (for some ϕ) we now only require that the metric coefficients be integrable.

This formulation incorporates automatically part of the Cauchy data, but it leaves a problem over the part of the Cauchy data specifying the value of ϕ itself on S. This is because, in the above integral formulation of the wave equation, ϕ is only defined up to its value on a set of measure zero, so that if ϕ is not continuous the concept of a limiting value of ϕ on S is not meaningful. We circumvent this by demanding that ϕ have first derivatives (defined almost everywhere) which are square integrable; in other words, that $\phi \in L_2^1(M)$, the closure of the space of differentiable functions with square integrable derivatives with respect to the norm

$$\|\phi\|_M^1 = \left[\int_M \left(\sum_i (\phi_{,i})^2 + \phi^2\right) d^4x\right]^{1/2}.$$

Then in this case, the trace part of the Sobolev embedding theorem (Adams 1975), Theorem 5.4 (2)) shows that, for dim $M > 2$, the restriction ofϕ to S is well defined as an element of $L_2(S)$. We require that this is given by Cauchy data.

Having set up the problem, the pattern of the argument now follows the method of energy estimates, introduced originally by Friedrichs (1954) and developed extensively since then (see, for instance, Egorov & Shubin 1992). An *energy*, in this context, is any positive definite functional involving ϕ and its first derivatives associated with a spacelike hypersurface. Friedrichs noted that (a) the existence of an energy whose growth could be estimated from the differential equation ensured that solutions were unique; and (b) uniqueness for the adjoint problem, with data prescribed on a future surface and solved towards the past, ensured existence of solutions for the original problem.

To see whether this argument works in our low-differentiability case, I adopt a modification of the choice of energy-functional used by Hawking & Ellis (1973), namely

$$E(\tau) := \int_{S_\tau} S^{ij} t_i n_j \sqrt{-g} d^3x$$

where S_T is one of a family of spacelike hypersurfaces parametrised by $_T$ (of which S is required to be a member), n is the future normal to S_τ, t_i is a timelike vector field (this plays a crucial role) and

$$S^{ij} := (g^{ik}g^{jl} - \tfrac{1}{2}g^{ij}g^{kl})\phi_{,k}\phi_{,l} - \tfrac{1}{2}g^{ij}\phi^2.$$

In the Hawking & Ellis version, t_i is replaced by n_i and their construction will break down in our case because it involves $n_{i,j}$ which, because of the non-differentiability of g, may be unbounded. It turns out, however, that even though g may have unbounded derivatives, in many important cases one can find timelike vector fields t_i for which $t_{i;j}$ is bounded. The important requirement is condition (v)(b) in the theorem which follows:

Theorem. *Suppose given* (M, g) *and a point p in M such that*

(a) g_{ij} *and* g^{ij} *are continuous.*
(b) *gij is* C^1 *in* $M \setminus J^+(p)$.
(c) *weak derivatives* $g_{ij,\,90k}$ *exist and are locally square integrable on M.*
(d) *there exist functions* $R^i{}_{jkl}$ *which, interpreted as distributions, coincide with the Riemann tensor defined distributionally from g and* $g_{ij,\,k}$.
(e) *there is a non-empty open set* $C \subset \mathbb{R}^4$, *and positive functions* $M, N : \mathbb{R}^+ \to \mathbb{R}^+$ *such that, if* γ *is a curve with* $d\gamma/ds \in C$ *for all s then*
  (i) γ *is future timelike*
  (ii) *the integrals*

$$I_\gamma(a) := \int_0^a |\Gamma^i_{jk}(\gamma(s))|^2 ds \quad \text{and} \quad J_\gamma(a) := \int_0^a |R^i{}_{jkl}(\gamma(s))| ds$$

(*where* Γ *is defined using the weak derivatives* $g_{ij,k}$) *are convergent, with*

$$I_\gamma(a) < M(a), \quad J_\gamma(a) < N(a)$$

and $M(a) \to 0$, $N(a) \to 0$ *as* $a \to 0$.

Then p is-inessential.

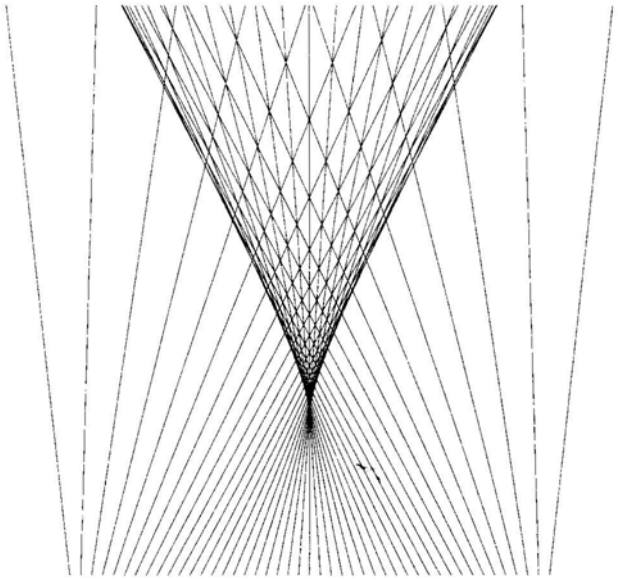

Figure 1. Space-time diagram of world lines of shell-crossing dust.

The conditions (a) and (c), introduced by Geroch & Traschen (1987), are the minimal conditions for R^i_{jkl} to be definable as a distribution by the usual coordinate formula in terms of g_{ij} and g^{ij}. The set C defines a range of timelike directions which are transverse to any shocks or caustics that may be present (this is illustrated in the next section). We refer to (e) by saying that the space-time is curve-integrable. The proof is given in (Clarke 1998). The existence of a suitable t is obtained by taking the tangent vector to a congruence of timelike geodesics, and using the geodesic deviation equation, together with condition (e), to show that this is well behaved. The argument then follows the course already described.

The sort of singularities to which this can be applied is typified by the dust caustic (shell-crossing) singularities first investigated by Yodzis *et al.* (1973, 1974), and shown in Fig. 1. Here the density rises as one approaches the caustic inversely as the square root of the space-time distance from the caustic, ensuring that the spacetime is path integrable.

5. Future prospects

Work in progress suggests that the above theorem can easily be generalised to curved cosmic strings, provided they are such that the path integrability condition holds in a coordinate system where a "wedge" is opened up, with its edge on the string, so as to compensate for the defecit angle. (In other words, one uses cartesian coordinates defined in terms of an angular coordinate whose range is the geometrical angle traversed as one circles the string.) One needs a little more than this, however, for the cases of practical interest, since for both curved cosmic strings (Clarke *et al.* 1990) and the radiating cosmic string (Wilson 1997) the curvature increases inversely as the space-time distance, so a more delicate analysis is needed.

In the course of a generic gravitational collapse with high angular momentum (where the naked Kerr singularity might be a possible end-point, if cosmic censorship were not true) we might expect the generation of very short wavelength gravitational waves, which might then focus each other in the way familiar from exact solutions. Such caustic-like focussing could well fall within the scope of the theorem that we have already applied to dust caustics, and these situations could form a bridge in our understanding, between the very simple, idealised dust caustics, and a fully general situation as needed for a cosmic censorship theorem. Much work still needs to be done, however, to traverse this gap.

References

Adams, R. A. 1975, *Sobolev Spaces*, Academic Press.
Clarke, C. J. S. 1998, *Class. Quantum Grav.* **15**, 975.
Clarke, C. J. S., Ellis G. F. R., Vickers J. A. 1990, *Class. Quantum Grav.* **7**, 1.
Clarke, C. J. S., Vickers, J. A. V. and Wilson, J. P. 1996, *Class. Quantum Grav.* **13**, 2485.
Egorov, Yu. V., Shubin, M. A. 1992, *Partial Differential Equations I*, Springer Verlag.
Friedrichs, K. O. 1954, *Comm. Pure & Applied Maths.* **7**, 345.
Geroch, R., Traschen, J. 1987, *Phys. Rev.* D**36**, 1017.
Hawking, S. W., Ellis, G. F. R. 1973, *The large scale structure of space-time*, (Cambridge University Press).
Wilson, J. P. 1997, *Classical and Quantum Gravity*, **14**, 3337.
Yodzis P., Mueller zum Hagen H., Seifert HJ. 1973, *Comm. Math. Phys.* **34**, 135; 1974, **37**, 29.

J. Astrophys. Astr. (1999) **20**, 259–268

Some Aspects of Four-Dimensional Black Hole Solutions in Gauss-Bonnet Extended String Gravity

S. O. Alexeyev* & M. V. Sazhin**, = 20 *Sternberg Astronomical Institute, Moscow State University Universitetskii Prospect*, *13*, *Moscow 119899*, *Russia*
**email: alexeyev@grg2.phys.msu.su*
***email: sazhin@sai.msu.su*

Abstract. An internal singularity of a string four-dimensional black hole with second order curvature corrections is investigated. A restriction to a minimal size of a neutral black hole is obtained in the frame of the model considered. Vacuum polarization of the surrounding space-time caused by this minimal-size black hole is also discussed.

Key words. Black holes—string theory—higher order curvature corrections.

1. Introduction

At the present time the physics of black holes contains a lot of unsolved (and even non-understood) problems. One of them is the question on the nature of the black hole inner singularities. Studying them we hope to clarify some important aspects of the Cosmic Censorship applications (Penrose 1992; Hawking & Penrose 1996; Poisson 1997; Wald 1997; Burko 1997) Moreover we can also examine the boundaries of the applicability of the General Relativity. Another interesting and completely unsolved question is: what is the endpoint of the black hole evaporation? (Hawking & Penrose 1996; Hawking & Ellis 1973; Novikov & Frolov 1986). In order to find the more comprehensive solutions of these problems it would be desirable to use the non-minimal gravity model which is the effective low energy limit of some great unification theory. That is why during the last years the four-dimensional dilatonic black holes attracts great attention because this type of black holes represents the solution of the string theory at its low energy limit (Gibbons & Maeda 1988; Mignemi & Stewart 1993; Garfincle, Horowitz & Strominger 1991; 1992; Natsuume 1994; Bento & Bertolani 1996; Kanti *et al.* 1996; Kanti & Tamvakis 1997; Torii, Yajima & Maeda 1997; Alexeyev & Pomazanov 1997a; Alexeyev 1997).

It is important to note that the string theory predicts the Einstein equations to be modified by higher order curvature corrections in the range where the curvature of space-time has the near-Planckian values. At the present time the form of the higher order curvature corrections in the string effective action is not investigated completely (Bento & Bertolani 1995). We do not know the general structure of the expansion, and, hence, the direct summing up is impossible. But as we deal with the expansion, the most important correction is the second order curvature one which is the product of the Gauss-Bonnet and dilatonic terms. It increases the order of the differential equations till the second one and the existence of the dilatonic term makes the contribution of the

Gauss-Bonnet term to be dynamic. So, the action is (for simplicity, only bosonian part is taken into account)

$$S = D\frac{1}{16\pi}\int d^4x\sqrt{-g}\left[m_{Pl}^2\left(-R + 2\partial_\mu\phi\partial^\mu\phi\right) - e^{-2\phi}F_{\mu\nu}F^{\mu\nu} + \lambda\left(e^{-2\phi} = 80S_{GB} = 80\right)\right], \quad (1)$$

$$S_{GB} = R_{ijkl}R^{ijkl} - 4R_{ij}R^{ij} + R^2.$$

Here ϕ is the dilatonic field, $F_{\mu\nu} = q \sin\theta d\theta \wedge d\varphi$ is the Maxwell term and λ is the string coupling constant.

The most careful investigation of the discussed model started only a few years ago (Gibbons & Maeda 1988; Garfincle, Horowitz & Strominger 1991; 1992; Mignemi & Stewart 1993; Natsuume 1994; Bento & Bertolani 1996; Kanti *et al.* 1996; Kanti & Tamvakis 1997; Torii, Yajima & Maeda 1997; Alexeyev & Pomazanov 1997b) and all its predictions are not investigated completely yet. It predicts a change of the solution behavior near singularities or in the regions where the influence of the higher order curvature corrections becomes strong. As the differential equations have a very complicated form, the solutions were obtained by the perturbative (Mignemi & Stewart 1993; Natsuume 1994) or numerical (Kanti *et al.* 1996; Kanti & Tamvakis 1997; Torii, Yajima & Maeda 1997; Alexeyev & Pomazanov 1997a) methods. For example, using these methods a new solution called "neutral Gauss-Bonnet black hole" was found (Kanti *et al.* 1996; Kanti & Tamvakis 1997; Torii, Yajima & Maeda 1997; Alexeyev & Pomazanov 1997a).

The main purpose of our work is to discuss external and internal black hole solutions with dilatonic hair and their main properties. This means that we are interesting in static, spherically symmetric, asymptotically flat solutions providing a regular horizon. Therefore, the most convenient choice of metric (which is usually called as the "curvature gauge") is

$$\mathrm{d}s^2 = \Delta\mathrm{d}t^2 - \frac{\sigma^2}{\Delta}\mathrm{d}r^2 - r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\varphi^2), \quad (2)$$

where $\Delta = \Delta(r)$, $\sigma = \sigma(r)$. We use this curvature gauge (and the Einstein frame) for more convenient comparison with the Schwarzschild solution.

