

Proc. Roy. Soc. Lond. A. **314**, 529–548 (1970)

Printed in Great Britain

# The singularities of gravitational collapse and cosmology

By S. W. HAWKING

Institute of Theoretical Astronomy, University of Cambridge

AND R. PENROSE

Department of Mathematics, Birkbeck College, London

(Communicated by H. Bondi, F.R.S.—Received 30 April 1969)

A new theorem on space-time singularities is presented which largely incorporates and generalizes the previously known results. The theorem implies that space-time singularities are to be expected if either the universe is spatially closed or there is an ‘object’ undergoing relativistic gravitational collapse (existence of a trapped surface) or there is a point  $p$  whose past null cone encounters sufficient matter that the divergence of the null rays through  $p$  changes sign somewhere to the past of  $p$  (i.e. there is a minimum apparent solid angle, as viewed from  $p$  for small objects of given size). The theorem applies if the following four physical assumptions are made: (i) Einstein’s equations hold (with zero or negative cosmological constant), (ii) the energy density is nowhere less than minus each principal pressure nor less than minus the sum of the three principal pressures (the ‘energy condition’), (iii) there are no closed timelike curves, (iv) every timelike or null geodesic enters a region where the curvature is not specially aligned with the geodesic. (This last condition would hold in any sufficiently general physically realistic model.) In common with earlier results, timelike or null geodesic incompleteness is used here as the indication of the presence of space-time singularities. No assumption concerning existence of a global Cauchy hypersurface is required for the present theorem.

## 1. INTRODUCTION

An important feature of gravitation, for very large concentrations of mass, is that it is essentially *unstable*. This is due, in the first instance, to its  $r^{-2}$  attractive character. But, in addition, when general relativity begins to play a significant role, other instabilities may also arise (cf. Chandrasekhar 1964). The instability of gravitation is not manifest under normal conditions owing to the extreme smallness of the gravitational constant. The pull of gravity is readily counteracted by other forces. However, this instability *does* play an important dynamical role when large enough concentrations of mass are present. In particular, as the work of Chandrasekhar (1935) showed, a star of mass greater than about 1.3 times that of the Sun, which has exhausted its resources of thermal and nuclear energy, cannot sustain itself against its own gravitational pull, so a *gravitational collapse* ensues. It has sometimes been suggested also that, on a somewhat larger scale, some form of gravitational collapse may be taking place in quasars, or perhaps in the centres of (some?) galaxies. Finally, on the scale of the universe as a whole, this instability shows up again in those models for which the expansion eventually reverses, and the entire universe becomes involved in a gravitational collapse. In the reverse direction in time there is also the ‘big bang’ initial phase which is common

to most relativistic expanding models. This again may be regarded as a manifestation of the instability of gravitation (in reverse).

But what is the ultimate fate of a system in gravitational collapse? Is the picture that is presented by symmetrical exact models accurate, according to which a *singularity in space-time* would ensue? Or may it not be that any asymmetries present might cause the different parts of the collapsing material to miss each other, so possibly to lead to some form of *bounce*? It seems that until comparatively recently many people had believed that such an asymmetrical bounce might indeed be possible to achieve, in a manner consistent with general relativity (cf. particularly, Lindquist & Wheeler 1957; Lifshitz & Khalatnikov 1963). However, some recent theorems† (Penrose 1965a; Hawking 1966a, b; H; Geroch 1966) have ruled out a large number of possibilities of this kind. The present paper carries these results further, and considerably strengthens the implication that a singularity-free bounce (of the type required) does not seem to be realizable within the framework of general relativity.

In the first theorem (referred to as I; see Penrose 1965a; cf. also Penrose 1966; P; Hawking 1966c) the concept of the existence of a *trapped surface*‡ was used as a characterization of a gravitational collapse which has passed a ‘point of no return’. On the basis of a *weak energy condition*, ‡ the intention was to establish the existence of space-time singularities from the existence of a trapped surface. Unfortunately, however, theorem I required, as an additional hypothesis, the existence of a non-compact global Cauchy hypersurface. Although ‘reasonable’ from the point of view of classical Laplacian determinism, the assumption of the existence of a global Cauchy hypersurface is hard to justify from the standpoint of general relativity. Also, it is violated in a number of exact models. Furthermore, the non-compactness assumption used in theorem I applies only if the universe is ‘open’.

The second theorem (Hawking 1966a), and its improved version (referred to as II, see H; cf. also Hawking (1966c) and P), required the existence of a compact spacelike hypersurface with everywhere diverging normals. Thus it applies to ‘closed’, everywhere expanding, universe models. For such models II implies the existence of an initial (e.g. ‘big bang’ type) singularity. However, this condition on the normals may well not be applicable to the actual universe (particularly if there are local collapsing regions), even if the universe is ‘closed’. Also, the condition is virtually unverifiable by observation.

