Half Century of Black Hole Theory: From Physicists’ Purgatory to Mathematicians’ Paradise
Brandon Carter

Citation: AIP Conference Proceedings 841, 29 (2006); doi: 10.1063/1.2218167
View online: http://dx.doi.org/10.1063/1.2218167
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/841?ver=pdfcov
Published by the AIP Publishing

Articles you may be interested in
An introduction to general relativity, black holes and gravitational waves

Relation between the Parameters of a Gravitational Lens and the Frequencies of Blackhole Quasinormal Modes

The Final Merger of Comparable Mass Binary Black Holes

Extracting Waves From Binary Black Hole Systems

Spectral features from X-ray illuminated accretion discs as diagnostics of the black hole angular momentum
Half Century of Black-Hole Theory:  
From Physicists’ Purgatory to Mathematicians’ Paradise.  

Brandon Carter  
LuTh, Observatoire Paris-Meudon, France  

Abstract. Although implicit in the discovery of the Schwarzschild solution 40 years earlier, the issues raised by the theory of what are now known as black holes were so unsettling to physicists of Einstein’s generation that the subject remained in a state of semiclandestine gestation until his demise. That turning point – just half a century after Einstein’s original foundation of relativity theory, and just half a century ago today – can be considered to mark the birth of black hole theory as a subject of systematic development by physicists of a new and less inhibited generation, whose enthusiastic investigations have revealed structures of unforeseen mathematical beauty, even though questions about the physical significance of the concomitant singularities remain controversial.  

Keywords: Black holes; Singularities  
PACS: 04.70.-s; 04.20.Dw  

INTRODUCTION: SCHWARZSCHILD’S UNWELCOME SOLUTION  

This illustrated review is intended to provide a brief overview of the emergence, during the last half century, of the theory of ordinary (macroscopic 4-dimensional) black holes, considered as a phenomenon that (unlike the time reversed phenomenon of white holes) is manifestly of astrophysical importance in the real world. The scope of this review therefore does not cover quantum aspects such as the Bekenstein-Hawking particle creation effect, which is far too weak to be significant for the macroscopic black holes that are believed to actually exist in the observable universe. Nor does it cover the interesting mathematics of higher dimensional generalisations, a subject that is (for the time being) so far from relevance to the known physical world (in which – according to the second law of thermodynamics – the distinction between past and future actually matters) that its practitioners have formed a subculture in which the senior members seem to have forgotten (and their juniors seem never to have been aware of) the distinction between black and white holes, as they have adopted a regretably misleading terminology whereby the adjective “black” is abusively applied to any brane system that is hollow – including the case of an ordinary (black or white) hole, which, to be systematic, should be classified as a (black or white) hollow zero brane of codimension 3.  

The rapid general acceptance of the reality and importance of the positrons whose existence was implied by Dirac’s 1928 theory of the electron is in striking contrast with the widespread resistance to recognition of the reality and importance of the black holes
whose existence was implied by Einstein’s 1915 theory of gravity. It is symptomatic that black holes were not even named as such until more than half a century later. The sloth with which the subject has been developed over the years is illustrated by the fact that although the simplest black hole solution was already discovered (by Schwarzschild) in 1916, the simulation in Figure 1 of the present review (80 years later) provides what seems to the first serious reply to the very easy question of what it would actually look like, all by itself, with no illumination other than that from a uniform sky background.

Much of the responsibility for the delay in the investigation of the consequences of his own theory is attributable [1] to Einstein himself. Although his work had revolutionary implications, Einstein’s instincts tended to be rather conservative. It was as a matter of necessity (to provide an adequate account first of electromagnetism and then of gravitation) rather than preference that Einstein introduced the radically new paradigms involved first in his theory of special relativity, just a hundred years ago, and then in the work on general relativity that came to fruition ten years later. When cherished prejudices were undermined by the consequences, Einstein was as much upset as any of his contemporary colleagues. It could have been said of Albert Einstein (as it was said of his illustrious and like minded contemporary, Arthur Eddington) that he was always profound, but sometimes profoundly wrong.

The most flagrant example was occasionned by Friedmann’s prescient 1922 discovery of what is now known as the “big bang” solution of the general relativity equations, which Einstein refused to accept because it conflicted with his unreasonable prejudice in favor of a cosmological scenario that would be not only homogeneous (as actually sug-
gested by subsequently available data) but also static (as commonly supposed by earlier generations) despite the incompatibility (in thermal disequilibrium) of these alternative simplifications with each other and with the obvious observational consideration (known in cosmologically minded circles as the Cheseaux-Olbers paradox) that – between the stars – the night sky is dark. Einstein’s incoherent attitude (reminiscent of the murder suspect who claimed to have an alibi as well as the excuse of having acted in self defense) lead him not only to tamper with his own gravitation equations by inclusion of the cosmological constant, but anyway to presume without checking that Friedmann’s (actually quite valid) solution of the original version must have been mathematically erroneous.

Compared with his tendency to obstruct progress in cosmology, Einstein’s conservatism was rather more excusable in the not so simple case of what are now known as black holes. It is understandable that (like Eddington) he should have been unwilling to explore the limitations on the validity of his theory that are indicated by the weird and singular – or as Thorne [1] puts it “outrageous” – features that emerge when strong field solutions of the general relativity equations are extrapolated too far into the non linear regime.