2. Numerical results

For searching the solution in the maximal widest range of the radial coordinate it was necessary to use the most "strong" method for a numerical integration of the systems of the differential equations with particular points. This problem was solved by a modernization of the methods of integration over the additional parameter (Alexeyev & Pomazanov 1997a). This allowed us to investigate the internal structure and the particular points of the black hole using the analysis of the main determinant zeros of the linear system of the differential equations in the non-evident form. So, the system has the following matrix form

$$a_{i1}\Delta'' + a_{i2}\sigma' + a_{i3}\phi'' = b_i, \quad (3)$$

where $i = 1, 2, 3$, matrices $a_{ij} = A8\ b_i$ are:

$$
\begin{aligned}
a_{11} &= 0, \\
a_{12} &= -m_{Pl}^2\sigma^2 r + 4\mathrm{e}^{-2\phi}\lambda\phi'(\sigma^2 - 3\Delta), \\
a_{13} &= 4\mathrm{e}^{-2\phi}\lambda\sigma(\Delta - \sigma^2), \\
a_{21} &= m_{Pl}^2\sigma^3 r + 4\mathrm{e}^{-2\phi}\lambda\phi' 2\Delta\sigma, \\
a_{22} &= -m_{Pl}^2\sigma^2(\Delta' r + 2\Delta) - 4\mathrm{e}^{-2\phi}\lambda\phi' 6\Delta\Delta', \\
a_{23} &= 4\mathrm{e}^{-2\phi}\lambda 2\Delta\Delta'\sigma, \\
a_{31} &= 4\mathrm{e}^{-2\phi}\lambda\sigma(\Delta - \sigma^2), \\
a_{32} &= 2m_{Pl}^2\sigma^2\Delta r^2\phi' + 4\mathrm{e}^{-2\phi}\lambda\Delta'(-3\Delta + \sigma^2), \\
a_{33} &= -2m_{Pl}^2\sigma^2\Delta r^2\sigma, \\
b_1 &= -m_{Pl}^2\sigma^3 r^2(\phi')^2 + 4\mathrm{e}^{-2\phi}\lambda\sigma(\Delta - \sigma^2)2(\phi')^2, \\
b_2 &= -m_{Pl}^2\sigma^3(2\Delta' + 2\Delta r(\phi')^2) - \frac{1}{2}\mathrm{e}^{-2\phi}q^2\frac{\sigma}{r^3} \\
&\quad + 4\mathrm{e}^{-2\phi}\lambda 4\sigma\Delta\Delta'(\phi')^2 - 4\mathrm{e}^{-2\phi}\lambda\phi' 2(\Delta')^2\sigma, \\
b_3 &= 2m_{Pl}^2\sigma^3 r\phi'(\Delta' r + 2\Delta) - 4\mathrm{e}^{-2\phi}\lambda(\Delta')^2\sigma - 2\mathrm{e}^{-2\phi}q^2\frac{\sigma}{r^2}.
\end{aligned}
$$

Matrix (3) represents the linear system of ordinary differential equations (relatively the oldest derivatives) given in a non-evident form. Hence, according to the existence theorem the system (3) has a single solution only in the case of its main determinant to be not equal to zero. In the case of zero main determinant at some point of the solution trajectory, the uniqueness of the solution (3) will be violated. Different types of zeros of the main system determinant correspond to the different types of the particular points of the solution and only three types of zeros present in the asymptotically flat solution of equation (3). So, as the main determinant has the following structure

$$
D_{\text{main}} = \Delta[A\Delta^2 + B\Delta + C], \tag{4}
$$

where $A = A(\Delta, \Delta', \sigma, \phi, \phi')$, $B = B(\Delta, \Delta', \sigma, \phi, \phi')$ and $C = C(\Delta, \Delta', \sigma, \phi, \phi')$, zeros of the main determinant are

$$
\begin{aligned}
&\text{(a)} \quad \Delta = 0, \quad C \neq 0 \\
&\text{(b)} \quad A\Delta^2 + B\Delta + C = 0, \quad \Delta \neq 0, \quad C \neq 0 \\
&\text{(c)} \quad \Delta = 0, \quad C = 0.
\end{aligned} \tag{5}
$$

The numerical results are shown in Fig. 1. It presents a dependence of the metric functions Δ (a), σ (b) and the dilaton function $e^{-2\phi}$ (c) against the radial coordinate r. The solution exists in the definite range of the magnetic charge q: $0 \leq q \leq r_h/\sqrt{2}$, which corresponds to the charged solution in the first order (Gibbons & Maeda 1988; Garfincle, Horowitz & Strominger 1991; 1992).

The behavior of the solution outside the horizon has the usual form. Under the influence of the Gauss-Bonnet term the structure of the solution inside the horizon changes such that a new limiting value named “critical magnetic charge” appears. The structure of the inside solution is definded by this value. When $q_{cr} > q \geq r_h / \sqrt{2}$ the solution has the waiting behavior which corresponds to the first order one. When

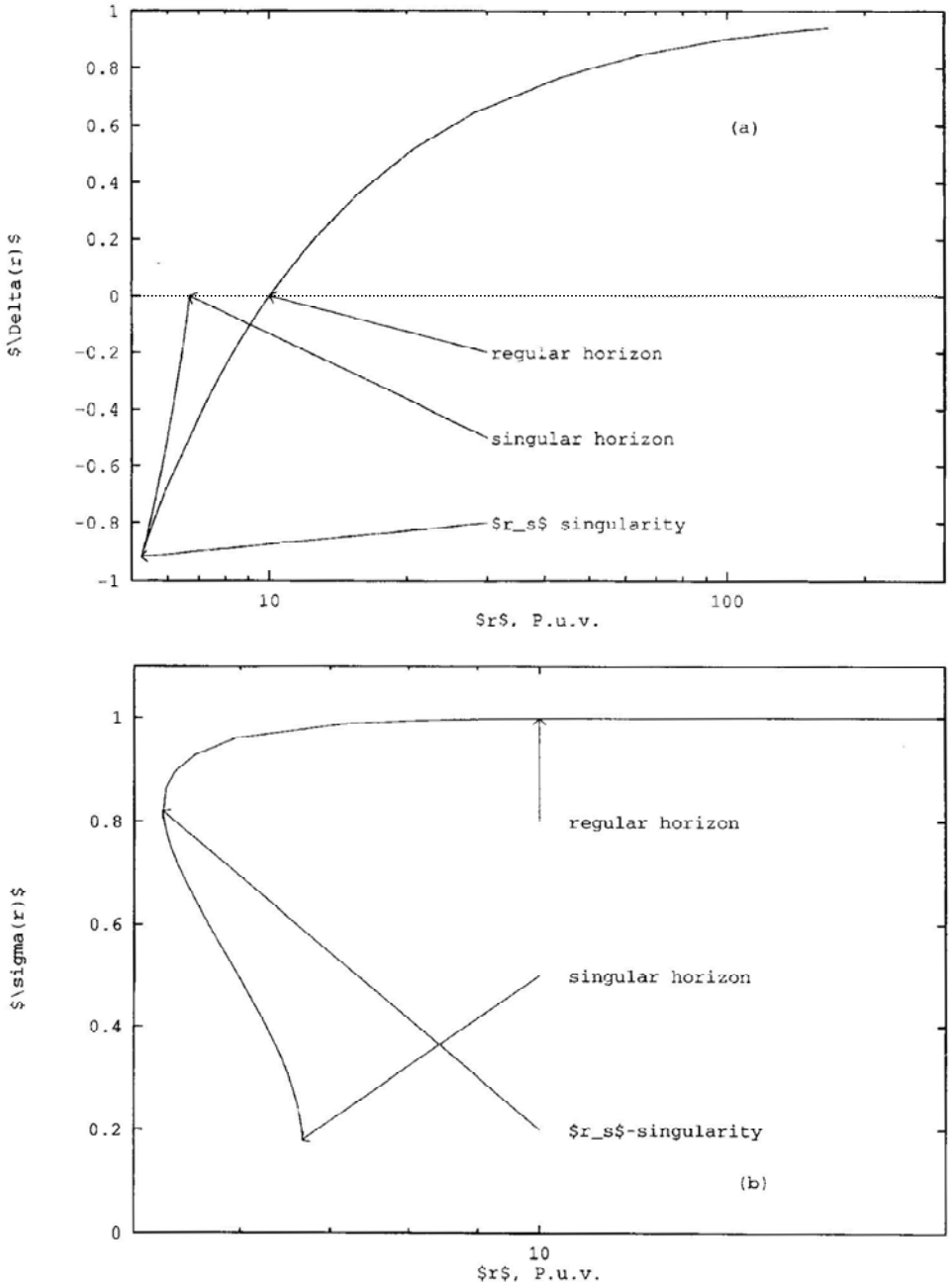

Figure 1(a–b). (*Continued*)

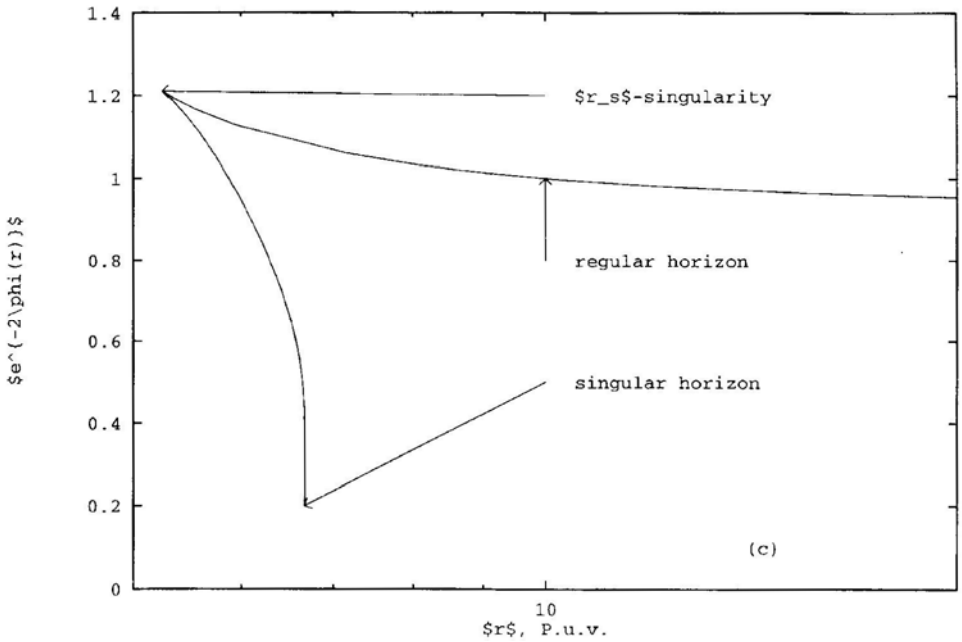

Figure 1(a, b & c). Dependence of the metric functions Δ (**a**), σ (**b**) and the dilaton function $e^{-2\phi}$ (**c**) against the radial coordinate r when $0 \le q \le q_{cr}$.

$0 \le q \le q_{cr}$ a new singular point so-called r_s appears (see Fig. 1). The type (b) of zero of the main system determinant is realized in this point. The solution turns to the other branch when reaching this point. Only two branches exist near the position r_s. There are no any branches between r_s and the origin. Therefore a new singular surface with the topology $S^2 \times R^1$ appears (it is an infinite "tube" in the t direction). Such a singularity is absent in the first order curvature gravity. This singularity exists in the different kinds of metric choice as we tested.

3. Minimal black hole

3.1 *Numerical investigations*

Figure 2 shows the graph of the metric function Δ versus the radial coordinate r at the different values of the event horizon r_h when $q = 0$. The curve (a) represents the case where r_h is rather large and is equal to 30.0 Planck unit values (P.u.v.). The curve (b) shows the changes in the behavior of Δ (r) when r_h is equal to 7.5 P.u.v. The curve (c) represents the boundary case with $r_h = r_{h\min}$ where all the particular points merge and the internal structure disappears. The curve (d) shows the case where $2M \ll r_{h\min}$ and any horizon is absent. Here it is necessary to note that for minimal and near-minimal values of r_h the metric functions can be approximated by the following formulas

$$\Delta = 1 - \frac{r_h}{r}, \qquad \sigma = 1 - \frac{s_h}{r^8}, \tag{6}$$

where $s_h = s_h(r_h)$.

