Singularities in GR: From Hilbert to Hawking

SCGR2023: Singularities and Curvature in General Relativity
19–23 June 2023, Nijmegen

Klaas Landsman

Institute for Astrophysics, Mathematics, and Particle Physics
and
Radboud Center for Natural Philosophy (under construction)
Radboud University Nijmegen

Singularities in GR: From Einstein to Penrose

SCGR2023: Singularities and Curvature in General Relativity
19–23 June 2023, Nijmegen

Klaas Landsman

Institute for Astrophysics, Mathematics, and Particle Physics
and
Radboud Center for Natural Philosophy (under construction)
Radboud University Nijmegen

Dedicated to Annegret Burtscher and Leo García-Heveling

History of GR: state of the art

History of GR so far has been largely centered around Einstein:

- *The Collected Papers of Albert Einstein*, especially:
  Vol. 4 (1912–1914), Vol. 6 (1914–1917), Vol. 7 (1918–1921)
  `https://einsteinpapers.press.princeton.edu/`
- Renn, J., ed. (2007). *The Genesis of General Relativity* (Springer)
- Janssen, M., Renn, J. (2022). *How Einstein Found His Field Equations: Sources and Interpretation* (Springer)
- Dongen, J. van (2010). *Einstein's Unification* (CUP)

Relatively new research about the "Renaissance of GR":

- Eisenstaedt, J. (2006). *The Curious History of Relativity: How Einstein's Theory of Gravity was Lost and Found Again* (Princeton)
- Blum, A.S., Lalli, R., Renn, J., eds. (2020). *The Renaissance of General Relativity in Context* (Springer)

Dennis Lehmkuhl (Bonn), 2023–2027: ERC Consolidator Grant
GR 1955–1975: Penrose, Hawking, Bondi, Ehlers, Wheeler, etc.

Mathematical GR

From 1920 through 1965, GR was considered to have so few empirically testable predictions that its practitioners in English-speaking countries were largely banished to mathematics departments. When the discoveries of cosmological background radiation, quasars and pulsars made it clear that GR does model important science at astronomical scales, the theory still appeared remote from the microcosmic concerns of most research physicists. (. . .)

The isolation of GR from the rest of theoretical physics was intensified by the special nature of its mathematical tools. Particle physicists could recognize that condensed-matter people were doing quantum field theory; nuclear and molecular physicists used the same quantum mechanics. In the early days, the heavily indexed tensors of GR betokened a kinship with continuum mechanics (similarly exiled to engineering departments), but when relativists fell under the spell of index-free differential forms and algebraic topology, their isolation became complete. (Fulling, 2006)

History of mathematical GR (after Riemann)

Reich, K. (1994). *Die Entwicklung des Tensorkalküls: Vom absoluten Differentialkalkül zur Relativitätstheorie* (Birkhäuser)

Goodstein, J.R. (2018). *Einstein's Italian Mathematicians: Ricci, Levi-Civita, and the Birth of General Relativity* (American Mathematical Society)

Renn, J., Stachel, J. (2007). Hilbert's Foundation of Physics: From a theory of everything to a constituent of general relativity. *The Genesis of General Relativity. Volume 4*, ed. Renn, J., pp. 857–973 (Springer)

Scholz, E., ed. (2001). *Hermann Weyl's Raum - Zeit - Materie and a General Introduction to His Scientific Work* (Birkhäuser)

Stachel, J. (1992). The Cauchy problem in general relativity–The early years. *ibid.*, pp. 407–418

Lichnerowicz, A. (1992). Mathematics and general relativity: A recollection. *Studies in the History of General Relativity*, pp. 103–108 (Birkhäuser).

Choquet-Bruhat, Y. (2014). Beginnings of the Cauchy problem. `arXiv:1410.3490`.

Choquet-Bruhat, Y. (2018). *A Lady Mathematician in this Strange Universe: Memoirs* (World Scientific)

Ringström, H. (2015). Origins and development of the Cauchy problem in general relativity. *Classical and Quantum Gravity* 32:124003.

Primary literature on singularities in GR

Hilbert, D. (1917). Die Grundlagen der Physik (Zweite Mitteilung). *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen*,

Einstein, A. (1918). Kritisches zu einer von Hrn. de Sitter gegebenen Lösung der Gravitationsgleichungen. *Sitzungsberichte der Königlich Preußischen usw.*

Lemaître, G. (1933). L'univers en expansion. *Annales de la Société scientifique de Bruxelles* A 53, 51–85

Raychaudhuri, A.K. (1955). Relativistic cosmology: I. *Physical Review* 98, 1123–1126.

Komar, A. (1956). Necessity of singularities in the solution of the field equations of general relativity. *Physical Review* 104, 544–546.

Misner, C.W. (1963). The flatter regions of Newman, Unti, and Tamburino's generalized Schwarzschild space. *Journal of Mathematical Physics* 4, 924–937

Penrose, R. (1965). Gravitational collapse and space-time singularities. *Physical Review Letters* 14, 57–59

Hawking, S.W. (1966). *Singularities and the geometry of spacetime*. Adams Prize Essay ($\Leftarrow$ 1965 Cambridge PhD thesis, $\Rightarrow$ Hawking & Ellis, 1973)

Penrose, R. (1979). Singularities and time-asymmetry. *General Relativity: An Einstein Centenary Survey*, eds. Hawking & Israel, pp. 581–638 (CUP).

Secondary (historical) literature on singularities in GR

Tipler, F.J., Clarke, C.J.S., Ellis, G.F.R. (1980). Singularities and horizons–A review article. *General Relativity and Gravitation: One Hundred Years After the Birth of Albert Einstein*, Vol. 2. ed. Held, A., pp. 97–206 (Plenum Press).

Earman, J., Eisenstaedt, J. (1999). Einstein and singularities. *Studies in History and Philosophy of Modern Physics* 30, 185–235

Earman, J. (1999). The Penrose–Hawking singularity theorems: History and implications. *The Expanding Worlds of General Relativity (Einstein Studies Vol. 7)*, eds. Goenner, H. et al., pp. 236–267 (Birkhäuser)

Senovilla, J.M.M. (1998). Singularity theorems and their consequences. *General Relativity and Gravitation* 30, 701–848.