At the outset Einstein’s interest in the spherical vacuum solution of his 1915 gravitational field equations was entirely restricted to the weak field regime, far outside the “horizon” at \( r = 2m \) in the simple exact solution

\[
ds^2 = r^2 (d\theta^2 + \sin^2 \theta \, d\phi^2) + \frac{dr^2}{(1-2m/r)} - (1-2m/r)dt^2,
\]

(1)

that was obtained within a year, but that was immediately orphaned by the premature death of its discoverer, Karl Schwarzschild, after which its embarrassing physical implications were hardly taken seriously by anyone – with the notable exception of
Oppenheimer [2] – until the topic was taken up by a less inhibited generation subsequent to the death of Einstein himself, just half a century ago, at Princeton in 1955. It was only then (and there) that John Wheeler inaugurated the systematic development of the subject – for which he coined the name “black hole” theory – in a series of pioneering investigations that started [3] by addressing the crucial question of stability, while not long afterwards, on the other side of the “iron curtain” another nuclear arms veteran, Yacob Zel’dovich, initiated an independent approach [4] to the same problem (using the alternative name “frozen star” which in the end did not catch on).

**OUTCOME OF STELLAR EVOLUTION: CHANDRA’S UNWELCOME LIMIT**

The question of gravitational trapping of light had been raised in the eighteenth century by Michel and Laplace, whose critical mass \( m \approx \rho^{-1/2} \) assumed the standard mass
FIGURE 4. Conformal Projection diagrams showing firstly the combined black and white hole geometry obtained (belatedly [6] in 1960, by Wheeler in collaboration with Kruskal) as the artificial analytic (vacuum) extension of the Schwarzschild solution, and secondly a more astrophysically natural extension with homogeneous interior (found in 1939 [2] by Oppenheimer in collaboration with Snyder) in which the shaded region is the “domain of outer communications” and the unshaded region is the prototype example of a hole qualifiable as black in the strict sense.

density that is understood (on the basis of quantum theory as developped by 1930) to result in hadronic matter from balance between Fermi repulsion and electrostatic attraction which (in Planck units, with proton and electron masses $m_p \approx 10^{-19}$, $m_e \approx 10^{-22}$ gives $\rho \approx m_p n$ with $n \approx \lambda^{-3}$ for the Bohr radius $\lambda \approx e^2/m_e$ with $e^2 \approx 1/137$. However most theorists refused to face the issue of gravitational collapse even after progress in quantum theory lead to Chandrasekhar’s 1931 discovery of the maximum mass $m \approx m_p^{-2}$ for cold body – which is attained when relativistic gas pressure $P \approx n/\lambda \approx n^{4/3}$ provides the support required by virial condition $P \approx m_p^{2/3} \rho^{4/3}$.

For a lower mass $m \lesssim m_p^{-2}$, stellar evolution at finite temperature $\Theta$, with gas pressure $P \approx n\Theta$ subject to $\rho \approx \Theta^3/m_p^3 m^2$, can terminate in cold equilibrium supported by non-relativistic Fermi pressure $P \approx n^{5/3}/m_e$ giving $\rho \approx m_p^3 m^5 m^2$ for a white dwarf, or $P \approx n^{5/3}/m_p$ giving $\rho \approx m_p^8 m^2$ for neutron star, as shown in Figure 3.

However a self gravitating mass of hot gas will be radiation dominated with $P \approx \Theta^4$, whenever its mass exceeds the Chandrasekhar limit, $m \gtrsim m_p^{-2}$, so that, as first understood by Chandra’s Cambridge research director, Arthur Eddington, its condition for (thermally supported) equilibrium will be given by $\rho \approx \Theta^3/m^1/2$. What Chandra could never get Eddington to accept is that, for such a large mass, no cold equilibrium state will be available, so after exhaustion of fuel for thermonuclear burning (at $\Theta \approx e^4 m_p$ ) gravitational collapse will become inevitable.

SPHERICAL COLLAPSE PAST THE HORIZON

Eddington’s example shows how, as has described in detail by Werner Israel [5] (and in striking contrast with the open mindedness of Michel and Laplace a century and a half earlier) physicists of Einstein’s generation tried to convince themselves that nature would never allow compacification within a radius comparable to the Schwarzschild value. While Einstein lived, even after Chandrasekhar’s discovery had shown that such

a fate might often be difficult to avoid, the implications were taken seriously only by Oppenheimer and his colleagues, who showed [2] how, as shown in Figure 4, the solutions of Schwarzschild and Friedmann could be combined to provide a complete description of the collapse of a homogeneous spherical body through what is now called its event horizon all the way to a terminal singularity.

Despite the persuasion of such experienced physicists as Wheeler and Zel’ dovich, and the mathematical progress due to younger geometers such as Robert Boyer and particularly Roger Penrose, the astrophysical relevance of the region near and within the horizon continued to be widely disbelieved until (and even after) the 1967 discovery [7] by Israel of the uniqueness of the Schwarzschild geometry as a static solution: many people (for a while including Israel himself [5]) still supposed (wrongly) that the horizon was an unstable artefact of exact spherical symmetry. It is therefore not surprising that the question of what such a black hole would actually look like was not addressed until much more recently, particularly considering that nothing would be seen at all without some source of illumination.

The realisation that many spectacular astrophysical phenomena ranging in scale from supermassive quasars in distant parts of the universe down to stellar mass X ray sources within our own galaxy may be attributed to accretion discs [8, 9, 10] round more or less massive black holes has however provided the motivation for increasingly realistic numerical simulations (Figures 5 and 6) of what would be seen from outside in the presence of an illuminating source of this kind [11, 12].

As the most easily calculable example, I have shown in the appendix how to work out the case shown in Figure 1 of an isolated spherical black hole for which the only source of illumination is a uniform distant sky background, viewed as a function of proper time,

$$\tau = -\frac{4m}{3} \left(\frac{r}{2m}\right)^{3/2}, \quad (2)$$

34
by a (doomed) observer falling towards the singularity inside the black hole, with zero energy and angular momentum.