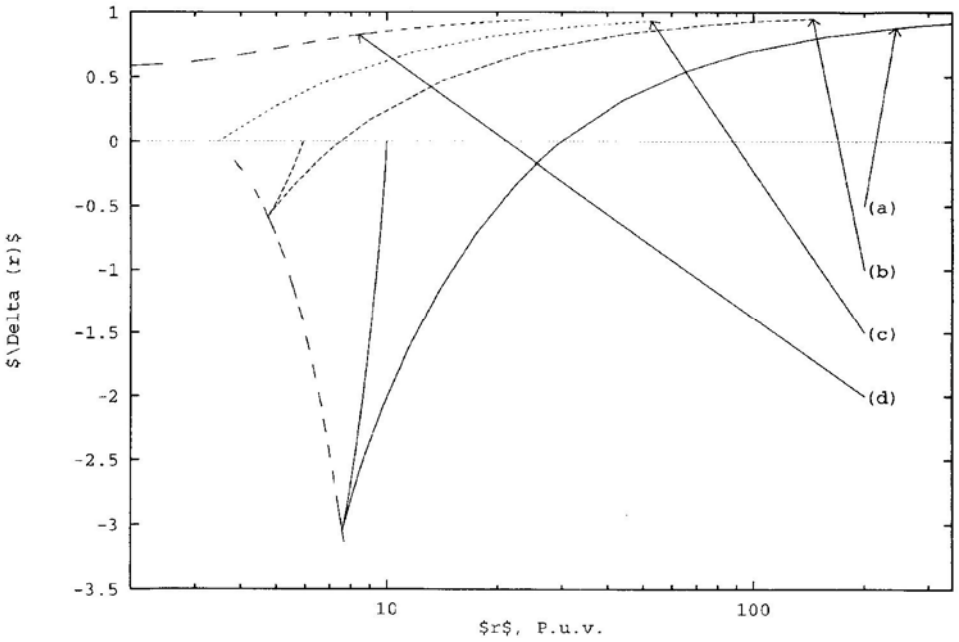

Figure 2. The dependence of the metric function Δ versus the radial coordinate r at the different values of the event horizon r_h when $q = 0$. The curve (**a**) represents the case where r_h is rather large and is equal to 30.0 Planck unit values (P.u.v.). The curve (**b**) shows the changes in the behavior of $\Delta(r)$ when r_h is equal to 7.5 P.u.v. The curve (**c**) represents the boundary case with $r_h = r_{hmin}$ where all the particular points merge and the internal structure disappears. The curve (**d**) shows the case where $2M \ll r_{h\min}$ and any horizon is absent.

3.2 *Analytical investigations*

One can also obtain the existence of a minimal dilatonic black hole from the analytical calculations. This type of a system singular point corresponds to the type (5a) of zeros of the main system determinant. Substituting the asymptotic expansions near the position r_h (which have the usual quasi Schwarzschild form) into the system (3) and after some manipulations described in (Alexeyev & Pomazanov 1997b), one obtains the following infinum value of the minimal event horizon

$$r_h^{inf} = \sqrt{\lambda}\sqrt{4\sqrt{6}}. \tag{7}$$

The analogous formula but in the other interpretation was studied earlier by Kanti *et al.* (1996). In the case $q > 0$ such restriction does not exist and regular horizon can take any meaning in the range $[0...\infty)$.

Consequently, the point $r_{h\min}$ represents the event horizon and the singularity in the same point. One should remember that λ is a combination of the fundamental string constants. That is why formula (7) can be reinterpreted as a restriction to the minimal black hole size (mass) in the given model. This restriction appears in the second order curvature gravity and is absent in the minimal Einstein-Schwarzschild gravity.

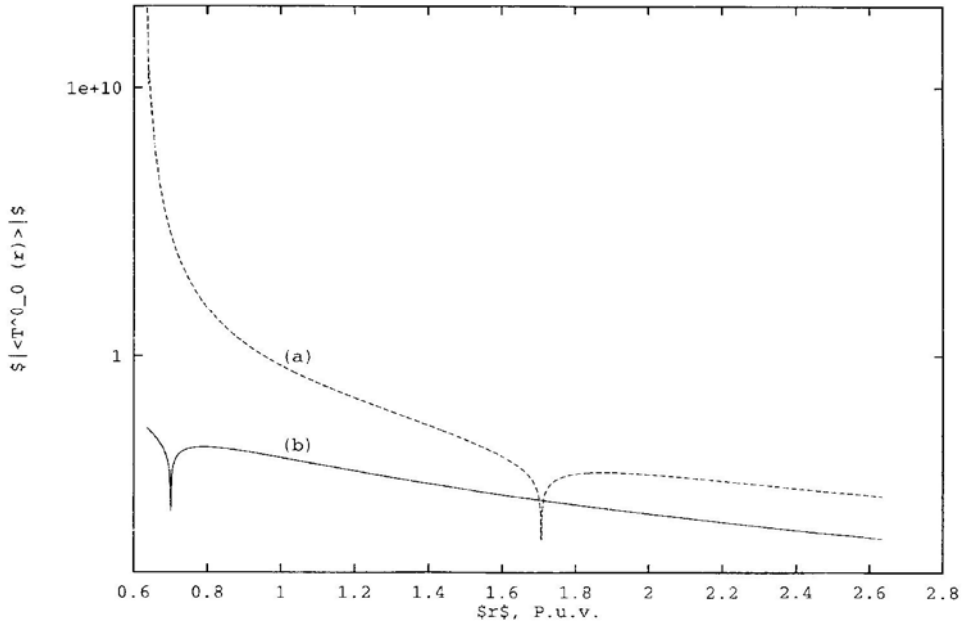

Figure 3. The dependencies of $|\langle T_0^0\rangle(r)|$ at string case with the second order curvature corrections (**a**) and at Schwarzschild (**b**) case.

3.3 *Vacuum polarization*

One of the main features of the black hole strong gravitational field is its influence to the structure of the surrounding space-time (Candelas 1980; Frolov & Zel'nikov 1984; Novikov & Frolov 1986; Anderson, Hiscock & Loranz 1995; Visser 1996a; 1996b; 1996c; 1997; Herman & Hiscock 1996; Anderson, Taylor & Hiscock 1997). In the semiclassical level this effect can be described by the vacuum polarization and stress-energy tensor expectation values. Our purpose is to compare the $\langle T_{\mu\nu}\rangle$ of the Schwarzschild black hole with the same value of the string minimal and near-minimal black hole. This work is in progress now. Nevertheless, we would like discuss some interesting preliminary results.

We analyze the contribution of the massive fields to the vacuum polarization of the string minimal black hole. In this case the contribution will be considerable enough because its mass has the order of the Planck mass $m_{pl} = \sqrt{\hbar c/G}$. Working almost near the boundaries of applicability, we can study the expansions for Hartle-Hawking vacuum average of the stress-energy tensor. Using Frolov-Zel'nikov expansions (Novikov & Frolov 1986; Frolov & Zel'nikov 1984)

$$\langle T^{(s)}_{\mu\nu}\rangle = -\frac{2}{|g|^{1/2}}\frac{\delta W^{(s)}_{ren}}{\delta g^{\mu\nu}}, \tag{8}$$

which in one-loop approximation reads (here $L^{(s)}$ is a sum of different combinations of $R^3{}_{ijkl}$ terms, see Ref. (Novikov & Frolov 1986).

$$W^{(s)}_{ren} = \frac{1}{(4\pi m)^2}\frac{1}{8\ 7!}\int d^4x|g|^{1/2}L^{(s)} + O(\epsilon^2). \tag{9}$$

In our coordinates (Δ, σ, see expansions (6)) the stress-energy tensor average values (8) have the following form (because of the great size of the formulas we show only the $\langle T_0^0 \rangle$ component in the case when the spin of a particle is equal to 1/2): $\langle T_0^0 \rangle = T_1/T_2$,

$$
\begin{aligned}
T_1 = & -8775r^{80}r_h^2 + 9117r^{79}r_h^3 + \frac{51940}{31}r^{75}r_h s_h - \frac{53900}{31}r^{74}r_h^2 s_h \\
& - 17962560r^{74}s_h + 108404640r^{73}r_h s_h - 179223493r^{72}r_h^2 s_h \\
& + 89133599r^{71}r_h^3 s_h - \frac{6468}{31}r^{68}s_h^2 + \frac{33026}{31}r^{67}r_h s_h^2 - \frac{19208}{31}r^{66}r_h^2 s_h^2 \\
& - 583036720r^{66}s_h^2 + 2813006280r^{65}r_h s_h^2 - 4012428650r^{64}r_h^2 s_h^2 \\
& + 1783041602r^{63}r_h^3 s_h^2 + \frac{32928}{31}r^{60}s_h^3 - \frac{567518}{31}r^{59}r_h s_h^3 + \frac{528220}{31}r^{58}r_h^2 s_h^3 \\
& - 2729994080r^{58}s_h^3 + 11437564864r^{57}r_h s_h^3 - 14744667622r^{56}r_h^2 s_h^3 \\
& + 6035301470r^{55}r_h^3 s_h^3 - \frac{68257}{31}r^{52}s_h^4 + \frac{900522}{31}r^{51}r_h s_h^4 - \frac{842800}{31}r^{50}r_h^2 s_h^4 \\
& - 2818841780r^{50}s_h^4 + 10549836302r^{49}r_h s_h^4 - 12479930891r^{48}r_h^2 s_h^4 \\
& + 4749439473r^{47}r_h^3 s_h^4 + \frac{74039}{31}r^{44}s_h^5 - \frac{405230}{31}r^{43}r_h s_h^5 + \frac{359660}{31}r^{42}r_h^2 s_h^5 \\
& - 540258212r^{42}s_h^5 + 1847352418r^{41}r_h s_h^5 - 2027148281r^{40}r_h^2 s_h^5 \\
& + 720283059r^{39}r_h^3 s_h^5 - \frac{46109}{31}r^{36}s_h^6 - \frac{98490}{31}r^{35}r_h s_h^6 + \frac{115640}{31}r^{34}r_h^2 s_h^6 \\
& + 3278996r^{34}s_h^6 - 4899564r^{33}r_h s_h^6 + 1843020r^{32}r_h^2 s_h^6 + \frac{18963}{31}r^{28}s_h^7 \\
& + \frac{86534}{31}r^{27}r_h s_h^7 - \frac{87612}{31}r^{26}r_h^2 s_h^7 - 1820140r^{26}s_h^7 + 3130052r^{25}r_h s_h^7 \\
& - 1456108r^{24}r_h^2 s_h^7 - \frac{7203}{31}r^{20}s_h^8 - \frac{882}{31}r^{19}r_h s_h^8 + 47124r^{18}s_h^8 + 24438r^{17}r_h s_h^8 \\
& + \frac{2597}{31}r^{12}s_h^9 + \frac{98}{31}r^{11}r_h s_h^9 - 21148r^{10}s_h^9 - 2790r^9 r_h s_h^9 - \frac{539}{31}r^4 s_h^{10} \\
& + 4796r^2 s_h^{10} + \frac{49}{31}r^{-4}s_h^{11} - 436r^{-6}s_h^{11}, \\
T_2 = & \, r^{88} - 11r^{80}s_h + 55r^{72}s_h^2 - 165r^{64}s_h^3 + 330r^{56}s_h^4 - 462r^{48}s_h^5 + 462r^{40}s_h^6 \\
& - 330r^{32}s_h^7 + 165r^{24}s_h^8 - 55r^{16}s_h^9 + 11r^8 s_h^{10} - s_h^{11}.
\end{aligned}
$$

One obtains the Schwarzschild value of $\langle T_0^0 \rangle$ by the supposition of $S_h = 0$ (Novikov & Frolov 1986).

Comparing our stress-energy tensor with the Schwarzschild one, we see that their numerical values coincide asymptotically at the infinity, but strongly differs in the neighborhood of the event horizon (for the Planckian masses the difference is about 15 orders). Therefore taking into account the terms of the string theory shows its

influence to be stronger near the event horizon where the gravitational field is not a weak one. So, we arrive at a conclusion that the application of the General Relativity gives good results at large distances from this minimal black hole, but in the neighborhood of this object one must use quantum gravity models.

4. Discussion and conclusions

From all the discussed numerical and analytical arguments it is necessary to conclude that string theory gives the new knowledge on the black hole space-times. It provides new types of the internal singularities which are absent in the same models of the General Relativity.

The most interesting consequence of the second order curvature corrections is the existence of a minimal black hole. It exists only in the neutral (quasi-Schwarzschild) case. Using Green-Witten-Schwartz formulas (Green, Schwarz & Witten 1987, Chapter 13), we can calculate the value of the string coupling constant λ, therefore, we can find the numerical value of this minimal black hole mass. It has the value about $0{\cdot}4\ m_{pl}$ (and, in that case, $S_h = 1/40$). Speculating about this phenomenon (if this object is stable) we can say that in General Relativity the most realistic models describing the black holes in our Universe are Schwarzschild one and Kerr one. Sometimes ago (Page 1976; Chambers, Hiscock & Taylor 1997) it was established that spinning black hole looses its angular momentum during the Hawking radiation. Therefore, our minimal (quasi-Schwarzschild) black hole being the endpoint of the evaporation can represent the relic remnant of the black holes formed during the initial stages of our Universe formation. This is a very interesting problem and it requires additional investigations.

Acknowledgements

S. A. would like to thank the Organizing Committee for financial support. This work was also supported by RFBR travel grant No. 97-02-27719. M.V.S. acknowledges the Center for Cosmoparticle Physics “Cosmion” (Moscow, Russia) for financial support.