Senovilla, J.M.M. (2022). The influence of Penrose's singularity theorem in general relativity. *GRG* 54:151 (*Proceedings SCRI21*)

Landsman, K. (2022). Penrose's 1965 singularity theorem: From geodesic incompleteness to cosmic censorship. *GRG* 54:115 (*Proceedings SCRI21*)

Landsman, K. (2022). *Foundations of General Relativity: From Einstein to Black Holes, 2nd edition* (Radboud University Press)

Lehmkuhl, D. (2023). The prediction and interpretation of singularities and black holes: From Einstein and Schwarzschild to Penrose and Wheeler (draft)

Roger Penrose, Oxford, July 2, 2022

Hilbert (1917)

- Hilbert, Einstein, Weyl, and many others saw singularities as points/regions inside spacetime (motivated by Schwarzschild $r = 2m$ and de Sitter solution) which even now is not a foolish attitude compared with algebraic geometry & stratified spaces

Hilbert about $r = 2m$ in the Schwarzschild metric:

> *A line element or a gravitational field g_{ij} is regular at a point if it is possible to introduce by a reversible, one-one transformation a coordinate system, such that in this system the corresponding functions g'_{ij} are regular at that point, i.e. they are continuous and arbitrarily differentiable at the point and in a neighbourhood of the point, and the determinant g' is different from 0*

Mistake (even then): a coordinate transformation is supposed to be defined (and invertible) even in the region where one of the coordinate systems fails (which would even make the Euclidean metric in $\mathbb{R}^n$ singular at $r = 0$ if is expressed in polar coordinates!)

Einstein

Except for a few brief periods, Einstein was uninterested in analysing the nature of the spacetime singularities that appeared in solutions to his gravitational "eld equations for general relativity. The existence of such monstrosities reinforced his conviction that general relativity was an incomplete theory which would be superseded by a singularity-free unified field theory. Nevertheless, on a number of occasions between 1916 and the end of his life, Einstein was forced to confront singularities. His reactions show a strange asymmetry: he tended to be more disturbed by (what today we would call) merely apparent singularities and less disturbed by (what we would call) real singularities. Einstein had strong a priori ideas about what results a correct physical theory should deliver. In the process of searching through theoretical possibilities, he tended to push aside technical problems and jump over essential difficulties. Sometimes this method of working produced brilliant new ideas (such as the Einstein–Rosen bridge) and sometimes it lead him to miss important implications of his theory of gravity (such as gravitational collapse). (Earman & Eisenstaedt, 1999)

Einstein (and others) on Schwarzschild solution

- $r = 0$: artefact of vacuum solution, solved by matter source
- $r = 2m$: regarded as a real "singularity" by Einstein, Hilbert, and many others: *discontinuity* (Schwarzschild), *magic circle* (Eddington), *barrier* (Kottler), *limit circle* (Brillouin), *death* (Nordmann) – despite Eddington (1924) & Lemaître (1933)

  Finally settled by Finkelstein (1958), Kruskal (1960), etc.
- But not worrying because physically irrelevant for 3 reasons:

1. Stars could not have $R < 2m$ (Schwarzschild, 1916, 2nd), Einstein (1922): pressure as $r \to 0$ would be infinite (used as *reductio ad absurdum* argument) so no vacuum at $r = 2m$
2. Even if $R < 2m$ for some star, $r = 2m$ would be empirically inaccessible (infinite redshift argument, Eddington)
3. As for $r = 0$: if $r = 2m$ is really a singularity it again indicates that there should be a matter source (Einstein, Eddington)

Einstein on de Sitter space (1918)

$$\boxed{ds^2 = -\cos^2(r/R)dt^2 + dr^2 + R^2\sin^2(r/R)d\Omega}$$

Like $r = 2m$, Einstein regarded $r = \pi R/2$ as a real (*echte*) singularity:

- Singularity := point x (*in* space-time) where some component $g_{\mu\nu}(x)$ or $g^{\mu\nu}(x)$ or its first derivative is discontinuous, such that no choice of coordinates can remove this discontinuity
- Singularity is *real* if it can be connected to some "regular" point P_0 by curve of finite proper length (which Einstein thought was *not* the case for Schwarzschild $r = 2m$ but which is correct for $r = \pi R/2$)

Serious attempt–but both the definition and its application are flawed:

1. $r = \pi R/2$ is a coordinate singularity (dS solution is regular)
2. Points that can be connected by a curve with *infinite* proper lenght can also be connected by curve of *finite* proper length (C.J.S. Clarke, 1993: "wiggle" curve to make it approximately lightlike)
3. No causality requirement; so definition is physically obscure

But (with some goodwill) foreshadows incomplete causal geodesics!

Lemaître: The expanding Universe (1933)

The equations of the Friedmann universe admit solutions where the radius of the universe tends to zero for non-zero mass. This contradicts the generally accepted result that a given mass cannot have a radius smaller than $2m$.

- Resolves $r = 2m$ "singularity" in Schwarzschild metric
- Misses $r = 2m$ as event horizon (clarified by Finkelstein, 1958)
- Early cosmological "singularity theorem" (similar to Tolman, Robertson, de Sitter, Synge, 1930–1935), and the first based on *energy conditions* ($T^i_i < 0$ and $|T^i_i| \leq T^4_4$ for $i = 1,2,3$):

  Metric $ds^2 = -dt^2 + \sum_i h_{ii}(t)(dx^i)^2$, $R := \det(h)^{1/6} = \det(-g)^{1/6}$
- If $\dot{R}(t) < 0$ for some $t > 0$, then $R(t_0) = 0$ for some $t_0 < t$
- No attempt to define singularity and unclear what coordinate dependence of R means (hindsight: $R = 0$ is real singularity)