In such a case the redshift \( Z \) determining the observed energy \( E/(1+Z) \) of a photon emitted from the sky background with the uniform average energy \( \bar{E} \) say will be given by the formula

\[
Z = \frac{-\cos \alpha}{\sqrt{r/2m}}, \quad \alpha > \beta, \quad (3)
\]

where \( \alpha \) is the apparent angle of reception, which must of course exceed the apparent angle \( \beta \) subtended by the black hole. This means that the redshift will be positive (so that the sky will appear darker than normal) due to the Doppler effect, for photons coming in from behind the observer (with \( \alpha > \pi/2 \)). However photons received in the range \( \beta < \alpha < \pi/2 \) will be blueshifted by an amount that will diverge, as shown in Figure 7, as the singularity is approached.

**DISCOVERY OF HORIZON STABILITY AND OF KERR SOLUTION**

Following the demise of Einstein (and the development of nuclear weapons) a new (less inhibited) generation of physicists, lead by Wheeler and Zel’dovich, came to recognise the likelihood – and need in any case for testing – of stability with respect to non-spherical perturbations of what was termed a “black hole”. Work by Vishweshwara [13], Price [14], and others confirmed that “anything that can be radiated away will be radiated away” – leaving a final equilibrium state characterised only by mass and angular
momentum. The (still open) mathematical question of the extent to which this remains true (with singularities hidden inside horizon) for very large deviations from sphericity was raised by the “cosmic censorship” conjecture formulated by Roger Penrose [15, 16] but in any case the relevance of black holes for astrophysical phenomena (notably quasars) was generally accepted in astronomical circles from 1970 onwards.

The generic form of what was afterwards recognised to be the final black hole equilibrium state in question was discovered in 1963, when Roy Kerr announced [17, 18] that “among the solutions ... there is one which is stationary ... and also axysymmetric. Like the Schwarzschild metric, which it contains, it is type D ... m is a real constant ...

The metric is

\[ ds^2 = (r^2 + a^2 \cos^2 \theta)(d\theta^2 + \sin^2 \theta \, d\phi^2) + 2(adu + a\sin^2 \theta \, d\phi)(dr + a\sin^2 \theta \, d\phi) \]

\[ - \left(1 - \frac{2mr}{r^2 + a^2 \cos^2 \theta}\right)(du + a\sin^2 \theta \, d\phi)^2, \]  

where \( a \) is a real constant. This may be transformed to an asymptotically flat coordinate system ... we find that \( m \) is the Schwarzschild mass and \( ma \) the angular momentum ”.

Since the black hole concept had still not been clearly formulated then, it was at first (wrongly) supposed that the physical relevance of this vacuum solution would be as the exterior to a compact self gravitating body like a neutron star, as suggested by Kerr’s (off the mark) conclusion [17] that it would be “desirable to calculate an interior solution.”

What actually makes the Kerr metric so important however, as can be see from Figure 9 (using C.P. diagrams, which were originally developed for this purpose) is the feature first clearly recognised [19, 20] by Bob Boyer in 1965, which is that for \( a^2 \leq m^2 \) the distant sky limit known as “asymptopia” is both visible and accessible only in a non-singular “domain of outer communications” bounded by past and future null (outer)

FIGURE 7. Plot of reception angle \( \alpha \) against countdown proper time \( \tau \) for arrival at singularity of radially falling zero-energy observer in units such that \( 2m = \sqrt{2/3} \) so that null orbit radius \( r = 3m \) is crossed when \( \tau = -1 \). Constant brightness contours indicate intensity of light received from uniform background, which will be inversely proportional to 4th power of redshift factor \((1 + Z)\), having unshifted value for rays arriving at angle \( \pi/2 \).
The topology within the black (and white) hole regions was first elucidated \cite{21} in terms of Conformal Projections on the symmetry axis in 1966 and then completely \cite{22, 23, 24} by Boyer, Lindquist and myself in 1967 and 1968 – the year when the much needed term “black hole” was finally introduced by Wheeler to describe the region from which light cannot escape to “asymptopia”. (A “white hole” region would be one that could not receive light from “asymptopia”.) In the generic rotating case (unlike the static Schwarzschild limit) the well behaved domain outside the black hole horizon includes an “ergosphere” region where, as shown in Figure 8, the Killing vector generating the stationarity symmetry becomes spacelike, so that (globally defined) particle energies can be negative.

In contrast with the good behavior of the outer region, $r > m + c$, I found that, as well as having the irremovable ring shaped curvature singularity already noticed by Kerr where $r^2 + a^2 \cos^2 \theta \to 0$, the inner parts of the rotating Kerr solutions would always be causally pathological, due to the existence near the ring singularity of a small region (see Figure 10) where the axial symmetry generating Killing vector becomes timelike\cite{23, 25}. This feature gives rise to a causality violating “time machine region” (a feature so “outrageous” as to be unmentionable even by Thorne \cite{1}) that would extend all the way out to “asymptopia” (meaning $r \to +\infty$) in the – presumably unphysical – case for which $a^2 > m^2$. (I would emphasize that this kind of time machine, like those recently considered by Ori \cite{32}, would survive even if one takes the covering space, unlike a time machine of the wormhole kind discussed by Thorne \cite{1} which is merely an artefact of multiply connected space time topology).