References

Alexeyev, S. O. 1997, *Gravitation and Cosmology*, **3**(**4**), 161.
Alexeyev, S. O., Pomazanov, M. V. 1997a, *Phys. Rev*., **D55**, 2110.
Alexeyev, S. O., Pomazanov, M. V. 1997b, *Gravitation and Cosmology*, 3(**11**), 191.
Anderson, P. R., Hiscock, W. A., Loranz, D. J. 1995, *Phys. Rev. Lett.*, **74**, 4365.
Anderson, P. R., Taylor, B. E., Hiscock, W. A. 1997, *Phys. Rev.*, **D55**, 6116.
Bento, M. C, Bertoliami, 0. 1996, *Phys. Lett*., **B368,** 198.
Burko, L. M. 1997, *Phys. Rev. Lett*., **79**, 4958.
Candelas, P. 1980, *Phys. Rev*., **D21**, 2185.
Chambers, C. M., Hiscock, W. A., Taylor, B. E. 1997, *Phys. Rev. Lett*., **78**, 3249.
Frolov, V. P., Zel’nikov, A. I. 1984, Proceedings, Quantum Gravity (Moscow).
Garfincle, D., Horowitz, G., Strominger, A. 1991, *Phys. Rev*., **D43**, 3140.
Garfincle, D., Horowitz, G., Strominger, A., 1992, *Phys. Rev*., **D45**, 3888.
Gibbons, G. W., Maeda, K. 1988, *Nucl. Phys*., **B298**, 741.

Green, M. B., Schwarz, J. H., Witten, E. 1987, "*Superstring theory*", Cambridge University Press.
Hawking, S. W., Ellis, G. F. R. 1973, "*Large Scale Structure of Space Time*", Cambridge University Press.
Hawking, S. W., Penrose, R., 1996, (*The Isaak Newton Institute series of lectures*), (Princeton University Press), 141.
Herman, R., Hiscock, W. A. 1996, *Phys. Rev*., **D53**, 3285.
Kanti, P., Mavromatos, N. E., Rizos, J., Tamvakis, K., Winstanley, E. 1996, *Phys. Rev*, **D54**, 5049.
Kanti, P., Tamvakis, K. 1997, *Phys. Lett*., **B392**, 30.
Mignemi, S., Stewart, N. R. 1993, *Phys. Rev*., **D47**, 5259.
Natsuume, M., 1994, *Phys. Rev*., **D50**, 3945.
Novikov, I. D., Frolov, V. P. 1986, "*Physics of black holes*", Moscow, Nauka (in Russian).
Page D. N. 1976, *Phys. Rev*., **D14**, 3260.
Penrose, R. 1992, *Trieste 1992 Proceedings*, 314.
Poisson, E., To be published in the Proceedings of Workshop on the Internal Structure of Black Holes and Space-Time Singularities, Haifa, Israel, 29 Jun – 3 Jul 1997; Report No. gr-qc/9709022.
Torii, T., Yajima, H., Maeda, K. 1997, *Phys. Rev*., **D55**, 739.
Visser, M.1996, *Phys. Rev*.,**D54**, 5103; 1996, *ibid*., **D54**, 5116; 1996, *ibid*., **D54**, 5123; 1997, *ibid*., **D56**, 926.
Wald, R. M. 1997, To appear in 'The Black Hole Trail', ed. by B. Iyer; Report No. gr-qc/9710068.

J. Astrophys. Astr. (1999) **20**, 269–280

Black Hole Dynamics: A Survey of Black Hole Physics from the Point of View of Perturbation Theory

Nils Andersson, *Department of Mathematics, University of Southampton, Southampton SO171 IBJ, UK*

Abstract. This article is a brief survey of the contribution of perturbative studies to our understanding of black hole physics. For natural reasons, I will not be able to discuss all details required for an exhaustive understanding of a field that has been active for the last forty years. Neither will I be able to cover all problem areas where perturbation theory has been applied. My aim is simply to provide the interested reader with a few pointers that can serve as useful starting points for an odyssey through the literature.

1. A brief history of . . .

Speculations about black holes – objects that are so compact that not even light can escape their gravitational pull – date back to the late 1780s, when John Michell was the first to combine Newtonian gravity with the corpuscular theory of light and suggest that massive stars may be "invisible". The idea soon went out of fashion as the wave-theory of light ascended, but it was brought back into the arena (albeit in a slightly different guise) with Einstein's general theory of relativity. The first black hole solution to Einstein's field equations was, of course, discovered by Schwarzschild almost immediately after the publication of Einstein's original paper. Thus it became clear that, from a theoretical point of view, black holes may exist (see Israel (1987) for a detailed account of the development of the concept of black holes).

Even though it may not have been generally appreciated at the time, several discoveries in the 1930s made a strong case for the actual existence of these exotic objects. Firstly, Chandrasekhar's 1931 proof that there is an upper limit on the mass of white dwarfs ($M \leq 1.4\mathrm{M}_{\odot}$). Secondly, Chadwick's 1932 discovery of the neutron and the subsequent idea — due to Baade and Zwicky — that entire stars made up of these particles may exist. Such neutron stars would be limited to a mass less than something like $3\mathrm{M}_{\odot}$. Finally, the seminal work on gravitational collapse by Oppenheimer and Snyder from 1939, that provided the first demonstration of how the implosion of a star forms a black hole.

Still, there were no observational evidence for the actual existence of black holes in the universe. Indeed, it was not clear how one would go about trying to observe an, in principle, invisible object. The first indications that there are black holes out there followed the launch of the UHURU satellite, and the true opening of the X-ray window, in 1970. During its three year lifetime the satellite discovered several hundred discrete X-ray sources in the sky. Among them was the first, and still one of the strongest candidates for a black hole: Cygnus X1. The lower limit on the mass of

the invisible companion in this binary system is estimated to be at least 6 $M_\odot$. Thus, it is unlikely to be a neutron star, and by implication it can only be a black hole.

Recent observations, as described in other chapters of this volume, have provided further detailed evidence for black holes. Among these, the stunning pictures of the core of M87 taken by the Hubble Space Telescope that indicate a central mass of three billion solar masses, the observation of water masers and a disklike structure in NGC 4258 and the monitoring of stars close to the centre of our own galaxy immediately come to mind. The data indicate that most galaxies harbor a gigantic black hole at the centre. In a similar way, the case for solar-mass black holes have continued to strengthen with detailed observation of low-mass X-ray binaries. Given the recent observational advances there is every reason to expect that we will soon have a much improved understanding of astrophysical black holes. And within a few years a new generation of gravitational-wave detectors should come on-line. These promise to open up a new window to the universe. This could well lead to future astronomers being able to monitor black holes more or less daily. It is a fascinating prospect.

2. Stability

The relevance of the collapse work of Oppenheimer & Snyder (1939) was not immediately realized. One can easily identify two reasons for this. The most obvious one is the break out of the second world war. For several years, the world was in deep turmoil and not focused on basic (somewhat ethereal) research. A second reason is that the scientific community was not yet prepared to accept the inevitability of gravitational collapse. For example, the true non-singular nature of the event horizon had not yet been understood.

Oppenheimer & Snyder had considered a very idealized scenario — the collapse of a dust cloud: spherical, nonspinning, with no internal pressure or shocks etcetera. In this case the result indicated that the formation of a black hole was inevitable. But what about a less idealized – more realistic – case? What happens when a non-ideal star collapses? This question prompted John Wheeler and his students to pick up the torch after it had been left smouldering for almost twenty years. Their investigations led to what Kip Thorne has branded the "golden age of black hole physics" (Thorne 1994), and the answer to the question of non-ideal collapse was succinctly summarized by Richard Price in 1972: *Whatever can be radiated is radiated!* A collapsing star always settles down to a black hole without "hair", i.e. a black hole that can be fully described in terms of its mass, electric charge and angular momentum.

But we have jumped far ahead in the story. In 1957 John Wheeler and Tullio Regge published what can be considered the first paper on black hole perturbation theory (Regge & Wheeler 1957). Their motivation was to investigate whether a black hole was stable to external perturbations or whether it "would explode if an ant sneezed in it's vicinity" (as Vishveshwara so aptly has described it (Vishveshwara 1998). To address the stability issue Regge and Wheeler derived the equations that describe a slightly deformed black hole. This is done by assuming linear perturbations

$$g_{\mu\nu} = g_{\mu\nu}^{\text{background}} + h_{\mu\nu}, \qquad \text{where } |h_{\mu\nu}| \ll 1. \tag{1}$$

The perturbations of a Schwarzschild black hole can then be divided into two classes. The first induces inertial frame-dragging (rotation) and is often referred to as axial (or odd parity). The second class corresponds to perturbations that remain unchanged after sign of φ. These are called polar (or even parity) perturbations. In the case of non-rotating black holes these two classes decouple and can be studied separately.

Axial perturbations are governed by what is now known as the Regge-Wheeler equation

$$\frac{\partial^2\psi}{\partial r_*^2} - \frac{\partial^2\psi}{\partial t^2} - V\psi = 0, \tag{2}$$

where

$$V = \left(1 - \frac{2M}{r}\right)\left[\frac{l(l+1)}{r^2} - \frac{2M(1-s^2)}{r^3}\right]. \tag{3}$$

Here s is the spin-weight of the perturbing field, l is the integer index of the spherical harmonic used to describe the angular properties of the perturbation, and the tortoise coordinate is defined as

$$\frac{\partial}{\partial r_*} = \left(1 - \frac{2M}{r}\right)\frac{\partial}{\partial r}. \tag{4}$$

This translates the part of spacetime accessible to a causal observer into the range $-\infty \leq r_* \leq \infty$. (In other words, the event horizon is pushed all the way to $r_* = -\infty$ A similar equation (the Zerilli equation (Zerilli 1970)) can be derived for polar perturbations. In other words, the description of perturbed black holes involves solving a wave equation with a deceivingly simple effective potential.

Given the above differential equation we can address the main question for black hole stability: Will a perturbation of a black hole become unbounded if evolved according to the linear equations? A satisfactory answer to this question was first provided by Vishveshwara (1970a). He showed how one can derive (by multiplying (2) with its complex conjugate and integrating) the following "energy integral"

$$\int_{r_*=-\infty}^{r_*=+\infty}\left[\left|\frac{\partial\psi}{\partial t}\right|^2 + \left|\frac{\partial\psi}{\partial r_*}\right|^2 + V|\psi|^2\right]\mathrm{d}r_* = \text{constant}. \tag{5}$$

Since V is positive definite this bounds $\partial\psi/\partial t$ and excludes exponentially growing solutions. This implies that there are no unstable modes of a nonrotating black hole.

However, the simple argument above leaves a few loopholes through which an instability might sneak in. For example, perturbations that grow linearly (or slower) with t are not ruled out. Also, we have only provided a bound for integrals of'. ψ. The perturbation may still blow up in an ever narrowing spatial region. For non-rotating black holes these gaps were filled by Kay & Wald (1987), who proved that ψ) remains pointwise bounded when evolved from smooth, bounded initial data. In other words, Schwarzschild black holes are stable for initial data that has compact support on the Kruskal extension.

Studies of the stability of rotating black holes (which is the most relevant case from an astrophysical point of view) are not as straightforward. Because of the nature of

the perturbation equations (see an example later) one cannot readily derive an energy integral like equation (5). That it is possible to do this, and hence prove mode-stability of Kerr black holes, was shown by Bernard Whiting (1989) using an intricate set of coordinate transformations. A complete proof of the stability of Kerr black holes (a la Kay & Wald (1987)) is still outstanding.

3. Quasinormal modes

It was soon realized that the perturbation equation can also be used to provide information about how a black hole interacts with its environment. One can, for example, study how a black hole reacts if one were to "kick" it in some way. Work in this direction was pioneered by Vishveshwara (1970b). As he has subsequently described it (Vishveshwara 1998): The question was "*how do you observe a solitary black hole? To me the answer seemed obvious. It had to be through scattering of radiation, provided the black hole left its fingerprint on the scattered wave …. So, I started pelting the black hole with Gaussian wave packets. If the wave packet was spatially wide, the scattered one was affected very little. It was like a big wave washing over a small pebble. But when the Gaussian became sharper, maxima and minima started emerging, finally levelling off to a set pattern when the width of the Gaussian became comparable to or less than the size of the black hole. The final outcome was a very characteristic decaying mode, to be christened later as the quasinormal mode. The whole experiment was extraordinarily exciting.*"

During the 1970s the perturbation equations were used to study black holes in many dynamical situations, such as small bodies falling into (or being scattered by) a black hole (Davis *et al.* 1971), and slightly nonspherical gravitational collapse (Cunningham, Price & Moncrief 1979). It was found that the emerging radiation shows similar features in all cases (cf. Fig. 1). The initial response consists of a broad-band burst, followed by the quasinormal-mode ringing and finally, at late times, a power-law fall-off. Remarkably, the last two features are independent of the nature of the perturbing agent. They reflect the detailed nature of a black hole spacetime.