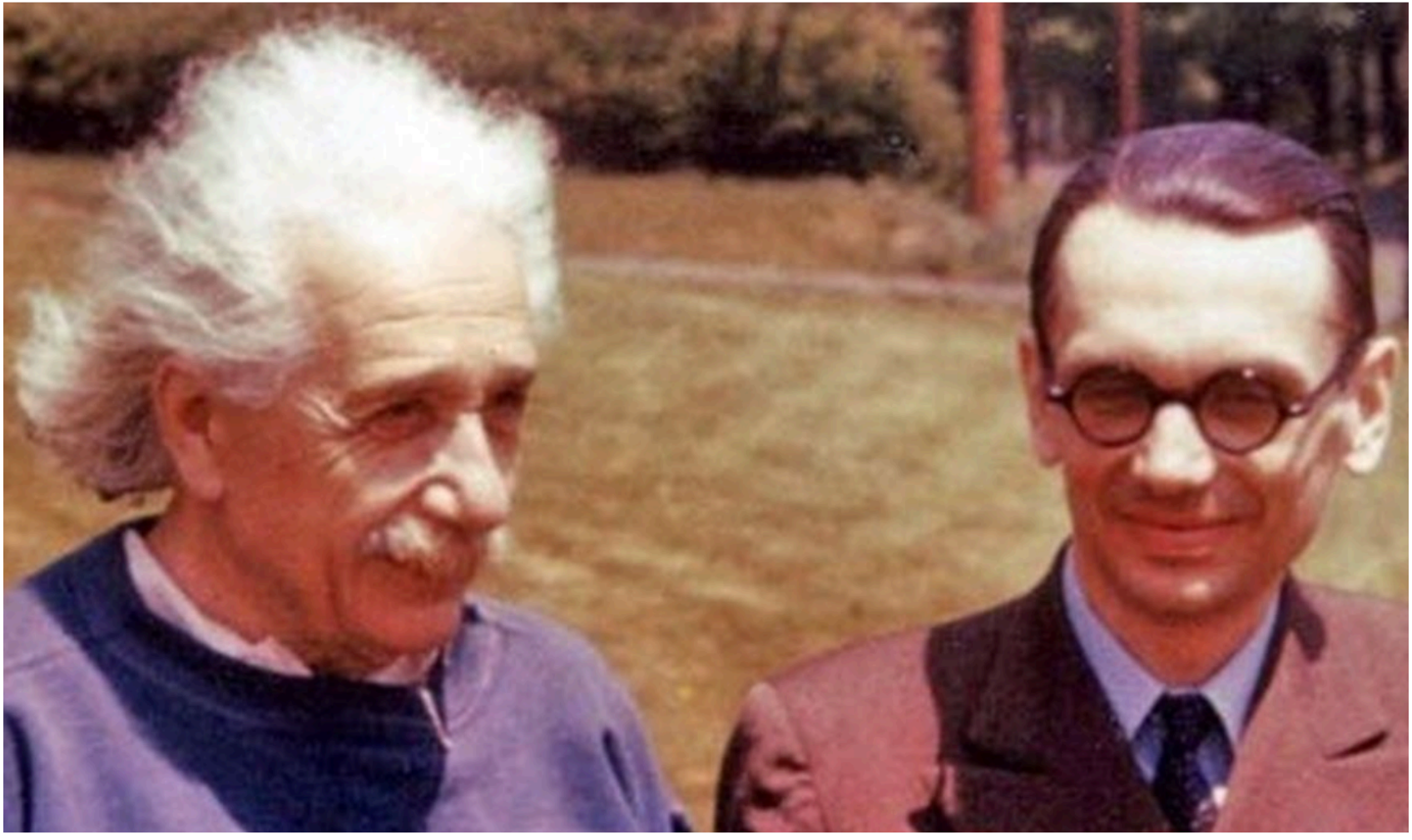
The matter has to find, though, a way of avoiding the vanishing of its volume (...) Forces which prevent the mutual interpenetration of elementary particles are without doubt capable of stopping the contraction [when $R \sim$ solar system]

"Renaissance" of general relativity (1950–1970)

After solar eclipse sensation in 1919, GR gradually decoupled from mainstream physics (especially from QM). But revival $\sim$ 1950:

- Gödel's solution (1949) with rotating fluid sparked further mathematical work (on global geometry, congruences, and relativistic fluid mechanics): Lichnerowicz, Choquet-Bruhat (France); Raychaudhuri (India); Komar, Markus, Taub, Misner (US); Jordan, Heckmann, Schücking, Ehlers (Germany), ...
- Cold War: huge science funding in USA and USSR, nuclear (bomb) physicist moved to GR (Wheeler, Zeldovich) for astrophysics reasons
- GR conferences: Bern (1955), Chapel Hill (1957), Warsaw (1962)
- New astrophysics (Quasars, 1963; CMB, 1965; Pulsars, 1967)

$\Rightarrow$ GR "schools": Princeton (Wheeler), Austin (Schild), Cambridge (Sciama, Hoyle), Syracuse (Bergmann), King's College London (Bondi), Warsaw (Infeld, Trautman), Dublin (Lanczos, Synge, Schrödinger), Moscow (Zeldovich, Landau–Lifshitz–Khalatnikov)



Einstein and Gödel in Princeton, $\sim$ 1949

Wheeler and Sciama had a widespread impact on the development of the understanding of singularities, particularly through their students. Wheeler's students included Misner, Shepley, Thorne, and Geroch [and Christodoulou], while Sciama's included Ellis, Hawking, Carter, Rees, and Clarke. However, Sciama has claimed, on occasion, that his most important contribution to relativity has been in influencing Roger Penrose to work on the subject! (Tipler–Clarke–Ellis, 1980)

Consolidating textbooks of this era (1950–1970)

- Penrose, R. (1972). *Techniques of Differential Topology in Relativity*
- Weinberg, S. (1972). *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*
- Misner, C.W., Thorne, K.S., Wheeler, J.A. (1973). *Gravitation*
- Hawking, S.W., Ellis, G.F.R. (1973). *The Large Scale Structure of Space-Time*

Raychaudhuri (1955) and Komar (1956)

The Friedmann solution, which employs the assumptions that the universe is spacially isotropic and that the state of matter may be represented by incoherent dust, yields the result that the universe is not stationary, but is rather in a state either of expansion from a singular point in time (which would correspond to creation), or of contraction toward a singular point in time (which would correspond to annihilation). The question naturally arises whether such singular points are a consequence of the particular symmetry presupposed in Friedmann's model, or whether perhaps for more general distributions of matter one need not expect instants of creation or annihilation of the universe. The purpose of this paper is to show that singularities are to be expected under very general hypotheses, and in particular that the singular instant of creation (or annihilation) necessarily would occur at a finite time in the past (or future, respectively) (Komar, Introduction)

- R. assumes *anisotropic* dust, K. just energy conditions.
- Same notion of singularity as Lemaître: $\det(-g) = 0$
- Proofs based on Raychaudhuri equation for $\det(-g)^{1/6}$

Misner (1963)

Causal geodesic (in)completeness and global Lorentzian geometry were "in the air", e.g. K. Gödel (1949–52), L. Markus (1955-63), W. Kundt (1963), R. Hermann (1964), M. Fierz & R. Jost (1965), ...

Misner just summarizes 'the present state of the art' on singularities:

the clue to clarity is to refuse ever to speak of a singularity but instead to phrase everything in terms of the properties of differentiable metric fields on manifolds. If one is given a manifold, and on it a metric which does not at all points satisfy the necessary differentiability requirements, one simply throws away all the points of singularity. The starting point for any further discussion is then the largest submanifold on which the metric is differentiable. This is done because there is not known any useful way of describing the singularities of a function except by describing its behavior at regular points near the singularity. The first problem then is to select a criteria which will identify in an intuitively acceptable way a "nonsingular space." Evidently, differentiability is only a minimum prerequisite, since everything becomes differentiable when the singular points are discarded. The problem is rather to recognize the holes left in the space where singular (or even regular) points have been omitted.