In so far as the (physically relevant) black hole cases characterised by $a^2 \leq m^2$ are concerned, the good news \cite{23, 25} (for believers in causality) is that the closed timelike lines are all contained within the inner region $r < m - c$. The boundary of the time machine region is constituted by the “inner horizon”, where $r = m - c$. which acts as a Cauchy hypersurface from the point of view of initial data for formation of the black hole.
by gravitational collapse. Unlike the outer horizon \( r = m + c \), whose stability throughout the allowed range \( 0 \leq a^2 < m^2 \) has been even confirmed by Whiting [26], it was to be expected [27, 28] that a Cauchy horizon of the kind occurring at \( r = m - c \) would be unstable, and it has been shown that outcome is likely to be the formation of a curvature singularity of the weak kind designated by the term “mass inflation” [29, 30, 31].

**SEDUCTIVE MATHEMATICAL FEATURES OF KERR TYPE METRICS**

In his original 1963 letter [17], and with Alfred Schild [18] in a sequel, Kerr obtained the useful alternative form

\[
\text{ds}^2 = g_{\mu\nu} \text{d}x^\mu \text{d}x^\nu = \eta_{\mu\nu} + 2(m/U)n_\mu n_\nu \tag{6}
\]

with null covector \( n_\mu \text{d}x^\mu = du + a\sin^2 \theta \, d\phi \), for \( U = (r^2 + a^2 \cos^2 \theta)/r \), in a flat background. The latter was obtained in the Minkowski form,

\[
\eta_{\mu\nu} \text{d}x^\mu \text{d}x^\nu = \text{d}\tilde{t}^2 + \text{d}\tilde{z}^2 + \text{d}\tilde{\rho}^2 - \text{d}r^2 \tag{7}
\]

by setting \( \tilde{t} = u - r \), \( \tilde{z} = a \cos \theta \), \( \tilde{x} + i\tilde{y} = (r - ia)e^{i\phi}\sin \theta \), which gave

\[
n_\mu \text{d}x^\mu = \text{d}\tilde{t} + \frac{\tilde{z}\text{d}\tilde{z}}{r} + \frac{(r\tilde{x} - a\tilde{y})\text{d}\tilde{x} + (r\tilde{y} + a\tilde{x})\text{d}\tilde{y}}{r^2 + a^2} \tag{8}
\]

(This form of pure vacuum solution was generalised to higher dimensions by Myers and Perry[33]. It is perhaps of greater current cosmological interest – in view of the evidence that the expansion of the universe is accelerating – that this form has also been extended to include a cosmological constant in a 4 dimensional De Sitter background by myself [24, 34], while further generalisations to a De Sitter background in 5 and higher dimensions [35, 36] have been obtained more recently.)

As well as time and axial symmetry, the Kerr solution has a discrete PT symmetry that was predictable from Papapetrou’s “circularity” theorem [37], and made manifest in 1967 [22] by the Boyer Lindquist transformation

\[
\text{d}t = \text{d}u - (r^2 + a^2)\Delta^{-1}\text{d}r, \quad \text{d}\phi = -\text{d}\phi + a\Delta^{-1}\text{d}r \tag{9}
\]

with

\[
\Delta = r^2 - 2mr + a^2. \tag{10}
\]

This gives Kerr’s null form as

\[
n_\mu \text{d}x^\mu = \text{d}\tilde{t} - a\sin^2 \theta \, \text{d}\phi + \rho^2 \Delta^{-1}\text{d}r, \quad \rho = \sqrt{r^2 + a^2 \cos^2 \theta}, \tag{11}
\]

The metric itself is thereby obtained in the convenient form

\[
\text{ds}^2 = \rho^2 \left( \frac{\text{d}r^2}{\Delta} + \text{d}\theta^2 \right) + (r^2 + a^2)\sin^2 \theta \, \text{d}\phi^2 + \frac{2mr}{\rho^2}(\text{d}t - a\sin^2 \theta \, \text{d}\phi)^2 - \text{d}r^2 \tag{12}
\]
in which there are cross terms involving the non-ignorable differentials, \(dr\) and \(d\theta\), but – as the price for this simplification – if \(a^2 \leq m^2\) there will be a removable coordinate singularity on the null “horizon” where \(\Delta\) vanishes.

Whereas the possibility of making the foregoing simplification was predictable in advance, there was no reason to anticipate the discovery [23, 38] that, in addition to the ordinary “circular” symmetry generated by Killing vectors, \(k^\mu \partial / \partial x^\mu = \partial / \partial t\) and \(h^\mu \partial / \partial x^\mu = \partial / \partial \phi\), the Kerr metric would turn out to have the hidden symmetry that is embodied in the canonical tetrad

\[
g_{\mu\nu} = \sum_{i=1}^{3} \vartheta^{i}_\mu \vartheta^{i}_\nu - \vartheta^{0}_\mu \vartheta^{0}_\nu \tag{13}
\]

specified by

\[
\vartheta^{i}_\mu dx^\mu = \left(\rho / \sqrt{\Delta}\right) dr, \quad \vartheta^{2}_\mu dx^\mu = \rho d\theta, \tag{14}
\]

\[
\frac{\vartheta^{3}_\mu dx^\mu}{\sin \theta} = \left(\frac{r^2 + a^2}{\rho} - adt\right) d\phi - a \sin^2 \theta d\phi, \quad \frac{\vartheta^{0}_\mu dx^\mu}{\sqrt{\Delta}} = \frac{dt - a \sin^2 \theta d\phi}{\rho}. \tag{15}
\]