Since their serendipitous discovery, the quasinormal modes have attracted considerable attention. Still, the quasinormal-mode spectra of various black holes were not completely unveiled until rather recently (and there are still some outstanding questions). This reflects the fact that the mode-problem is not trivial. The reason for the underlying difficulty is, however, easily explained. The effective potential V is of short range, and corresponds to a single potential barrier. This means that the black hole problem is in many ways similar to one of potential scattering in quantum mechanics. The quasinormal modes are solutions to the equation that do not depend on the character of waves falling onto the black hole. In effect, they must be solutions to (2) that satisfy the causal condition of purely ingoing waves crossing the event horizon, while at the same time behaving as purely outgoing waves reaching spatial infinity (the quasinormal modes are analogous to the resonances in quantum scattering). Assuming a time-dependence $e^{-i\omega t}$, a general causal solution to (2) is prescribed by the asymptotic behaviour

$$\psi \sim \begin{cases} e^{-i\omega r_*} & \text{as } r \to 2M, \\ A_{\rm out} e^{i\omega r_*} + A_{\rm in} e^{-i\omega r_*} & \text{as } r \to +\infty. \end{cases} \tag{6}$$

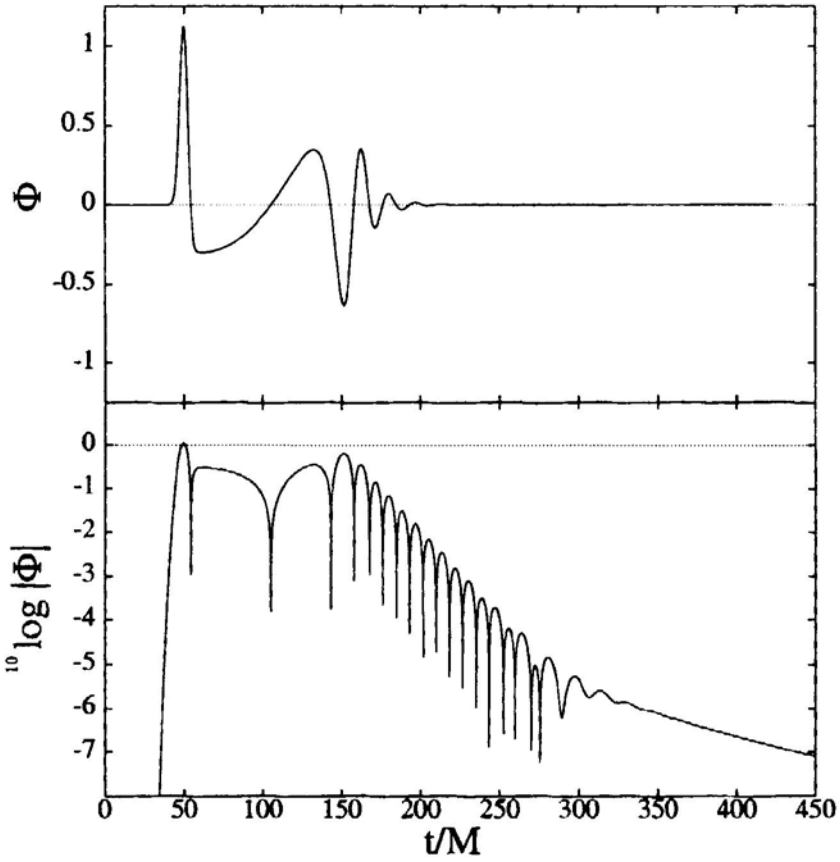

Figure 1. A recreation of Vishveshwara's scattering experiment: The response of a Schwarzschild black hole as a Gaussian wavepacket of scalar waves impinges upon it. The first bump (at $t = 50\ M$) is the initial Gaussian passing by the observer on its way towards the black hole. Quasinormal-mode ringing clearly dominates the signal after $t \approx 150\ M$. At very late times (after $t \approx 300\ M$) the signal is dominated by a power-law fall-off with time.

In this description, the quasinormal modes correspond to $A_{\rm in} = 0$. To identify a quasinormal-mode solution we must be able to determine a solution that behaves as $e^{i\omega r_*}$ as $r_* \to \infty$, with no admixture of ingoing waves. We require that the mode is damped according to an observer at a fixed location (since we have already proved that the black hole is stable — no unstable mode-solutions exist). This means that Im $\omega_n < 0$. Thus, the problem involves identifying solutions that diverge exponenttially as $r_* \to \infty$, on a constant t hypersurface. A similar problem arises at the horizon. There are by now several accurate methods for handling this difficulty and unveiling the entire spectrum of quasinormal modes (Leaver 1985; Nollert & Schmidt 1992; Andersson 1997).

Although we are not going to discuss the details of the mode-spectrum here, it is worthwhile giving a flavour of the astrophysically most important modes. First of all, the detection of a mode-signal from a black hole in anything other than gravitational waves is extremely unlikely. The reason for this is that, at the frequencies of a typical quasinormal mode an electromagnetic wave is not expected to travel far in the

interstellar medium. So we focus our attention on gravitational perturbations. For a quadrupole perturbation of a Schwarzschild black hole, the fundamental gravitational-wave quasinormal mode has frequency

$$f \approx 12\,\mathrm{kHz}\left(\frac{\mathrm{M}_\odot}{M}\right), \tag{7}$$

while the associated e-folding time is

$$\tau \approx 0.05\,\mathrm{ms}\left(\frac{M}{\mathrm{M}_\odot}\right). \tag{8}$$

The various overtones all have shorter e-folding times. Hence, the quasinormal modes of a black hole are very short lived. We can compare a black hole to other resonant systems in nature by defining a quality factor

$$Q \approx \frac{1}{2}\left|\frac{\mathrm{Re}\,\omega_n}{\mathrm{Im}\,\omega_n}\right|. \tag{9}$$

For the quasinormal modes we then find $Q \approx l$. This should be compared to the result for the fundamental fluid pulsation mode of a neutron star: $Q \sim 1000$, or the typical value for an atom: $Q \sim 10^6$. The Schwarzschild black hole is clearly a very poor oscillator in comparison to these other systems.

It is worthwhile mentioning a few outstanding problems regarding quasinormal modes. The calculation of quasinormal-mode frequencies is no longer a challenge. But there are still facets of the mode problem that require a deeper understanding. There have been a few studies of the excitation of quasinormal modes, i.e. to what extent the various modes are excited by given initial data (Leaver 1986; Sun & Price 1988; Andersson 1995), but this work needs to be extended considerably if we are to achieve a complete understanding. This is particularly important for rapidly spinning black holes, for which some quasinormal modes become almost undamped (Leaver 1985). Such modes would seem to be ideal for gravitational-wave detection (Finn 1992), but this will only be the case if they are actually excited to an appreciable level. At the present time this has not been demonstrated to be the case. Also, recent work has uncovered some peculiarities in the behaviour of the quasinormal mode frequencies as the black hole gets near to either maximal rotation or maximal electric charge (Andersson & Onozowa 1996; Onozowa 1997). Specifically, the overtone modes seem to spiral onto a limiting complex frequency as the black hole becomes extreme. It is interesting to speculate about the reason for this behaviour. It may be that this is an issue of little physical relevance, but it may also be that the underlying physics dictates this strange behaviour and that we can learn something new by studying it further (Kallosh *et al.* 1998).

4. Late-time tails

As already mentioned, the response of a black hole to any external perturbation is dominated by a power-law fall-off at very late times (see Fig. 2). This feature was first noticed by Richard Price (1972) in his studies of gravitational collapse. He deduced that a perturbation corresponding to a certain multipole l will fall off as an

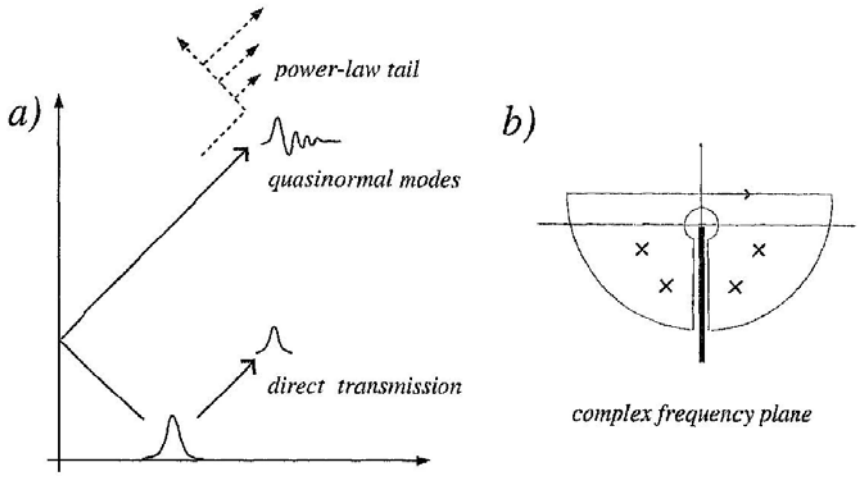

Figure 2. (a) Schematic description of a black hole's response to initial data of compact support. The directly transmitted wave (from a source point y) arrives at a distant observer (at r_*) roughly at $t - r_* + y = 0$. The black holes response, that is dominated by quasinormal mode ringing, reaches the observer at roughly $t - r_* - y = 0$. At very late times the signal falls off as an inverse power of time. This power-law tail arises because of multiple backscattering off the spacetime curvature, (**b**) Integration contours in the complex frequency plane. The original inversion contour for the Green's function lies above the real frequency axis. When analytically continued in the complex plane this contour can be replaced by the sum of 1) the quasinormal modes [the singularities of the Green's function; the first few are represented by crosses in the figure] 2) an integral along the branch cut (a thick line along the negative imaginary ω axis in the figure), that leads to the power law tail, and 3) high frequency arcs (that one would expect vanish at most times, but they should also lead to roughly "flat space propagators" at early times).

inverse power of time

$$\psi \sim t^{-2l-3} \qquad \text{as } t \to +\infty \text{ for constant } r. \tag{10}$$

This power-law tail arises because of backscattering off the slightly curved spacetime far away from the central object (Ching *et al.* 1995). The tail is generic and, in fact, independent of the existence of a horizon (Gundlach *et al.* 1994). Similar tails will arise in the spacetime exterior to a star, a black hole or an imploding/exploding shell of matter as long as the mass involved is the same. Furthermore, since the behaviour in the far zone dictates the nature of the tail, Kerr black holes must have tails similar to the Schwarzschild one (Andersson 1997).

It should also be made clear that the late-time tail is radiative: It exists both at the future horizon and at future infinity (Leaver 1986; Gundlach *et al.* 1994).

$$\psi \sim \begin{cases} u^{-l-2} & \text{as } u \to +\infty, \\ v^{-2l-3} & \text{as } v \to +\infty. \end{cases} \tag{11}$$

It is instructive to discuss the way that the tail and the quasinormal modes arise in a dynamical scenario in a little bit more detail. Suppose we are given some (perturbative) scalar field on a spacelike hypersurface $t = 0$ (say), and that we want to

predict the future behaviour of the field. That is, we want to solve

$$\Box\Phi = 0, \tag{12}$$

for a specific set of initial data. In spherical symmetry it is useful to introduce

$$\Phi_{\ell m} = \frac{u_\ell(r_*, t)}{r} Y_{\ell m}(\theta, \varphi). \tag{13}$$

The function $u_\ell(r_*,t)$ then solves the Regge-Wheeler equation (with $s = 0$), and the future evolution of a field given at some initial time ($t = 0$) follows from

$$u_\ell(r_*, t) = \int G(r_*, y, t)\partial_t u_\ell(y, 0)\mathrm{d}y + \int \partial_t G(r_*, y, t) u_\ell(y, 0)\mathrm{d}y. \tag{14}$$

Where G is the appropriate (retarded) Green's function, and $G(r_*\ , y, t) = 0$ for $t \leq 0$.

The above problem is usually analyzed in the frequency domain (after Fourier decomposition). Then we can use complex frequencies to deduce the character of the Green's function. This way we find that the quasinormal modes are the poles of the Green's function, and we can account for them by means of the residue theorem (Leaver 1986; Andersson 1997). In order to do this quantitatively, we need more information than the mode-frequency itself. In the general case we need also the exact form of the mode-eigenfunctions, and subsequently we must evaluate integrals of products of these functions. This is a truly difficult task, and even though it may lead to an accurate representation of parts of the black holes response (Leaver 1986; Sun & Price 1988), it is not very instructive. It is usually better to proceed via approximations. One such approach is based on assuming that (**i**) the observer is situated far away from the black hole, (**ii**) the initial data has considerable support only far away from the black hole, and (**iii**) the initial data has no support outside the observer. These assumptions facilitate an analytic approximation of quasinormal-mode excitation. Specifically we find that the mode-contribution to the Green's function is (Andersson 1997)

$$G^{QNM}(r_*, y, t) = \mathrm{Re}\left[\sum_{n=0}^{\infty} \frac{A_{\mathrm{out}}(\omega_n)}{\omega_n \alpha_n} e^{-i\omega_n(t-r_*-y)}\right], \tag{15}$$

$$A_{\mathrm{in}}(\omega) \approx (\omega - \omega_n)\alpha_n, \tag{16}$$

and the sum is over all modes in the fourth quadrant of the complex ω-plane. This approximate result is quite useful. First of all, we can now readily estimate the mode-excitation in situations where our underlying assumptions are relevant. We can also gain some insight in the convergence of the mode-sum at different times (Andersson 1997). For example, based on the above result it would seem natural to introduce the concept of a "dynamical" mode excitation. To ensure causality G should vanish for $t - r_* + y > 0$. This translates into a lower limit of integration $y = r_* - t$. Similarly, the contribution from the high-frequency arcs in the lower half of the ω-plane will diverge unless $t - r_* - y > 0$. This introduces an upper limit of integration $y = t - r_*$. Once these limits are used, we find that the mode-contribution converges at all times, and represents the main part of the signal very well (Andersson 1997).