Misner (1963)

- Differences between Riemannian and Lorentzian geometry around Hopf–Rinow theorem; highlights that implication geodesically complete $\Rightarrow$ inextendible is valid in both
- Adds that in GR it is 'commonly accepted' that an 'essentially singular space' is geodesically incomplete and inextendible
- Defines curvature singularity as unboundedness of some scalar polynomial in $(\nabla_\tau)R_{\mu\nu\rho\sigma}$ along some open geodesic segment
- Curvature singularity is *sufficient* condition for singularity
- Proposes: Curvature singularity $\Rightarrow$ space-time is "singular" $\Rightarrow$ space-time is geodesically incomplete as well as inextendible
- States necessary and sufficient conditions for singularity but no definition (N.B. geodesic incompleteness is just *necessary*!)

[17]I. S. Shklovsky, Cosmic Radio Waves (Harvard University Press, Cambridge, Massachusetts, 1960), pp. 271-316.
[18]C. Hazard, M. B. Mackey, and A. J. Simmins, Nature 197, 1037 (1963).
[19]M. Schmidt, Nature 197, 1040 (1963).
[20]A. T. Moffet, Science 146, 764 (1964).
[21]M. Ryle and A. Sandage, Astrophys. J. 139, 419 (1964).
[22]I. S. Shklovsky, reference 17, pp. 369, 372.
[23]E. Fermi, Phys. Rev. 75, 1169 (1949).
[24]E. N. Parker, Phys. Rev. 109, 1328 (1958).
[25]T. H. Stix, Plasma Physics Laboratory Report No. MATT-239, Princeton University, 1964 (unpublished).
[26]R. W. Fredricks, F. L. Scarf, and W. Bernstein, to be published.
[27]V. Sarabhai, J. Geophys. Res. 68, 1555 (1963).
[28]E. N. Parker, Astrophys. J. 133, 1014 (1961); Interplanetary Dynamical Processes (John Wiley & Sons, Inc., New York, 1963).
[29]T. Gold, Gas Dynamics of Cosmic Clouds, edited by H. C. van de Hulst and J. M. Burgers (North-Holland Publishing Company, Amsterdam, 1955).
[30]E. N. Parker, Space Sci. Rev. 1, 62 (1962).
[31]M. Neugebauer and C. W. Snyder, Science 138, 1095 (1962).
[32]A. Bonetti, H. S. Bridge, A. J. Lazarus, B. Rossi, and F. Scherb, J. Geophys. Res. 69, 4017 (1963).
[33]E. J. Smith, L. Davis, P. J. Coleman, and C. P. Sonett, Science 139, 1099 (1962).
[34]N. F. Ness, C. S. Scearce, and J. B. Seek, J. Geophys. Res. 69, 3531 (1964).
[35]P. J. Coleman, L. Davis, and C. P. Sonett, Phys. Rev. Letters 5, 43 (1960).
[36]D. A. Bryant, T. L. Cline, U. D. Desai, and F. B. McDonald, J. Geophys. Res. 67, 4983 (1962).
[37]A similar increase in the low-energy proton intensity, in coincidence with the passage of a blast wave, was observed on 12 November 1960 [J. F. Steljes, H. Carmichael, and K. G. McCracken, J. Geophys. Res. 66, 1363 (1961)]. The increases proton intensity may have been partly or wholly of interplanetary origin, but the increase was of such a form that it is not possible to rule out a solar origin.

GRAVITATIONAL COLLAPSE AND SPACE-TIME SINGULARITIES

Roger Penrose
Department of Mathematics, Birkbeck College, London, England
(Received 18 December 1964)

The discovery of the quasistellar radio sources has stimulated renewed interest in the question of gravitational collapse. It has been suggested by some authors[1] that the enormous amounts of energy that these objects apparently emit may result from the collapse of a mass of the order of $(10^6$-$10^8)M_\odot$ to the neighborhood of its Schwarzschild radius, accompanied by a violent release of energy, possibly in the form of gravitational radiation. The detailed mathematical discussion of such situations is difficult since the full complexity of general relativity is required. Consequently, most exact calculations concerned with the implications of gravitational collapse have employed the simplifying assumption of spherical symmetry. Unfortunately, this precludes any detailed discussion of gravitational radiation—which requires at least a quadripole structure.

The general situation with regard to a spherically symmetrical body is well known.[2] For a sufficiently great mass, there is no final equilibrium state. When sufficient thermal energy has been radiated away, the body contracts and continues to contract until a physical singularity is encountered at $r=0$. As measured by local comoving observers, the body passes within its Schwarzschild radius $r=2m$. (The densities at which this happens need not be enormously high if the total mass is large enough.) To an outside observer the contraction to $r=2m$ appears to take an infinite time. Nevertheless, the existence of a singularity presents a serious problem for any complete discussion of the physics of the interior region.

The question has been raised as to whether this singularity is, in fact, simply a property of the high symmetry assumed. The matter collapses radially inwards to the single point at the center, so that a resulting space-time catastrophe there is perhaps not surprising. Could not the presence of perturbations which destroy the spherical symmetry alter the situation drastically? The recent rotating solution of Kerr[3] also possesses a physical singularity, but since a high degree of symmetry is still present (and the solution is algebraically special), it might again be argued that this is not representative of the general situation.[4] Collapse without assumptions of symmetry[5] will be discussed here.

Consider the time development of a Cauchy hypersurface C^3 representing an initial matter distribution. We may assume Einstein's field equations and suitable equations of state governing the matter. In fact, the only assumption made here about these equations of state will be the non-negative definiteness of Einstein's energy expression (with or without cosmological term). Suppose this matter distribution undergoes gravitational collapse in a way which, at first, qualitatively resembles the spherically symmetrical case. It will be shown that, after a certain critical condition has been fulfilled, deviations from spherical symmetry cannot prevent space-time singularities from arising. If, as seems justifiable, actual physical singularities in space-time are not to be permitted to occur, the conclusion would appear inescapable that inside such a collapsing object at least one of the following holds: (a) Negative local energy occurs.[6] (b) Einstein's equations are violated. (c) The space-time manifold is incomplete.[7] (d) The concept of space-time loses its meaning at very high curvatures—possibly because of quantum phenomena.[2] In fact (a), (b), (c), (d) are somewhat interrelated, the distinction being partly one of attitude of mind.