In terms of this canonical tetrad, the Kerr-Schild form of the metric is expressible as

\[
g_{\mu\nu} = \eta_{\mu\nu} + 2mr(\vartheta^{0}_\mu + \vartheta^{i}_\mu)(\vartheta^{0}_\nu + \vartheta^{i}_\nu), \tag{16}
\]

while the Killing-Yano 2-form brought to light by Roger Penrose and his coworkers is expressible as

\[
f_{\mu\nu} = 2a \cos \theta \left[ \vartheta^{i}_\mu \vartheta^{0}_\nu \right] - 2r \left[ \vartheta^{0}_\mu \vartheta^{i}_\nu \right]. \tag{17}
\]
The property of being a Killing-Yano 2-form means that it is such as to satisfy the very restrictive condition condition

\[ \nabla_\mu f_{\nu\rho} = \nabla_{[\mu} f_{\nu\rho]} , \]

thus providing a symmetric solution

\[ K_{\mu\nu} = f_{\mu\rho} f_{\nu\sigma} = a_{\mu} k_{\nu} + ah_{\mu} \]

of the Eisenhart type Killing tensor equation, as well as secondary and primary solutions

\[ \tilde{k}_\mu = K_{\mu\nu} k_\nu = a^2 k_\mu + ah_\mu \]

and \[ k_\mu = \frac{1}{3} \epsilon^{\mu\nu\rho\sigma} \nabla_\nu f_{\rho\sigma} \]

of the ordinary Killing vector equation \[ \nabla_\mu \tilde{k}_\nu = 0 . \]

For affine geodesic motion, \[ p_\nu \nabla_\nu p_\mu = 0 , \]

one thus obtains (energy and axial angular momentum) constants \[ \mathcal{E} = k_\nu p_\nu \]

and \[ \mathcal{M} = h_\nu p_\nu \], with (angular momentum) \[ \mathcal{J}_\mu = f_\mu \nu p_\mu \]

obeying \[ p_\nu \nabla_\nu \mathcal{J}_\mu = 0 . \]

There will also [40] be corresponding (self adjoint) operators

\[ \mathcal{E} = ik_\nu \nabla_\nu , \quad \mathcal{M} = ih_\nu \nabla_\nu , \quad \mathcal{K} = \nabla_\mu K^{\mu\nu} \nabla_\nu , \]

whose action on a scalar field commute with that of the the Dalembertian \[ \Box = \nabla^\nu \nabla_\nu \]: in other words \[ \{ \mathcal{E} , \Box \} = 0 , \quad \{ \mathcal{M} , \Box \} = 0 , \]

and (consistently with the integrability condition \[ K^{\rho}_{\mu\nu} R_{\nu\rho} = 0 \] also \[ \{ \mathcal{K} , \Box \} = 0 . \]

The ensuing integrability of the geodesic equation [23] and of the scalar wave equation is equivalent to their solubility by separation of variables [23, 38]. The possibility of extending these rather miraculous separability properties to the neutrino equation [41] and even to the massive spin 1/2 field [42, 43] as governed by the Dirac operator \[ \mathcal{D} = \gamma^\mu \nabla_\mu \] is attributable to corresponding spinor operator conservation laws

\[ [ \mathcal{E} , \mathcal{D} ] = 0 , \quad [ \mathcal{M} , \mathcal{D} ] = 0 , \quad [ \mathcal{J} , \mathcal{D} ] = 0 , \]

of energy, axial angular momentum, and (unsquared) total angular momentum, as respectively given [44] by

\[ \mathcal{E} = ik_\nu \nabla_\nu + \frac{1}{4} i(\nabla_\nu k_\nu) \gamma^\mu \gamma^\nu , \quad \mathcal{M} = ih_\nu \nabla_\nu + \frac{1}{4} i(\nabla_\nu h_\nu) \gamma^\mu \gamma^\nu , \]

and

\[ \mathcal{J} = i\gamma^\mu (\gamma^5 f_\mu \nu \nabla_\nu - k_\mu ) . \]

Such a neat commutation formulation is not (yet?) available for Teukolsky’s extension [45, 46] of solubility by separation of variables to massless spin 1 and spin 2 fields representing electromagnetic and gravitation perturbations – of which the latter are particularly important for Bernard Whiting’s demonstration [26] of stability. An even more difficult problem is posed by the charged generalisation [47] of the Kerr black hole metric, which retains many of its convenient properties (and is noteworthy for having the same gyromagnetic ratio as the Dirac electron [23, 48]) but which gives rise to a system of coupled electromagnetic and gravitational perturbations that has so far been found to be entirely intractible.
The overwhelming importance of Kerr solution derives from its provision of the generic representation of the final outcome of gravitational collapse, as was made fairly clear in 1971 by the prototype *no-hair theorem* [49, 50] proving that no other vacuum black hole equilibrium state can be obtained by continuous axisymmetric variation from the spherical Schwarzschild solution that had been shown by the earlier work of Israel [7] (before the generic definition of a black hole was available) to be only static possibility.

Conceivable loopholes (such as doubts about the axisymmetry assumption) in the reasoning leading to this conclusion (which was rapidly – perhaps too uncritically – accepted in astronomical circles) were successively dealt with by the subsequent mathematical work of Stephen Hawking [51], David Robinson [52] and other more recent contributors [53, 54, 55] to what has by now become a rather complete and watertight uniqueness theorem for pure vacuum black hole solutions in 4 spacetime dimensions. It should however be remarked [56] that there are some mathematical loose ends (concerning assumptions of analyticity and causality) that still need to be tidied up.