In the complex-frequency picture, the late-time tail is associated with a branch cut (usually taken along the negative imaginary axis) in $G(r_*, y, t)$. Analysis of this branch cut contribution leads to (Leaver 1986; Andersson 1997)

$$G^{\rm tail}(r_*, y, t) = (-1)^{\ell+1} \frac{(2\ell+2)!}{[(2\ell+1)!!]^2} \frac{4M(r_* y)^{\ell+1}}{t^{2\ell+3}} \tag{17}$$

to leading order. But this result is only accurate for very late times. To study the regime where quasinormal ringing gives way to the power-law fall-off we must include several higher order terms.

5. The coalescence of spinning black holes

With the advent of gravitational-wave astronomy not more than a few years away, an enormous effort is focused on predictions of waveforms for what are anticipated to be the most relevant scenarios. Without accurate theoretical templates one will not be able to dig out weak signals from a typically noisy datastream, and may miss out on many interesting events. Perturbation theory is playing a relevant role in this modeling. It has for example been used to address issues related to the convergence of the post-Newtonian expansion used to describe the inspiral phase of a binary system (see Sasaki's contribution to this volume). Somewhat surprisingly, the perturbation approach has also provided relevant results for the eventual coalescence of two black holes.

As was realized by Richard Price & Jorge Pullin (1994) a few years ago, the final stages of the collision of two black holes can be approximated using perturbation theory. The idea behind what is now commonly referred to as the "close-limit" approximation is very simple. Assume that the two black holes are surrounded by a common horizon. If so, they can be considered as a single perturbed black hole. For example, the standard Misner initial data set can be viewed as representing a "Schwarzschild background + something else". In this picture, the full problem is reduced to an initial-value problem for the Zerilli equation. This problem can readily be solved, and the results compare favourably with fully nonlinear numerical relativity simulations (Anninos *et al.* 1995). Why is this, seemingly naive, approximation such a success? A reasonable explanation (corroborated by the full nonlinear calculation) is that the spacetime is only strongly distorted in the region close to the horizon. Because of the existence of the potential barrier outside the black hole most of this perturbation be scattered back onto the black hole. The waves that reach a distant observer mainly originate from the region outside the peak of the potential, where the initial perturbation is much smaller and linearized theory is a reasonable approximation.

The close-limit approach was first used to study head-on collisions of two non-rotating black holes. It has subsequently been used to investigate other cases, such as (**i**) boosted (radially or perpendicularly) holes and (**ii**) the aptly named "cosmic screw" (two colliding black holes with opposite spins) (Nollert 1996). All these cases can be viewed as perturbations of a final Schwarzschild black hole. This is, however, not the generic case since one would expect that the coalescence of two black holes forms a (rapidly) spinning black hole. Hence, one would like to be able to formulate a close-limit approximation based on perturbations of a final Kerr black hole. Several groups are presently working towards this goal. Several pieces are required before the

answer can be puzzled together. First one must formulate the general initial-value problem and translate it into "Kerr + perturbations". This is not trivial. One main difficulty is associated with the fact that for Kerr black holes in Boyer-Lindquist coordinates, the spacelike slices are not conformally flat. This means that the standard initial-value formulation fails, and one must come up with a workable alternative (Gleiser *et al.* 1998; Campanelli *et al.* 1998; Krivan & Price 1998). A second part needed for the close-limit calculation is an evolution scheme for perturbations of Kerr black holes. Such a scheme has recently been put together, and it has been tested in several different ways.

Perturbations of rotating black holes are described by an equation first derived by Saul Teukolsky (1972). For a scalar field this equation can be written (using Boyer-Lindquist coordinates)

$$\left[\frac{(r^2+a^2)^2}{\Delta}-a^2\sin^2\theta\right]\frac{\partial^2\psi}{\partial t^2}+\frac{4Mar}{\Delta}\frac{\partial^2\psi}{\partial t\partial\varphi}+\left[\frac{a^2}{\Delta}-\frac{1}{\sin^2\theta}\right]\frac{\partial^2\psi}{\partial\varphi^2} \tag{18}$$

$$-\frac{\partial}{\partial r}\left(\Delta\frac{\partial\psi}{\partial r}\right)-\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right)=0, \tag{19}$$

were $\Delta \equiv r^2 - 2Mr + a^2$. Here it should be noted that: (**i**) We can always separate the azimuthal angle in the standard way ($\psi \sim e^{im\varphi}$), but the separation of the radius and the polar angle lead to a frequency-dependent separation constant. This means that time evolutions are best done using two space dimensions, i.e. coordinates (t, r, θ). (**ii**) The reason why the "energy integral" approach to stability fails becomes clear. It fails because of the existence of the ergosphere (due to the change of sign of the coefficient of $\partial^2\psi/\partial\varphi^2$). Inside

$$r_{es} = M + \sqrt{M^2 - a^2\cos^2\theta}, \tag{20}$$

observers cannot be stationary, and energy can be negative (as viewed from infinity).

The recent codes for evolving the Teukolsky equation (for scalar and gravitational perturbations) were written by William Krivan, Pablo Laguna & Philip Papadopoulos (1996) (with the present author as some kind of partially interacting external observer). These codes have to date been used to revisit problems that had previously been approached in the frequency domain. This served to test the codes and also give a new perspective on somewhat familiar results. This way it has been verified that the Kerr late-time tail is identical to Schwarzschild, but different multipoles are mixed due to (**i**) rotational effects, and (**ii**) imperfect initial data (Krivan, Laguna & Papadopoulos 1996). It was also demonstrated that there will typically, for a given m, be two distinct regimes of mode-ringing in the Kerr case (Krivan *et al.* 1997). This happens because the quasinormal mode frequencies of a Kerr black hole are no longer symmetrically placed relative to the Imω axis as in the case of Schwarzschild. Instead, if ω is a mode corresponding to l and m the conjugate $-\bar{\omega}$ will be a mode for l and $-m$. As $a \to M$ some quasinormal modes become very slowly damped. We have verified that these modes lead to very long-lived mode ringing for rapidly rotating black holes. This is an important step towards assessing the relevance of these modes for gravitational-wave detection. Finally, we have demonstrated how a slight amplification due to superradiance can be extracted (Andersson, Laguna & Papadopoulos 1998), but also that it demands very special initial data for this effect to be observable in a

time-evolution situation. Taken together, these results show that the Teukolsky codes provide reliable tools that can be used to study many different problems in black hole physics.

6. Final words

In this short article I have tried to describe some of the ways in which perturbation theory has been used to shed light on the physics of black holes. I hope I have managed to show that one can learn a lot about black holes simply by studying their dynamics at the perturbative level. While I have mainly reviewed past efforts it seems logical to end this contribution with a look to the future. To me, it seems likely that the perturbative approach will continue to play a relevant role. First of all, fully nonlinear black hole calculations will always require benchmark tests to prove their reliability. In the appropriate limit the ideal code test is a comparison to perturbative results. There are also several ways in which perturbative methods can be used within a nonlinear calculations, e.g. to extract waves in the far zone. These are important applications, but perturbation theory is too powerful to serve as a simple slave to numerical relativity. A perturbative analysis has a considerable predictive power. If we have learned any lessons from the past – from the unveiling of the relevance of the quasinormal modes in most situations to the success of the close-limit approximation for black-hole collisions – it should be that the perturbation approach tends to be more accurate than one has any reason to expect. I would not be at all surprised if future results continue to bring home this message.

References

Andersson, N. 1992, *Proc. R. Soc. London A*, 439, 47.
Andersson, N., Laguna, P., Papadopoulos, P.1998, *Phys. Rev. D*, **58**, 087503.
Anninos, P., Price, R. H., Pullin, J., Seidel, E., Suen, W. M. 1995, *Phys. Rev. D*, **52**, 4462.
Andersson, N., Onozawa, H. 1996, *Phys. Rev.D*, **54**,7470.
Andersson, N. 1995, *Phys. Rev. D*, **51**, 353.
Andersson, N. 1997, **55**, 468.
Ching E. S. C, Leung, P. T., Suen, W. M., Young, K. 1995, *Phys. Rev. D*, **52**, 2118.
Cunningham, C. T., Price, R. H., Moncrief, V. 1979, *Astrophys. J.*, **224**, 643(1978); **230**, 870.
Davis, M., Ruffini, R., Press, W. H., Price, R. H. 1971, *Phys. Rev. Lett*, **27**, 1466.
Finn, L. S. 1992, *Phys. Rev. D*, **46**, 5236.
Gleiser, R. J., Nicasio, C. O., Price, R. H., Pullin, J. 1998, *Phys. Rev. D*, **57**, 3401; 1998;
Campanelli, M., Lousto, C. O., Baker, J., Khanna, G., Pullin, J. 1998, *Phys. Rev. D*, **58**, 084019;
Krivan, W., Price, R. H. 1998, *Phys. Rev. D*, **58**, 104003.
Gundlach, C., Price, R. H., Pullin, J. 1994, *Phys. Rev. D*, **49**, 883; **49**, 890.
Israel, W. 1987, pp. 199–276 in *300 Years of Gravitation*, (ed.) SW Hawking & W Israel (Cambridge: Cambridge University Press).
Kallosh, R., Rahmfeld, J., Wong, W. K. 1998, *Phys. Rev. D,* **57**, 1063.
Kay, B. S., Wald, R. M. 1987, *Class. Quantum. Grav*., **4**, 893.
Krivan, W, Laguna, R, Papadopoulos, P. 1996, *Phys. Rev. D*, **54**, 4728.
Krivan, W., Laguna, R, Papadopoulos, R, Andersson, N. 1997, *Phys. Rev. D*, **56**, 3395.
Leaver, E. W. 1985, *Proc. Roy. Soc. London A*, **402**, 285.
Leaver, E. W 1986, *Phys. Rev. D*, **34**, 384.
Nollert, H. R, Schmidt, B. G. 1992, *Phys. Rev. D*, **45**, 2617.

Nollert, H. P. 1996 in *Proceedings of the 18th Texas Symposium on Relativistic Astrophysics*, Chicago, USA, December 15–20.
Oppenheimer, J. R., Snyder, H. 1939, *Phys. Rev*., **56**, 455.
Onozawa, H. 1997, *Phys. Rev. D*, **55**, 3593.
Price, R. H. 1972, *Phys. Rev. D*, **5**, 2419; **5**, 2439.
Price, R. H., Pullin, J. 1994, *Phys. Rev. Lett*., **72**, 329.
Regge, T., Wheeler, J. A. 1957, *Phys. Rev*., **108**, 1063.
Sun, Y., Price, R. H., 1988, *Phys. Rev. D*, **34**, 384.
Teukolsky, S. A., 1972, *Phys. Rev. Lett*., **29**, 1114.
Thorne, K. S. 1994, *Black holes and timewarps: Einstein's outrageous legacy* (Picador, London).
Vishveshwara, C. V. 1970a, *Phys. Rev. D*, **1**, 2870.
Vishveshwara, C. V. 1970b, *Nature*, **227**, 936.
Vishveshwara, C. V. 1998, *On the blackhole trail*, The Vaidya-Raichaudhuri Endowment Lecture.
Whiting, B. F. 1989, *J. Math. Phys*., **30**, 1301.
Zerilli, F. J. 1970, *Phys. Rev. Lett*., **24**, 737 (1970); *Phys. Rev. D*, **2**, 2141.

J. Astrophys. Astr. (1999) **20**, 281–289

Analytic Black Hole Perturbation Approach

Misao Sasaki, *Department of Earth and Space Science, Osaka University, Toyonaka 560 0043, Japan*
email: misao@vega.ess.sci.osaka-u.ac.jp

Abstract. In this talk, I review an analytic method for calculating gravitational radiation from a small mass particle orbiting a massive black hole. This method allows a systematic evalutation of the gravitational radiation to a very high order in post-Newtonian expansion, hence gives us useful information on the evolution of coalescing compact binary stars.

Key words. Gravitational radiation — black hole — post-Newtonian expansion.