Before examining the asymmetrical case, consider a spherically symmetrical matter distribution of finite radius in C^3 which collapses symmetrically. The empty region surrounding the matter will, in this case, be a Schwarzschild field, and we can conveniently use the metric $ds^2=-2drdv+dv^2(1-2m/r)-r^2(d\theta^2+\sin^2\theta d\varphi^2)$, with an advanced time parameter v to describe it.[8] The situation is depicted in Fig. 1. Note that an exterior observer will always see matter outside $r=2m$, the collapse through $r=2m$ to the singularity at $r=0$ being invisible to him.

After the matter has contracted within $r=2m$, a spacelike sphere S^2 ($t=$const, $2m>r=$const) can be found in the empty region surrounding the matter. This sphere is an example of what will be called here a trapped surface—defined generally as a closed, spacelike, two-surface T^2 with the property that the two systems of null geodesics which meet T^2 orthogonally converge locally in future directions at T^2. Clearly trapped surfaces will still exist if the matter region has no sharp boundary or if spherical symmetry is dropped, provided that the deviations from the above situation are not too great. Indeed, the Kerr solutions with $m>a$ (angular momentum ma) all possess trapped surfaces, whereas those for which $m\leq a$ do not.[9] The argument will be to show that the existence of a trapped surface implies—irrespective of symmetry—that singularities necessarily develop.

The existence of a singularity can never be inferred, however, without an assumption such as completeness for the manifold under consideration. It will be necessary, here, to suppose that the manifold M_+^4, which is the future time development of an initial Cauchy hypersurface C^3 (past boundary of the M_+^4 region), is in fact null complete into the future. The various assumptions are, more precisely, as follows: (i) M_+^4 is a nonsingular (+−−−) Riemannian manifold for which the null half-cones form two separate systems ("past" and "future"). (ii) Every null geodesic in M_+^4 can be extended into the future to arbitrarily large affine parameter values (null completeness). (iii) Every timelike or null geodesic in M_+^4 can be extended

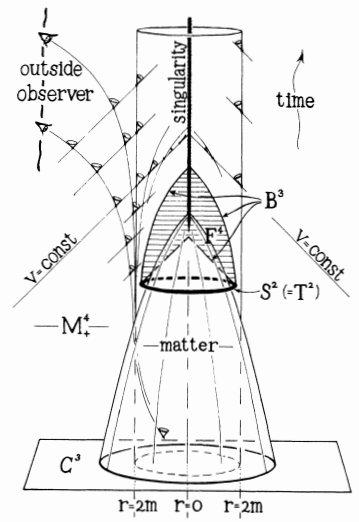


FIG. 1. Spherically symmetrical collapse (one space dimension surpressed). The diagram essentially also serves for the discussion of the asymmetrical case.

into the past until it meets C_3 (Cauchy hypersurface condition). (iv) At every point of M_+^4, all timelike vectors t^μ satisfy $(-R_{\mu\nu}+\frac{1}{2}Rg_{\mu\nu}-\lambda g_{\mu\nu})t^\mu t^\nu \geq 0$ (non-negativeness of local energy). (v) There exists a trapped surface T^2 in M_+^4. It will be shown here, in outline, that (i), $\cdots$, (v) are together inconsistent.

Let F^4 be the set of points in M_+^4 which can be connected to T^2 by a smooth timelike curve leading into the future from T^2. Let B^3 be the boundary of F^4. Local considerations show that B^3 is null where it is nonsingular, being generated by the null geodesic segments which meet T^2 orthogonally at a past endpoint and have a future endpoint if this is a singularity (on a caustic or crossing region) of B^3. Let l^μ (subject to $l^\mu{}_{;\nu}l^\nu=0$), ρ $(=-\frac{1}{2}l^\mu{}_{;\mu})$, and $|\sigma|$ $\{=[\frac{1}{2}l_{(\mu;\nu)}l^{\mu;\nu}-\frac{1}{4}(l^\mu{}_{;\mu})^2]^{1/2}\}$ be, respectively, a future-pointing tangent vector, the convergence, and the shear for these null geodesics,[10] and let A be a corresponding infinitesimal area of cross section of B^3. Then $[(A^{1/2})_{;\mu}l^\mu]_{;\nu}l^\nu = -(A^{1/2}\rho)_{;\mu}l^\mu = -A^{1/2}(|\sigma|^2+\Phi) \leq 0$ where $\Phi = -\frac{1}{2}R_{\mu\nu}l^\mu l^\nu$ $[\geq 0$ by (iv)$]$. Since T^2 is trapped, $\rho>0$ at T^2, whence A becomes zero at a finite affine distance to the future of T^2 on each null geodesic. Each geodesic thus encounters a caustic. Hence B^3 is compact (closed), being generated by a compact system of finite segments. We may approximate B^3 arbitrarily closely by a smooth, closed, spacelike hypersurface B^{3*}. Let K^4 denote the set of pairs (P,s) with $P\in B^{3*}$ and $0\leq s\leq 1$. Define a continuous map μ: $K^4 \to M_+^4$ where, for fixed P, $\mu\{(P,s)\}$ is the past geodesic segment normal to B^{3*} at $P=\mu\{(P,1)\}$ and meeting C^3 [as it must, by (iii)] in the point $\mu\{(P,0)\}$. At each point Q of $\mu\{K^4\}$, we can define the degree $d(Q)$ of μ to be the number of points of K^4 which map to Q (correctly counted). Over any region not containing the image of a boundary point of K^4, $d(Q)$ will be constant. Near B^{3*}, μ is 1-1, so $d(Q)=1$. It follows that $d(Q)=1$ near C^3 also, whence the degree of the map $B^{3*}\to C^3$ induced by μ when $s=0$ must also be unity. The impossibility of this follows from the noncompactness of C^3.

Full details of this and other related results will be given elsewhere.

---

[1]F. Hoyle and W. A. Fowler, Monthly Notices Roy. Astron. Soc. 125, 169 (1963); F. Hoyle, W. A. Fowler, G. R. Burbidge, and E. M. Burbidge, Astrophys. J. 139, 909 (1964); W. A. Fowler, Rev. Mod. Phys. 36, 545 (1964); Ya. B. Zel'dovich and I. D. Novikov, Dokl. Akad. Nauk SSSR 155, 1033 (1964) [translation: Soviet Phys.—Doklady 9, 246 (1964)]; I. S. Shklovskiĭ and N. S. Kardashev, Dokl. Akad. Nauk SSSR 155, 1039 (1964) [translation: Soviet Phys.—Doklady 9, 252 (1964)]; Ya. B. Zel'dovich and M. A. Podurets, Dokl. Akad. Nauk SSSR 156, 57 (1964) [translation: Soviet Phys.—Doklady 9, 373 (1964)]. Also various articles in the Proceedings of the 1963 Dallas Conference on Gravitational Collapse (University of Chicago Press, Chicago, Illinois, 1964).