The demonstration uses ellipsoidal coordinates for the 2-dimensional space metric $d\hat{s}^2 = d\lambda^2/(\lambda^2 - c^2) + d\mu^2/(1 - \mu^2)$, in terms of which the generic stationary axisymmetric asymptotically flat vacuum metric is known from the work of Papapetrou [37] to be expressible in the form

$$ds^2 = \rho^2 d\hat{s}^2 + X(d\varphi - \omega dt)^2 - (\lambda^2 - c^2)(1 - \mu^2)dr^2.$$  \hspace{1cm} (24)

for which, by the introduction of an Ernst [57] type potential given by $X^2 \partial \omega / \partial \lambda = (1 - \mu^2) \partial Y / \partial \mu$, the relevant Einstein equations will be obtainable from the (positive...
definite) action $\int d\lambda \, d\mu (|\hat{\nabla}X|^2 + |\hat{\nabla}Y|^2)/X^2$.

The black hole equilibrium problem is thus [49, 24, 50] reduced to a non linear 2
dimensional elliptic boundary value problem for the scalars $X, Y$, subject to conditions
of regularity on the horizon (with rigid angular velocity $\Omega$ ) where $\lambda = c$ and to
appropriate boundary conditions on the axis where $\mu = \pm 1$ and at large radius $\lambda \to \infty$
in terms of angular momentum $ma$.

The uniqueness theorem states that this 2 dimensional boundary problem has no
solutions other that those given (with $\lambda = r - m, \mu = \cos \theta$) by the Kerr solution
having mass $m = \sqrt{c^2 + a^2}$ and horizon angular velocity $\Omega = a/2m(m + c)$. The proof
is obtained from an identity equating a quantity that is a positive definite function of
the relevant deviation (of some other hypothetical solution from the Kerr value) to a
divergence whose surface integral can be seen to vanish by the boundary conditions.

The original no-hair theorem (applying just to the small deviation limit) was based on
an infinitesimal divergence identity that I obtained by a hit and miss method [49] that
was generalised by Robinson [52] to the finite difference divergence identity that was
needed to complete the proof in the pure vacuum case. For the electromagnetic (Einstein
Maxwell) generalisation, the analogous step from an infinitesimal no-hair theorem[58]
to a fully non-linear uniqueness theorem was more difficult, and was not obtained until
our hit and miss approach was superceded by the more sophisticated methods that were
developed later on by Mazur [59, 60] and Bunting [61, 62].

FURTHER DEVELOPMENTS

After it had become clear that (in the framework of Einstein’s theory) the Kerr solutions
(with $a \leq m$ ) are the only vacuum black hole equilibrium states, the next thing to be
investigated was the way the black holes will evolve when the equilibrium is perturbed.
A particularly noteworthy result, based on concepts (see Figure 11 ) developed in
collaboration with Penrose [63] was the demonstration by Stephen Hawking [64, 51]
that the area of a black hole horizon (which is proportional to what Christodoulou [65]
had previously identified as irreducible mass) can never decrease. More particularly it
was shown [66] that the area would grow, not only when the hole swallowed matter but
more generally whenever the null generators of the horizon were subjected to shear. It
was remarked that this effect could be described in terms of an effective viscosity and
that the horizon could also be characterised [67, 68] by an effective resistivity.

Later astrophysical developments were concerned more with surrounding or infalling
matter – for example in accretion discs – than with the black hole as such, at least until
recently. However the prospect of detecting gravitational radiation in the foreseeable
future has encouraged a resurgence of interest in purely gravitational effects, particularly
those involved in binary coalescence. The climax of a coalescence is too complicated
to be dealt with except by advanced methods of numerical computation, but the quasi
stationary preliminary stages are more amenable [69, 70, 71, 72, 73, 74], as also are the
final stages of ringdown, which can be analysed in terms of quasi normal modes (and
their superpositions in power law tails of the kind first described by Price [14]) which
have been the subject of considerable attention, particularly concerning the influence of
rotation [75, 76, 77, 78, 79].
APPENDIX: NULL GEODESICS IN SPHERICAL CASE

Although it would be insufficient for the complete Kruskal (black and white hole) extension, in order to cover a purely black hole (Oppenheimer Sneyder type) extension of the Schwarzschild solution, it will suffice to use an outgoing null coordinate patch of the kind introduced for the Kerr metric (4) for which, when $a = 0$, the metric will be given in terms of $x^0 = u$, $x^1 = r$, $x^2 = \theta$, $x^3 = \phi$, simply by

$$ds^2 = -(1 - 2m/r) du^2 + 2 du dr + r^2 d\theta^2 + \sin^2 \theta d\phi^2. \quad (25)$$

Within such a system, an observer falling in freely from a large distance with zero energy and angular momentum will have a geodesic trajectory characterised by fixed values of the angle coordinates $\theta$ and $\phi$ and by a radial coordinate $r$ that is given implicitly as a monotonically decreasing function of proper time by (2) and as a monotonically decreasing function of the ignorable coordinate $u$ by the relation

$$\frac{u_0 - u}{2m} = \frac{2}{3} \sqrt{\frac{r}{2m}} \left( \frac{r}{2m} + 3 \right) - \frac{r}{2m} - 2 \ln \left( 1 + \sqrt{\frac{r}{2m}} \right), \quad (26)$$

in which $u_0$ is a constant of integration specifying the value of $u$ for which the trajectory terminates at the singular limit $r \to 0$. For such a trajectory the (future oriented) timelike unit tangent vector will be given by

$$e^0_{(0)} = \left( 1 + \sqrt{2m/r} \right)^{-1}, \quad e^1_{(0)} = -2m/r, \quad (27)$$

and the tetrad specifying a corresponding local reference frame can be completed in a natural manner by using the associated (outward oriented) orthogonal spacelike unit
vector, which will be given by
\[ e^0_{(1)} = \left(1 + \sqrt{2m/r}\right)^{-1}, \quad e^1_{(1)} = 1, \quad (28) \]

together with two other horizontally oriented unit vectors whose specification will not matter for the present purpose because of the rotation symmetry of the system.