1. Introduction

Gravitational radiation from coalescing compact binaries has become a subject of great interest because it is the most promising target of the future gravitational wave detectors such as LIGO (Abramovici *et al.* 1992; Thorne 1994)/VIRGO (Bradaschia *et al.* 1990). During the inspiral stage of a binary, the gravitational radiation has a characteristic waveform, called a "chirp" signal, with both the frequency and amplitude increasing rapidly until the final coalescing stage begins (Thorne 1994). Since how the chirp signal develops in time depends on the rate of gravitational radiation, and the rate depends on orbital parameters of a binary, it brings us rich information about mass, spin and other physical quantities of the binary stars. In addition, the inspiralling binary may become an accurate distance indicator to be used in cosmology because the detected amplitude of gravitational waves will give the accurate distance to the source. Furthermore, provided we have an accurate theoretical prediction of the evolutionary behavior of inspiralling binaries based on general relativity, the observed data can be used to test the validity of general relativity or constrain alternative theories of gravity. Thus, much effort has been recently made to construct accurate theoretical templates (Will 1994).

To construct theoretical templates, the post-Newtonian approximations are usually employed to solve the Einstein equations. However, since the calculation is very much involved as one proceeds to a high PN order, it is most desired to have another method to examine the result. In this respect, although its validity is limited to the case of $\mu \ll M$, where M is the total mass of the binary and μ is its reduced mass, the black hole perturbation method can give a high PN order luminosity formula in a relatively simple way, hence can give a useful cross-check of the standard PN calculations. In fact, very recently, Blanchet, Iyer and Joguet have succeeded in calculating the gravitational wave luminosity from a binary in quasi-circular orbit to 3.5PN order, i.e, $O(\upsilon^7)$ beyond Newtonian quadrupole formula where υ is the orbital velocity of the binary (Blanchet, Iyer & Joguet 1997). In their work, the result from

the black hole perturbation approach has served as a useful guideline. Furthermore, if the binary actually consists of a black hole and a compact star, the black hole perturbation approach will be directly applicable.

Now, let us briefly review the previous work on the analytic black hole perturbation approach. Poisson (1993) first developed such a method and calculated the luminosity to 2PN order from a particle in circular orbits around a non-rotating black hole. Then a more systematic method was developed by Sasaki (1994) and the 4PN order waveforms and luminosity were derived by Tagoshi & Sasaki (1994). Subsequently Tanaka, Tagoshi & Sasaki (1996) calculated the luminosity to 5.5PN order.

The method was extended to the case of a rotating black hole by Shibata *et al.* (1994) and the energy and angular momentum luminosities to 2.5PN order from a particle in circular orbit with small inclination angle were obtained. For slightly eccentric orbits around a Kerr black hole, Tagoshi did the calculation to 2.5PN order (Tagoshi 1995). For circular orbits around a Kerr black hole, the calculation to 4PN order was done by Tagoshi *et al.* (1996). An extention of the method to the case of a spinning particle was done by Tanaka *et al.* (1996) and the luminosity to 2.5PN order was obtained for circular orbits, which includes the effect of spin-spin coupling.

In the next section, I review the black hole perturbation approach based on the Teukolsky equations (Teukolsky 1973). Then I summarize the recent results and discuss some implications. We use the units, $c = G = 1$.

1.1 *Teukolsky formalism*

Let us consider the case when a particle of mass μ is in a circular orbit around a Kerr black hole of mass $M \gg \mu$. The gravitational waves radiated out to infinity from the system is then described by the fourth Newman-Penrose quantity ψ_4 (Teukolsky 1973) which is related to the two independent modes of gravitational waves h_+ and $h_\times$ at infinity as

$$\psi_4 = \tfrac{1}{2}(\ddot{h}_+ - i\ddot{h}_\times). \tag{1}$$

On the Kerr background, it can be decomposed as

$$\psi_4 = \frac{1}{(r - ia\cos\theta)^4} \sum_{\ell m\omega} R_{\ell m\omega}(r)\,{}_{-2}S^{a\omega}_{\ell m}(\theta) e^{im\varphi - i\omega t}, \tag{2}$$

where ${}_{-2}S^{a\omega}_{\ell m}$ is the $s = -2$ spin-weighted spheroidal harmonic. The radial function $R_{\ell m\omega}$ satisfies the inhomogeneous Teukolsky equation (Teukolsky 1973).

$$\left[\Delta^2 \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\Delta}\frac{\mathrm{d}}{\mathrm{d}r}\right) - U(r)\right] R_{\ell m\omega}(r) = T_{\ell m\omega}(r), \tag{3}$$

where $T_{\ell m\omega}$ is the source term determined by the energy momentum tensor of the particle, and

$$\begin{aligned} \Delta &= r^2 - 2Mr + a^2 = (r - r_+)(r - r_-), \\ U(r) &= -\frac{K^2 + 4i(r - M)K}{\Delta} + 8i\omega r + \lambda, \\ K &= (r^2 + a^2)\omega - ma, \end{aligned} \tag{4}$$

where $r_{\pm}=M\pm\sqrt{M^2-a^2}$ are the radii of the inner (–) and outer (+) horizons, respectively, and λ is the eigenvalue of ${}_{-2}S^{a\omega}_{\ell m}$

To solve equation (3), we employ the Green function method. The Green function with the correct physical boundary condition is constructed from the two independent homogeneous solutions, the ingoing wave solution $R^{\rm in}$ and the upgoing wave solution $R^{\rm up}$, which have the following asymptotic behaviors:

$$R^{\rm in}\to\begin{cases}B^{\rm trans}\Delta^2 e^{-ikr^*} & \text{for } r^*\to-\infty,\\ r^3B^{ref}e^{i\omega r^*}+r^{-1}B^{inc}e^{-i\omega r^*} & \text{for } r^*\to+\infty,\end{cases}$$
$$R^{\rm up}\to\begin{cases}C^{\rm up}e^{ikr^*}+C^{ref}\Delta^2e^{-ikr^*} & \text{for } r^*\to-\infty,\\ r^3C^{\rm trans}e^{i\omega r^*} & \text{for } r^*\to+\infty,\end{cases} \tag{5}$$

where $k=\omega-ma/2Mr_+$ and r^* is the tortoise coordinate defined by $r^*=\int r^2{\rm d}r/\Delta$. Then $R_{\ell m\omega}$ is given by

$$R_{\ell m\omega}(r)=\int_{r_+}^{\infty}{\rm d}r'G^{ret}(r,r')T_{\ell m\omega}(r'), \tag{6}$$

where

$$G^{ret}(r,r')=\frac{\theta(r-r')R^{\rm up}(r)R^{\rm in}(r')+(r\leftrightarrow r')}{2i\omega\Delta^2(r')W};$$
$$W=\left(\frac{1}{\Delta}\frac{\rm d}{{\rm d}r}R^{\rm up}\right)R^{\rm in}-\left(\frac{1}{\Delta}\frac{\rm d}{{\rm d}r}R^{\rm in}\right)R^{\rm up}. \tag{7}$$

Below we focus on the radiation emitted to infinity. $R_{\ell m\omega}$ at $r\to\infty$ takes the form,

$$R_{\ell m\omega}(r\to\infty)=\frac{r^3e^{i\omega r^*}}{2i\omega B^{inc}}\int_{r_+}^{\infty}{\rm d}rR^{\rm in}(r)T_{\ell m\omega}(r)\Delta^{-2}$$
$$\equiv r^3e^{i\omega r^*}\tilde{Z}_{\ell m\omega}. \tag{8}$$

Since the geodesic equation on the Kerr background is integrable, the source term $T_{\ell m\omega}$ is analytically known. Hence our task is to find a method to compute $R^{\rm in}$.

In the case of a circular orbit with radius $r=r_0$, $T_{\ell m\omega}(r)$ takes the form,

$$T_{\ell m\omega}\sim[a_0\delta(r-r_0)+a_1\delta'(r-r_0)+a_2\delta''(r-r_0)]\delta(\omega-m\Omega), \tag{9}$$

where Ω is the orbital angular frequency. Hence what we need to know are the behavior of $R^{\rm in}$ around $r=r_0$ and its incident amplitude B^{inc}. Also because of equation (9), the amplitude $\tilde{Z}_{\ell m\omega}$ takes the form,

$$\tilde{Z}_{\ell m\omega}=Z_{\ell m}\delta(\omega-m\Omega). \tag{10}$$

In terms of $Z_{\ell m}$, the gravitational wave form at infinity is given by

$$h_+-ih_\times=-\frac{2}{r}\sum_{\ell m}\frac{1}{\omega^2}Z_{\ell m}\,{}_{-2}Y_{\ell m}(\theta,\varphi)e^{-i\omega(t-r^*)}, \tag{11}$$

and the luminosity is given by

$$\frac{dE}{dt} = \sum_{\ell=2}^{\infty} \sum_{m=1}^{\ell} |\, Z_{\ell m} \,|^2 / 2\pi\omega^2, \tag{12}$$

where $\omega = m\Omega$.

Now we consider the post-Newtonian expansion of the orbital motion by assuming $\upsilon := r_0\,\Omega \ll 1$. Since the radius of the orbit and the angular velocity are related as

$$\frac{M}{r_0} \sim (r_0\Omega)^2 = v^2, \tag{13}$$

we have the small non-dimensional parameters, $r_0, \omega = O(\upsilon)$ and $M\omega = O(\upsilon^3)$. We note that the parameter $r_0\omega$ represents the slowness of the particle motion, hence plays a role of the post-Newtonian expansion parameter, while the parameter $M\omega$ represents the strength of the gravity, hence plays a role of the post-Minkowskian parameter. In the present case, since the particle is in bound orbits, these parameters are related to each other as shown above through the orbital velocity of the particle.

Thus our task reduces to calculating the ingoing-wave Teukolsky function $R^{\rm in}$ at $r\omega = O(\upsilon) \ll 1$ as well as to extracting out its incident amplitude B^{inc} to a required order of $M\omega = O(\upsilon^3)$. Recently, Mano, Suzuki & Takasugi have found a powerful method to compute $R^{\rm in}$ (Mano, Suzuki & Takasugi 1996) which may be useful for numerical analyses as well. However, here we describe a method previously developed by Shibata *et al.* (1994). An advantage of the latter method is that the procedure of the post-Newtonian and post-Minkowski expansions can be more clearly seen.

In this method, we first transform the Teukolsky equation to a Regge-Wheeler type equation (Sasaki & Nakamura 1982a, b):

$$R \to X = (r^2 + a^2)^{1/2} r^2 J_- J_- \left[\frac{1}{r^2} R\right], \tag{14}$$

where $J_- = d/dr - iK/\Delta$. Then we find X satisfies a Regge-Wheeler type equation which reduces to the radial part of the d' Alembertian operator in the flat space limit ($M \to 0$). Consequently, the $M\omega \to 0$ limit of the ingoing wave solution $X^{\rm in}$ becomes $\omega r\, j_\ell(\omega r)$. Corresponding to equation (5), we have the asymptotic forms of $X^{in}_{\ell\omega}$ as

$$X^{in}_{\ell\omega}(r) = \begin{cases} A^{\rm trans} e^{-ikr^*}, & r^* \to -\infty \\ A^{ref}_{\ell\omega} e^{i\omega r^*} + A^{inc}_{\ell\omega} e^{-i\omega r^*}, & r^* \to +\infty, \end{cases} \tag{15}$$

where A^{inc} is related to B^{inc} as

$$B^{inc} = -\frac{1}{\omega^2} A^{inc}. \tag{16}$$

An important step of the method is to separate out the "ingoing wave phase" from the function $X^{\rm in}$;

$$\begin{aligned} X^{\rm in} &= \omega\sqrt{r^2 + a^2}\,\xi_{\ell m}(\omega r) \exp[-i\phi(\omega r)]; \\ \phi(\omega r) &= \int dr \left(\frac{K}{\Delta} - \omega\right) \to \begin{cases} kr^*, & r^* \to -\infty \\ \omega(r^* - r), & r^* \to \infty. \end{cases} \end{aligned} \tag{17}$$

Then the ingoing wave boundary condition at horizon is guaranteed by the regularity of the function $\xi_{\ell m}$ at $r \to r_+$

The next step is to rewrite the equation in terms of the non-dimensional variable $z = r\omega$ and the parameter $\epsilon = 2M\omega$, and expand it in powers of ϵ. The result takes the form,

$$\left[\frac{\mathrm{d}}{\mathrm{d}z^2} + \frac{2}{z}\frac{\mathrm{d}}{\mathrm{d}z} + \left(1 - \frac{\ell(\ell+1)}{z^2}\right)\right]\xi_{\ell m} = \epsilon Q_1[\xi_{\ell m}] + \epsilon^2 Q_2[\xi_{\ell m}] + \cdots. \tag{18}$$

Expanding $\xi_{\ell m}$ with respect to ϵ as

$$\xi_{\ell m}(z) = \sum_{n=0}^{\infty} \epsilon^n \xi_{\ell m}^{(n)}(z), \tag{19}$$

and inserting this to equation (19), we obtain the recursive equations for $\xi_{\ell m}^{(n)}$ which can be solved iteratively (see Mino *et al.* (1997) for the method of iteration).