[2]J. R. Oppenheimer and H. Snyder, Phys. Rev. 56, 455 (1939). See also J. A. Wheeler, in Relativity, Groups and Topology, edited by C. deWitt and B. deWitt (Gordon and Breach Publishers, Inc., New York, 1964); and reference 1.

[3]R. P. Kerr, Phys. Rev. Letters 11, 237 (1963).

[4]See also E. M. Lifshitz and I. M. Khalatnikov, Advan. Phys. 12, 185 (1963).

[5]See also P. G. Bergmann, Phys. Rev. Letters 12, 139 (1964).

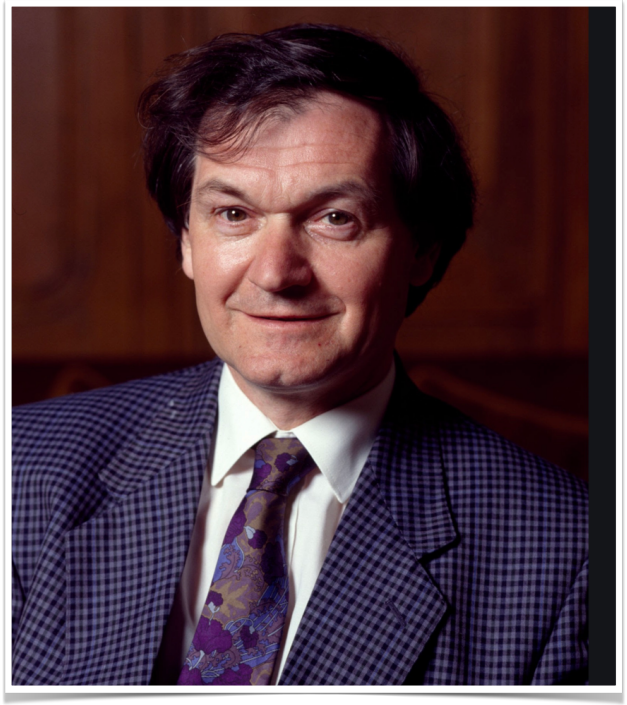
[6]The negative energy of a "C field" may be invoked to avoid singularities: F. Hoyle and J. V. Narlikar, Proc. Roy. Soc. (London) A278, 465 (1964). However, it is difficult to see how even the presence of negative energy could lead to an effective "bounce" if local causality is to be maintained.

[7]The "I'm all right, Jack" philosophy with regard to the singularities would be included under this heading!

[8]D. Finkelstein, Phys. Rev. 110, 965 (1959).

[9]The case $m<a$ is interesting in that here a singularity is "visible" to an outside observer. Whether or not "visible" singularities inevitably arise under appropriate circumstances is an intriguing question not covered by the present discussion.

[10]For the notation, etc., see E. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962).



Penrose (1965)

Turn against the tide: the "leaders" Wheeler, Lifshitz & Khalatnikov (and formerly Einstein, Eddington) expected singularities to be artefacts

The question has been raised as to whether [the Schwarzschild] singularity is, in fact, simply a property of the high symmetry assumed. The matter collapses radially inwards to the single point at the center, so that a resulting space-time catastrophe there is perhaps not surprising. Could not the presence of perturbations which destroy the spherical symmetry alter the situation drastically? (. . .) It will be shown that, after a certain critical condition has been fulfilled, deviations from spherical symmetry cannot prevent space-time singularities from arising. (. . .) It will be shown that (i)–(v) are inconsistent:

(i) Space-time (M, g) is a $4d$ time-orientable Lorentzian manifold

(ii) (M, g) is future null geodesically complete

(iii) M contains a non-compact Cauchy surface C_3

(iv) $(-R_{\mu\nu} + \frac{1}{2} g_{\mu\nu} R) t^\mu t^\nu \geq 0$ for any timelike vector t