Let us consider the observation of photon that arrives with trajectory deviating by an angle \( \alpha \) say from the outward radial direction. The first two components of its (null) momentum vector, \( p^\mu \) say, will evidently be given in terms of its energy \( \tilde{E} \) with respect to such a frame by
\[ p^i = \tilde{E} (e^i_{(0)} + \cos \alpha e^i_{(1)}), \quad (29) \]

for \( i = 0, 1 \). This locally observed energy is to be compared with the globally defined photon energy (as calibrated with respect to the asymptotic rest frame at large distance) that will be given in terms of the timelike Killing vector with components \( k^\mu = \delta^\mu_0 \) by
\[ E = -k^\mu p_\mu = (1 - 2m/r) p^0 - p^1, \quad (30) \]

the important feature of the latter being that it is conserved by the affine transport of the momentum vector along the null geodesic photon trajectory to which it is tangent. It can thus be seen that the corresponding locally observed energy \( \tilde{E} \) will be related to the globally defined energy constant \( E \) by
\[ \tilde{E} = E (1 + Z), \quad (31) \]

with the redshift \( Z \) given by (3), and that the associated component ratio will be given by
\[ \frac{p^1}{p^0} = (1 + \sqrt{2m/r}) \left(\frac{\cos \alpha - \sqrt{2m/r}}{\cos \alpha + 1}\right), \quad (32) \]

As the (unsurprising) spherical limit of the (still rather mysterious) separability of Kerr’s rotating generalisation, the evolution of the relevant affinely transported momentum components will be given in terms of the energy constant \( E \) and the associated squared angular momentum constant \( \mathcal{K} \) [23] (using a dot for differentiation with respect to the affine parameter) by
\[ p^0 = \dot{u} = \pm \sqrt{\mathcal{R} + \mathcal{P}}, \quad p^1 = \dot{r} = \pm \frac{\mathcal{P}}{r^2}, \quad (33) \]

where
\[ \mathcal{P} = \mathcal{E} r^2, \quad \mathcal{R} = \mathcal{P}^2 - \mathcal{K} (r^2 - 2mr). \quad (34) \]

We thereby obtain
\[ \frac{p^0}{p^1} = \frac{du}{dr} = (1 - 2m/r)^{-1} \left(1 \pm \frac{\mathcal{P}}{\sqrt{\mathcal{R}}}\right)^{-1}, \quad (35) \]
and hence, by comparison with (32),
\[ \pm \frac{\sqrt{R}}{\mathcal{P}} = \sqrt{r/2m} \left( \frac{\cos \alpha - \sqrt{2m/r}}{\sqrt{r/2m} - \cos \alpha} \right) \cdot (36) \]

This expression can be used to evaluate the squared angular momentum constant as a function of the locally defined energy \( \tilde{\mathcal{E}} \) and angle \( \alpha \) in the form
\[ \mathcal{K} = r^2 \tilde{\mathcal{E}}^2 \sin^2 \alpha \cdot (37) \]
in which the variable \( \tilde{\mathcal{E}} \) will itself be given via (31) in terms of the globally defined energy constant \( \mathcal{E} \) by the – red or blue – shift formula (3) so that for the square of the constant ratio of angular momentum to energy one obtains
\[ \frac{\mathcal{K}}{\mathcal{E}^2} = \left( \frac{r \sin \alpha}{1 - \sqrt{2m/r \cos \alpha}} \right)^2 \cdot (38) \]

With respect to an unconventional affine parameter orientation condition to the effect that the energy should always be non-negative, \( \mathcal{E} \geq 0 \), it can be seen (in view of the consideration that the squared angular momentum constant must necessarily be non-negative, \( \mathcal{K} \geq 0 \) ) that a null geodesic segment will be appropriately be describable as “incoming” or “outgoing” according to whether the upper or lower of the sign possibilities \( \pm \) is applicable, i.e. according to whether the right hand side of (36) is positive or strictly negative. It is however to be remarked that this convention will be consistent with the usual requirement that the affine parameter orientation be future oriented, giving \( \dot{u} \geq 0 \), only outside the horizon and for “ingoing” null segments within the horizon, where \( r < 2m \), but that for “outgoing” null segments within the horizon it would entail the opposite orientation convention, giving \( \dot{u} \leq 0 \). With respect to the usual parameter orientation condition giving \( \dot{u} \geq 0 \) the “outgoing” null segments within the horizon will need to be parametrised the other way round, which means that they will be characterised by negative energy \( \mathcal{E} < 0 \) and by the upper of the sign possibilities \( \pm \).

Whichever convention is used, it can be seen that within the horizon the radius \( r \) will always be a decreasing function of \( u \), even for the (relatively) “outgoing” null segments, and that only an “incoming” null segment can cross the horizon at a finite value of \( u \). It can be seen from (36) that outside the horizon (i.e. for \( r > 2m \) ) the criterion for a null segment to be classified as “incoming” is that it should have \( \cos \alpha \leq \sqrt{2m/r} \) and that the corresponding requirement within the horizon is \( \cos \alpha \leq \sqrt{r/2m} \).