Since the resulting expressions for $\xi_{\ell m}^{(n)}$ are quite complicated, as example we only give $\xi_{\ell m}^{(1)}$ here:

$$\begin{aligned}\xi_{\ell m}^{(1)} &= \frac{(\ell-1)(\ell+3)}{2(\ell+1)(2\ell+1)} j_{\ell+1} - \left(\frac{\ell^2-4}{2\ell(2\ell+1)} + \frac{2\ell-1}{\ell(\ell-1)}\right) j_{\ell-1} \\ &\quad + R_{\ell,0} j_0 + \sum_{m=1}^{\ell-2}\left(\frac{1}{m} + \frac{1}{m+1}\right) R_{\ell,m} j_m - 2D_\ell^{nj} + i j_\ell \ln z \\ &\quad + \frac{imq}{2}\left(\frac{\ell^2+4}{\ell^2(2\ell+1)}\right) j_{\ell-1} + \frac{imq}{2}\left(\frac{(\ell+1)^2+4}{(\ell+1)^2(2\ell+1)}\right) j_{\ell+1}.\end{aligned} \tag{20}$$

In the above expression, D_ℓ^{nj} is given by

$$D_\ell^{nj} = \tfrac{1}{2}[j_\ell \mathrm{Si}(2z) - n_\ell(\mathrm{Ci}(2z) - \gamma - \ln 2z)], \tag{21}$$

where $\gamma = 0.5772\ldots$ is the Euler constant, $\mathrm{Ci}(x) = -\int_x^\infty \mathrm{dt}\cos t/t$ and $\mathrm{Si}(x) = \int_0^x dt \sin t/t$, and $R_{m,k}$ is a polynomial of the inverse power of z defined by

$$\begin{aligned}R_{m,k} &= z^2(n_m j_k - j_m n_k) \\ &= -\sum_{r=0}^{[(m-k-1)/2]} (-1)^r \frac{(m-k-1-r)!\Gamma(m+\frac{1}{2}-r)}{r!(m-k-1-2r)!\Gamma(k+\frac{3}{2}+r)}\left(\frac{2}{z}\right)^{m-k-1-2r},\end{aligned} \tag{22}$$

for $m > k$ and it is defined by

$$R_{m,k} = -R_{k,m}, \tag{23}$$

for $m < k$. The incident amplitude A^{inc} is then readily evaluated to $O(\epsilon)$ as

$$\begin{aligned}A^{inc} &= \tfrac{1}{2} i^{\ell+1} e^{-\epsilon \ln \epsilon}[1 + \epsilon\alpha_{\ell m}^{(1)} + \cdots]; \\ \alpha_{\ell m}^{(1)} &= -\frac{\pi}{2} + \frac{2mq}{\ell^2(\ell+1)^2} + \frac{i}{2}\left[\psi(\ell) + \psi(\ell+1) + \frac{(\ell-1)(\ell+3)}{\ell(\ell+1)}\right], \\ \psi(\ell) &= \sum_{k=1}^{\ell-1}\frac{1}{k} - \gamma, \quad q = \frac{a}{M}.\end{aligned} \tag{24}$$

2. Gravitational wave luminosity to 5.5PN order

Here we only show the final result of the luminosity to 5.5PN order for circular orbits around a Schwarzschild black hole and discuss some implications (Tanaka, Tagoshi & Sasaki 1996). The results for various other cases are summarized in a review paper (Mino *et al.* (1997)).

The orbits are assumed to be at $r = r_0$ with the angular frequency $\Omega = (M/r_0^3)^{1/2}$. The result is

$$\begin{aligned}\left\langle \frac{dE}{dt} \right\rangle = \left(\frac{dE}{dt}\right)_N \Bigg[& 1 - \frac{1247}{336} v^2 + 4\pi v^3 - \frac{44711}{9072} v^4 - \frac{8191\pi}{672} v^5 \\ & + \left(\frac{6643739519}{69854400} - \frac{1712\gamma}{105} + \frac{16\pi^2}{3} - \frac{3424 \ln 2}{105} - \frac{1712 \ln v}{105} \right) v^6 \\ & - \frac{16285\pi}{504} v^7 + \left(- \frac{323105549467}{3178375200} + \frac{232597\gamma}{4410} - \frac{1369\pi^2}{126} \right. \\ & \left. + \frac{39931 \ln 2}{294} - \frac{47385 \ln 3}{1568} + \frac{232597 \ln v}{4410} \right) v^8 \\ & + \left(\frac{265978667519\pi}{745113600} - \frac{6848\gamma\pi}{105} - \frac{13696\pi \ln 2}{105} - \frac{6848\pi \ln v}{105} \right) v^9 \\ & + \left(- \frac{2500861660823683}{2831932303200} + \frac{916628467\gamma}{7858620} - \frac{424223\pi^2}{6804} \right. \\ & \left. - \frac{83217611 \ln 2}{1122660} + \frac{47385 \ln 3}{196} + \frac{916628467 \ln v}{7858620} \right) v^{10} \\ & + \left(\frac{83993097504011\pi}{1017080064000} + \frac{177293\gamma\pi}{1176} \right. \\ & \left. + \frac{8521283\pi \ln 2}{17640} - \frac{142155\pi \ln 3}{784} + \frac{177293\pi \ln v}{1176} \right) v^{11} \Bigg], \end{aligned} \tag{25}$$

where $(dE/dt)_N$ is the Newtonian quadrupole luminosity given by

$$\left(\frac{dE}{dt}\right)_N = \frac{32\mu^2 M^3}{5 r_0^5} = \frac{32}{5} \left(\frac{\mu}{M}\right)^2 v^{10}. \tag{26}$$

To compare the above result with those obtained by the standard post-Newtonian method (Blanchet, Iyer & Joguet 1997), we note that $v = (M\Omega)^{1/3}$.

As an application of the above result, let us calculate the total cycle of gravitational waves from a coalescing binary in a laser interferometer band and evaluate the error produced by the post-Newtonian formulas. It has been suggested that whether the error in the total cycle is less than unity or not gives a useful guideline to examine the accuracy of the post-Newtonian formulas as templates (Cutler *et al* 1993) (see also Poisson (1995)).

We ignore the finite mass effect in the post-Newtonian formulas and interpret M as the total mass and μ as the reduced mass of the system. The total cycle N of gravitational waves from an inspiralling binary is calculated by using the post-Newtonian energy loss formula, $(dE/dt)_n$, and the orbital energy formula $(dE/dv)_n$ which is

Table 1. The relative difference of cycle $\Delta N^{(n)}$ for typical coalescing compact binaries. The last line shows the cycle calculated by Newtonian quadrupole formula.

n	$(1.4M_\odot,1.4M_\odot)$	$(10M_\odot,10M_\odot)$	$(1.4M_\odot,10M_\odot)$	$(1.4M_\odot,70M_\odot)$
2	356	54	216	212
3	228	60	208	296
4	11	5	15	31
5	12	7	20	53
6	11	8	22	75
7	1.2	1.0	2.6	10
8	0.12	0.14	0.3	2.2
9	0.82	0.80	1.9	8.9
10	0.09	0.08	0.20	0.87
11	0.03	0.03	0.07	0.40
$N^{(0)}$	16000	600	3578	898

truncated at n/2PN order as

$$N^{(n)} = \int_{v_f}^{v_i} dv \frac{\Omega_\varphi}{\pi} \frac{(dE/dv)_n}{|(dE/dt)_n|}, \tag{27}$$

where $\upsilon_i = (M/r_i)^{1/2}$, $\upsilon_f = (M/r_f)^{1/2}$, and r_i and r_f are the initial and final orbital separations of the binary. We define the relative difference of cycle $\Delta N^{(n)}$ as $\Delta N^{(n)} \equiv |N^{(n)} - N^{(n-1)}|$. We adopt $r_f = 6M$ and r_i is the one at which the frequency of wave is 10Hz and which is given by $r_i/M \sim 347(M_\odot/M)^{2/3}$. The results for typical binary systems are given in Table 1.

For binaries whose total mass are less than $20M_\odot$, this table suggests that we need the 3PN ~ 4PN formula to obtain accurate wave forms. Although the 4PN order has not been achieved yet in the standard post-Newtonian analysis, this results show that the post-Newtonian approximation is applicable to the inspiral phase of coalescing compact binaries.

On the other hand, the convergence for the case of neutron star—black hole binaries, whose mass is above several ten $M_\odot$, is very slow. This is because r_i/M become smaller for a larger mass black hole, and the higher relativistic correction becomes more important. From Table 1, one might think that $N^{(n)}$ converges at $n = 11$ even for $(m_1,m_2) = (1.4M_\odot,70M_\odot)$. However this is not the case. It should be noted that Table 1 shows only the relative difference between the post-Newtonian approximated cycles. If we calculate the difference between the post-Newtonian formula and the fully relativistic one, we find that the 5.5PN formula is not accurate enough for the case $(m_1, m_2) = (1.4M_\odot, 70M_\odot)$ (Tanaka, Tagoshi & Sasaki 1996).

An analysis shows that the dominant error in the massive black hole-neutron star binary case comes *not* from the PN luminosity formula but from the PN expansion of the orbital energy. For such a binary, however, the test particle energy formula (or the one with a correction of $O(\mu/M)$ at low PN orders if necessary) can be used. Since the test particle energy formula on the black hole background is exactly known and it takes full account of relativistic corrections, probably we do not have to calculate the higher PN terms any more.

To conclude this section, as far as a binary in circular orbit with negligible spin is concerned, it seems that deriving the 4PN luminosity formula that includes all the

μ/m corrections is the only remaining issue. Once this is done, it will be possible to construct sufficiently accurate theoretical templates.

3. Summary

As we have seen in the previous section, although limited by the condition $\mu/M \ll 1$, the analytic black hole perturbation approach can provide very high PN order terms of the luminosity in a relatively simple and straightforward manner. Further, since calculations can be done numerically to evaluate the 'exact luminosity', we can estimate the errors of PN formulas and test the convergence of PN expansion. Thus this approach gives us useful information for the construction of theoretical wave templates.

In all of the previous papers on the black hole perturbation approach, the orbital evolution is assumed to be determined by the balance equations for the energy E and the z-component of the angular momentum L_z of the orbit. However, for general orbits on the Kerr background or for particles with spin, the orbital evolution is not determined by E and L_z alone. In such a case, one needs an explicit formula for the radiation reaction force to evaluate the orbital evolution. There is some progress in this direction (Kennefick & Ori 1996; Mino *et al.* 1997), but no useful formula has been obtained so far. Once we succeed in deriving the radiation reaction force term, it will make the black hole perturbation approach more fruitful and powerful.

Acknowledgements

I would like to thank H. Asada, Y. Mino, T. Nakamura, E. Poisson, M. Shibata, T. Tagoshi and T. Tanaka for fruitful collaborations on which this talk is based.

References

Abramovici, A., *et al.* 1992, *Science*, **256**, 325.
Blanchet, L., Iyer, B. R., Joguet, B. 1997, Talk given at GR15 (Pune).
Bradaschia, C. *et al*., 1990, *Nucl. Instrum. &Methods*, **A289**, 518.
Cutler, C. *et al.* 1993, *Phys. Rev. Lett*., **70**, 2984.
Kennefick, D., Ori, A.1996, *Phys. Rev*., **D53**, 4319.
Mano, S., Suzuki, H., Takasugi, E. 1996, *Prog. Theor Phys*., **95** 1079.
Mino,Y., Sasaki, M., Tanaka, T. 1997, *Prog. Theor. Phys. SuppL*, No. **128**, 373.
Mino, Y, Sasaki, M., Shibata, M., Tagoshi, H., Tanaka, T.1997, *Prog. Theor. Phys. Suppl*., No. **128**, 1.
Poisson, E. 1993, *Phys. Rev.* **D47**, 1497.
Poisson, E. 1995, *Phys. Rev*.,**D52**, 5719.
Sasaki, M. 1994, *Prog. Theor. Phys.* **92**, 17.
Sasaki, M., Nakamura, T.1982a, *Prog. Theor. Phys*., **67**, 1788;
Sasaki, M., Nakamura, T. 1982b, *Phys. Lett*, **89A**, 68.
Shibata, M., Sasaki, M., Tagoshi, H., Tanaka, T. 1994, *Phys. Rev. D* **51**, 1646.
Tagoshi, H. 1995, *Prog. Theor. Phys.* **93**, 307.
Tagoshi, H., Sasaki, M. 1994, *Prog. Theor. Phys.* **92**, 745.
Tagoshi, H., Shibata, M., Tanaka, T., Sasaki, M. 1996, *Phys. Rev. D*, **54** 1439.
Tanaka, T. Mino, Y, Sasaki, M., Shibata, M., 1996, *Phys. Rev. D*, **54**, 3762.

Tanaka, T., Tagoshi, H., Sasaki, M. 1996, *Prog. Theor. Phys.* **96**, 1087.
Teukolsky, S. A. 1973, *Astrophys. J.*, **185**, 635.
Thome, K. S. in *Proceedings of the Eighth Nishinomiya-Yukawa Memorial Symposium: Relativistic Cosmology*, (ed.) M. Sasaki (Universal Academy Press, Tokyo, 1994), p. 67, and references therein.
Will, C. M., in *Proceedings of the Eighth Nishinomiya-Yukawa Memorial Symposium: Relativistic Cosmology*, (ed.) M. Sasaki (Universal Academy Press, Tokyo, 1994), p. 83, and references therein.