(v) There exists a trapped surface in M

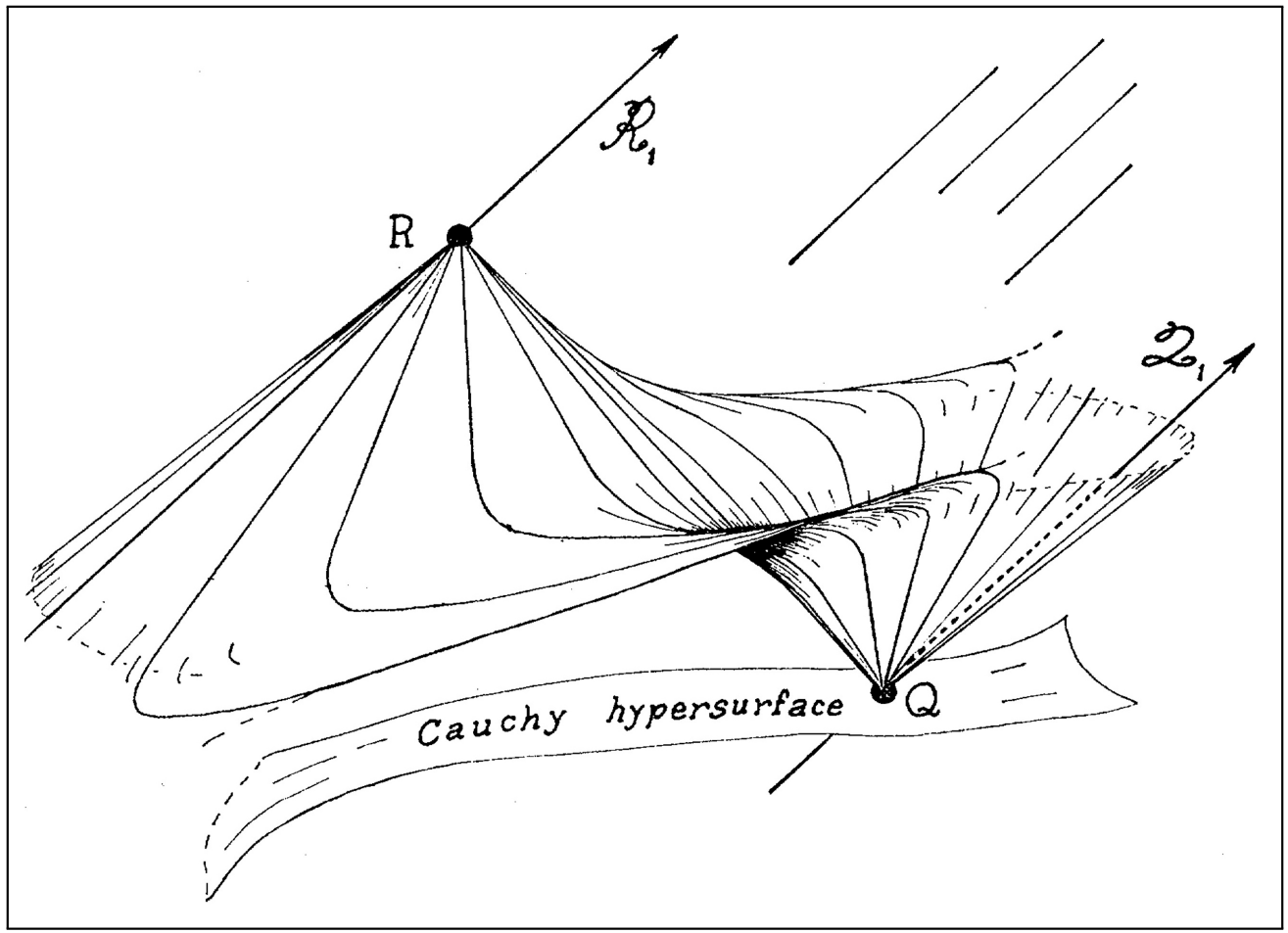
Penrose (1965): Trapped surfaces

Trapped surface one of *the* major innovations of Penrose (1965): Defined as *closed, spacelike, two-surface* T^2 *with the property that the two systems of null geodesics which meet* T^2 *orthogonally converge locally in future directions at* T^2

My conversation with [Ivor] Robinson stopped momentarily as we crossed a side road, and resumed again at the other side. Evidently, during those few moments an idea occurred to me, but then the ensuing conversation blotted it from my mind! Later in the day, after Robinson had left, I returned to my office. I remember having an odd feeling of elation that I could not account for. I began going through in my mind all the various things that had happened to me during the day, in an attempt to find what it was that had caused this elation. After eliminating numerous inadequate possibilities, I finally brought to mind the thought that I had had while crossing the street. (Penrose, quoted by Thorne, 1994)

Penrose (1965): Cauchy surfaces

Penrose (1965), following earlier 1965 paper on gravitational waves (which topic–not black holes!–originally inspired his global techniques), defines Cauchy surface C by property that every inextendible timelike or null geodesic meets C (which makes him one of the founders of concept of global hyperbolicity)



Penrose (1965): What is a singularity?

- Does not directly define what he means by a singularity
- Examples (gravitational collapse, Schwarzschild, and Kerr) suggest that he means: singularity = curvature singularity
- *Geodesic completeness* first appears in the statement of the theorem, in which it is a *reductio ad absurdum* assumption whose negation as a way out of the ensuing contradiction is left to the reader and is nowhere defined as a singularity
- Singular $\neq$ (causally) geodesically incomplete: 'The existence of a singularity can never be inferred, however, without an assumption such as [inextendibility] for the manifold under consideration.
- Among possibilities to avoid singularities despite his theorem, Penrose includes: (c) The space-time manifold is [extendible]

$\Rightarrow$ Penrose clearly recognizes that geodesic incompleteness might not be due to curvature singularity but to extendibility

Comments on Penrose (1965)

Source of singularity := incomplete causal geodesic? No:

- "Singularity" clearly always means *curvature singularity*
- Penrose argues from incompleteness to curvature singularity
- Penrose twice makes the point that although extendibility is a logically possible reason for null geodesic incompleteness, this is undesirable: the reason should be a curvature singularity

Penrose also clearly saw the gap between what his theorem actually proves and what it is often taken to prove (2020 Physics Nobel Prize!), namely the formation of a "black hole" (as yet undefined in 1965), given sufficient mass concentration (trapped surface) even in the absence of symmetry. In order to get black holes one should at least add:

1. Get rid of possible extendibility (strong cosmic censorship)
2. Existence of an event horizon (weak cosmic censorship)

The definition of a space-time singularity

Hawking claimed: singularity := incomplete causal geodesic

- any model must have a singularity, that is, it cannot be a geodesically complete C^1, piecewise C^2 manifold (1965 PhD)
- space-time is said to be singularity free if all timelike geodesics can be extended to arbitrary length (1966 PRL)
- [I] take timelike and lightlike geodesic incompleteness as our definition of a singularity of space-time (1966 Adams Prize Essay)

Timelike geodesic incompleteness has an immediate physical significance in that it presents the possibility that there could be freely moving observers or particles whose histories did not exist after (or before) a finite interval of proper time. This would appear to be an even more objectionable feature than infinite curvature and so it seems appropriate to regard such a space as singular. (. . .) The advantage of taking timelike and/or null incompleteness as being indicative of the presence of a singularity is [also] that on this basis one can establish a number of theorems about their occurrence. (Hawking & Ellis, 1973)

Impact of Penrose (1965)

After the publication of [Penrose's] paper in January 1965, the members of Dennis Sciama's general relativity group in the Department of Applied Mathematics and Theoretical Physics at Cambridge University (particularly Stephen Hawking, myself, and Brandon Carter) hurriedly tried to learn the new methods that Penrose had introduced. We were assisted in this by discussions with Felix Pirani and the group at King's College, London; with John Wheeler and Charles Misner, who visited Cambridge from the USA for an extended period; and with Roger Penrose and Bob Geroch, who was visiting Penrose at Birkbeck College, London. In particular we had a one day seminar in Cambridge attended by the members of the King's College group, where I and Brandon Carter summarized our understandings of the ingredients of Penrose's theorem. (. . .) Stephen arrived at [his] results by discussions with the Cambridge group that under Dennis Sciama's guidance met to discuss ideas at tea time each day, and with the London groups; as well as attending many seminars, we used to regularly catch the train to attend lectures on general relativity at King's College, London. (George Ellis, 2014)

Definition of a black hole

- Event horizons in GR: Rindler (1956), Finkelstein (1958)
- Penrose (1968): event horizon of observer := boundary of the chronological past of timelike curve traveled by observer
- Penrose (1969), *Gravitational collapse: The role of general relativity*, in a footnote (!): In a general space-time with a well-defined [future null infinity $\mathscr{I}^+$], the absolute event horizon would be defined as the boundary of the union of all timelike curves which escape to $\mathscr{I}^+$, [i.e.] $\dot{I}_-[\mathscr{I}^+]$ $[= \partial I^-(\mathscr{I}^+)]$ (N.B. Penrose had already introduced $\mathscr{I}^+$ in early 1960s to study gravitational waves)
- Hawking (1972), *Black holes in general relativity*: A black hole on a spacelike surface is defined to be a connected component of the region of the surface bounded by the event horizon $[= \partial J^-(\mathscr{I}^+)]$

N.B. All this defines a black *region* since it says nothing about *holes*, much as Penrose (1965) says nothing about being *black*!