It can be deduced from the expression (34) that the function \( R \) will remain positive whenever \( r \) is positive if \( \mathcal{K} / \mathcal{E}^2 \leq 27m^2 \). In such a case, the null geodesic will either be permanently “incoming”, proceeding all the way from “infinity” (i.e. the limit \( r \to \infty \) ) down to the internal singularity (i.e. the limit \( r \to 0 \) ), or else it will be permanently “outgoing”, proceeding all the way to the singularity or to infinity depending on whether it inside or outside the finite horizon radius value \( r = 2m \) to which it extends in the infinite past, i.e. as \( u \to -\infty \). As well the such “ingoing” and permanently “outgoing” possibilities, the critical case
\[ \mathcal{K} / \mathcal{E}^2 = 27m^2 \cdot (39) \]
includes also the exceptional possibility of a marginally outgoing – effectively “trapped” – null trajectory with fixed radius \( r = 3m \).

When the angular momentum exceeds this critical value, i.e. if \( \mathcal{K} / \mathcal{L}^2 > 27m^2 \), there will be a forbidden range \( r_- < r < r_+ \) of values of \( r \) for which \( \mathcal{R} < 0 \). It can be seen from 34 that relevant limits are explicitly obtainable, as the non negative solutions of the cubic equation

\[
\mathcal{L}^2 r^3 - \mathcal{K} (r - 2m) = 0,
\]

in the form

\[
r_{\pm} = 2\sqrt{\mathcal{K}/3\mathcal{L}^2} \cos \{\psi_{\pm}/3\},
\]

with

\[
\psi_{\pm} = \pi \mp \arcsin \sqrt{1 - 27\mathcal{L}^2 m^2/\mathcal{K}},
\]

which evidently entails the conditions \( 2m < r_- < 3m < r_+ \).

This means that for a value of \( \mathcal{K} / \mathcal{L}^2 \) above the critical bound (39) the possible null trajectories will be classifiable as “free” or “trapped”. The “free” geodesics are initially “incoming” from “infinity” but become “outgoing” after reaching the inner bound at \( r = r_+ \) so as to remain in the range \( r \geq r_+ \). The “trapped” geodesics are either permanently “ingoing” within the horizon or else are initially “outgoing” from just outside the horizon but become “ingoing” after reaching the outer bound \( r = r_- \) so as to remain within the range \( r \leq r_- \).

For a position in the range \( r \leq 3m \) the only kinds of “bright” geodesic, meaning those coming from the distant sky at “infinity”, are of the permanently “incoming” kind characterised by \( \mathcal{K} / \mathcal{L}^2 \leq 27m^2 \). whereas for a position in the range \( r \leq 3m \) there will also be “bright” geodesics of the “free” kind characterised by \( \mathcal{K} / \mathcal{L}^2 > 27m^2 \). Apart from the special case of the circular null geodesics at \( r = 3m \), all the other kinds of null geodesic can be classified as “dark” since they can be seen to have emerged from near the horizon limit radius \( r \to 2m \) in the distant past (the limit \( u \to -\infty \)) and so can be interpreted as trajectories of very highly redshifted radiation from the infalling matter that be presumed to originally formed the black hole whose static final state is under consideration here.

It can be seen that the ratio \( \mathcal{K} / \mathcal{L}^2 \) specified as a function of \( \cos \alpha \) by (38) will be monotonically increasing in the “incoming” range, i.e. for \(-1 \leq \cos \alpha < \sqrt{2m/r}\) where \( r > 2m \) and for \(-1 \leq \cos \alpha < \sqrt{r/2m}\) where \( r < 2m \). At the upper end of this “incoming” range the ratio \( \mathcal{K}/\mathcal{L}^2 \) the tends to a maximum that will be finite – with value \( r^3/(r-2m) \) – outside the horizon, but that will be infinite inside the black hole. The ratio \( \mathcal{K}/\mathcal{L}^2 \) will then decrease monotonically for the higher “outgoing” part of the range of \( \cos \alpha \).

The critical value (39) will be attained for two values of \( \cos \alpha \), of which the lower one, \( \cos \alpha = \mathcal{K}_- \) say, will be in the “incoming” range, and the higher one, \( \cos \alpha = \mathcal{K}_+ \) will be in the “outgoing” range. It can be seen from (38) that these values will be obtainable as the upper and lower roots of

\[
r^2 (1 - \mathcal{K}_-^2) = 27m^2 (1 - \sqrt{2m/r \mathcal{K}_+})^2,
\]

46
which will be real and distinct except at \( r = 3m \) where they will coincide. The range of angles characterising the “bright” geodesics will therefore be given by

\[
-1 \leq \cos \alpha < \cos \beta
\]  

(44)

so that \( \beta \) will be interpretable as the apparent angular radius of the black hole) with a bounding value \( \cos \beta \) that will be given by \( \cos \beta = \mathcal{X}^- \), within the radius of the circular null trajectory, i.e. for \( r < 3m \), while in the outer regions for which \( r > 3m \) it will be given by \( \cos \beta = \mathcal{X}^+ \).

The required solutions of (43) are expressible in terms of the dimensionless variable \( \tilde{r} = r/2m \) by \( \mathcal{X}^\pm = \left( 27 \sqrt{\tilde{r}} \pm [2\tilde{r}^2 - 3\tilde{r}] \sqrt{\tilde{r}^2 + 3\tilde{r}} \right) / (4\tilde{r}^3 + 27) \). It can thus be seen that (for the freely falling observer) the apparent angular size \( \beta \) of the black hole – as shown in the simulation of Figure 1, and as plotted against the proper time (2) in Figure 7 and Figure 12 – will be given as a function of the dimensionless radial variable \( \tilde{r} = r/2m \) by the analytic formula

\[
\cos \beta = \frac{27 \sqrt{\tilde{r}} + (2\tilde{r}^2 - 3\tilde{r}) \sqrt{\tilde{r}^2 + 3\tilde{r}}}{4\tilde{r}^3 + 27}.
\]  

REFERENCES