Black hole should be a "hole" (singularity) inside a "black" region!

Singularities and cosmic censorship

Penrose (1974, 1979) introduced new concept of a singularity which is not quite the same as that suggested by the singularity theorems

1. Singularity related to points which it can causally influence
2. *Incomplete causal geodesics* $\rightsquigarrow$ *Inextendible causal curves*

Combination of Penrose (1979) and Geroch–Horowitz (1979):

- Space-time (M, g) assumed strongly causal and may or may not be asymptotically flat with conformal completion $(\hat{M}, \hat{g})$
- $N \subset M$ or $N \subset \hat{M}$ is some "region of exposure" (of singularity)

N-naked singularity in M := future-inextendible future-directed causal curve c in M such that $I^-(c) \subset I^-(x)$ for some $x \in N$

Explanation: if curve c has endpoint $z \in I^-(x)$, then $I^-(c) \subset I^-(x)$ iff $z \in I^-(x)$, i.e. z can signal to x. If c ends in "singularity" $z \notin M$, then $I^-(c) \subset I^-(x)$ is still taken to mean that the "singularity" can signal to x

Singularities and cosmic censorship

$N \subset M$ or $N \subset \hat{M}$: *N-naked singularity* in M := future-inextendible fd causal curve c in M such that $I^-(c) \subset I^-(x)$ for some $x \in N$

N-cosmic censorship: Space-time has no N-naked singularities

- $N = I^-(\mathscr{I}^+)$: weak cosmic censorship à la Penrose (1969)
- $N = I^-(\mathscr{I}^+) \cap I^+(\mathscr{I}^-) = \mathrm{DOC}$: weak cosmic censorship à la Chruściel–Lopes Costa (2008) in their BH uniqueness theorem
- $N = M$: strong cosmic censorship à la Penrose (1974, 1979)

Theorem (Penrose): SC holds iff (M, g) is globally hyperbolic

Help! Ambiguity! *To which space-time (M, g) should we apply this?*

- Global hyperbolicity is *false* if applied to maximal analytic solutions

Analytic solutions of 1960s made way for current PDE ideology for GR:

assumptions about initial data—theorems about their MGHD

- Global hyperbolicity is *trivial* if applied to MGHD of any initial data

Cosmic censorship: From Penrose to PDE

Consider, for motivation, an initial-data set whose maximal evolution [= MGHD] is extendible to the future of S (...) This extended spacetime cannot, by definition of the maximal evolution, have S as a Cauchy surface. That is, from a point p in the extension there must exist a maximally extended past-directed timelike curve which cannot be assigned a past endpoint, and which fails to meet S. In this rather mild sense the extended spacetime must be nakedly singular. One might therefore imagine formulating

cosmic censorship as the assertion that every maximal evolution is inextendible

i.e. that, once the maximal evolution is completed, it is not possible to add any 'extra regions' as vantage points from which observers could detect that their spacetime is singular to the future of S. (Geroch & Horowitz, 1979)

The appropriate notion of cosmic censorship is that the generic solution to Einstein's equations is globally hyperbolic, i.e. [???] that the MGHD of a generic initial data set is inextendible. (Moncrief, 1981)

This is a very strong requirement: it blocks globally hyperbolic extensions of a MGHD (i.e. with "new" Cauchy surface), which Penrose would allow

Blueshift instability

The physical evolution toward NUT space is unstable and short wavelength perturbations (gravitons, photons, etc.) are accelerated to disruptive energies before the Cauchy horizon separating the cosmological and the NUT regions is attained. This instability is shown by the same behavior of time -like and null geodesics which shows that this space-time is not geodesically complete, and that no analytic continuation of it can be. (Misner & Taub, 1969, subm. 1967

There is a further difficulty confronting our observer who tries to cross [the Cauchy horizon] $H_+(\mathscr{H})$. As he looks out at the universe he is "leaving behind," he sees, in one final flash, as he crosses $H_+(\mathscr{H})$, the entire later history of the rest of his "old universe." If, for example, an unlimited amount of matter eventually falls into the star then presumably he will be confronted with an infinite density of matter along "$H_+(\mathscr{H})$". Even if only a finite amount of matter falls in, it may not be possible in generic situations to avoid a curvature singularity in place of $H_+(\mathscr{H})$. This is at present an open question. But it may be, that the place to look for curvature singularities is in this region rather than (or as well as?) at the "center." (Penrose, 1968)

If the initial data is generically perturbed then the Cauchy horizon does not survive as a non-singular hypersurface. It is strongly implied that instead, genuine space-time singularities will appear along the region which would otherwise have been the Cauchy horizon. (Simpson & Penrose, 1973)

Summary

- Einstein, Hilbert: singularities as points *within* space-time (misled by $r = 2m$ in Schwarzschild and $R = \pi R/2$ in dS)
- Cosmological "singularity theorems" without a good definition of a singularity in 1930s–1950s (Lemaître, Raychaudhuri, . . .)
- Incomplete causal geodesics as hallmark of singularities arose in 1960s (advocated by Misner and indirectly used by Penrose in 1965 as *necessary* but not *sufficient* for a singularity):

  Hawking (1965–66) emphatically turned this into a *definition*
- Penrose (1965) notes he still needs both an event horizon and inextendibility for his theorem to be relevant to "black holes"

$\Rightarrow$ Redefinition of singularities by Penrose (1969–1979) to fill these gaps via weak and strong cosmic censorship conjectures

- In turn redefined via PDE ideology for GR based on MGHD:

strong CS as inextendibility of MGHD serves Penrose's aims well
weak CS as completeness of $\mathscr{I}^+$ for MGHD seems less close to him

Epilogue (Earman, 1995)

Prior to the 1960s spacetime singularities were regarded as an embarrassment for GR because it was thought by Einstein and others that a singularity in the fabric of spacetime itself was an absurdity. But the embarrassment was a minor one that could be swept under the rug; for the then known models of GR containing singularities all embodied very special and physically unrealistic features. Two developments forced a major shift in attitude. First, the observation of the cosmic low temperature blackbody radiation lent credence to the notion that our universe originated in a big bang singularity. Second, a series of theorems due principally to Roger Penrose and Stephen Hawking indicated that, according to GR, singularities can be expected to occur under quite general conditions, both in cosmology and in gravitational collapse. Thus, singularities cannot be swept under the rug; they are, so to speak, woven into the pattern of the rug. These theorems might have been taken as turning what was initially was only a minor embarrassment into a major scandal. Instead, what occurred was a 180° reorientation: singularities were no longer relegated to obscurity; rather they were to be recognized as a central feature of GR.