The third and fourth results (referred to as III and IV; see Geroch (1966) and Hawking (1966b), respectively) again apply to ‘closed’ universe models (i.e. containing a compact, spacelike hypersurface), but which do not have to be assumed to be everywhere expanding. However, III required the somewhat unnatural assumption of the non-existence of ‘horizons’, while IV required that the given compact hypersurface be a global Cauchy hypersurface. Thus, III and IV could be objected to on grounds similar to those of I.

† We use H for referring to Hawking (1967) and P for referring to Penrose (1968).

‡ The precise meanings of these terms will be given in §3.

The fifth theorem (referred to as V; see **H**, also Hawking (1966c) and **P**) does not suffer from objections of this kind, but the requirement on which it was based—namely that the divergence of all timelike and null geodesics through some point  $p$  changes sign somewhere to the past of  $p$ —is somewhat stronger than one would wish. Theorem V would be considerably more useful in application if the above requirement referred only to *null* geodesics.

In this paper we establish a new theorem, which, with two reservations, effectively incorporates all of I, II, III, IV and V while avoiding each of the above objections. In its physical implications, our theorem falls short of completely superseding these previous results only in the following two main respects. In the first instance we shall require the non-existence of closed time like curves. Theorem II (and II alone) did not require such an assumption. Secondly, in common with II, III, IV and V, we shall require the slightly stronger energy condition given in (3.4), than that used in I. This means that our theorem cannot be directly applied when a *positive cosmological constant*  $\lambda$  is present. However, in a collapse, or ‘big bang’, situation we expect large curvatures to occur, and the larger the curvatures present the smaller is the significance of the value of  $\lambda$ . Thus, it is hard to imagine that the value of  $\lambda$  should qualitatively affect the singularity discussion, except in regions where curvatures are still small enough to be comparable with  $\lambda$ . We may take I as a further indication (though not a proof) of this. In a similar way, II may be taken as a strong indication that the development of closed timelike curves is not the ‘answer’ to the singularity problem. Of course, such causality violation would carry with it other very serious problems, in any case.

The energy condition (3.4) used here (and in II, III, IV and V) has a very direct physical interpretation. It states, in effect, that ‘gravitation is always attractive’ (in the sense that neighbouring geodesics near any one point accelerate, on the average, towards each other). Our theorem will apply, in fact, in theories other than classical general relativity provided gravitation remains attractive. In particular, we can apply our results in the theory of Brans & Dicke (1961), using the metric for which the field equations resemble Einstein’s (cf. Dicke 1962). The gravitational constant could, in principle, change sign in this theory, but only via a region at which it becomes infinite. Such a region could reasonably be called a ‘singularity’ in any case. On the other hand, gravitation does not always remain attractive in the theory of Hoyle & Narlikar (1963) (owing to the effective negative energy of the C-field) so our theorem is not directly applicable in this theory. We note, finally, that in Einstein’s theory (with ‘reasonable’ sources) it is only  $\lambda > 0$  which can prevent gravitation from being always attractive, the  $\lambda$  term representing a ‘cosmic repulsion’.

In common with all the previous results I, . . . , V, our theorem will not give very much information as to the nature of the space-time singularities that are to be inferred on the basis of Einstein’s theory. If we accept that ‘causality breakdown’ is unlikely to occur (because of philosophical difficulties encountered with closed timelike curves and because theorem II suggests that such curves probably do not

help in the singularity problem in any case), then we are led to the view that the instability of gravitation presumably† results in regions of enormously large curvature occurring in our universe. These curvatures would have to be so large that our present concepts of local physics would become drastically modified. While the quantum effects of gravitation are normally thought to be significant only when curvatures approach  $10^{33} \text{ cm}^{-1}$ , all our local physics is based on the Poincaré group being a good approximation of a local symmetry group at dimensions greater than  $10^{-13} \text{ cm}$ . Thus, if curvatures ever even approach  $10^{13} \text{ cm}^{-1}$ , there can be little doubt but that extraordinary local effects are likely to take place.

When a singularity results from a collapse situation in which a trapped surface has developed, then any such local effects would not be observable outside the collapse region. It is an open question whether physically realistic collapse situations, resulting in singularities, will sometimes arise *without* trapped surfaces developing (cf. Penrose 1969). If they do, it is likely that such singularities could (in principle) be observed from outside. Of course, the initial ‘big bang’ singularity of the Robertson–Walker models is an example of a singularity of the observable type. However, our theorem yields no information as to the observability of singularities in general. We cannot even rigorously infer whether the implied singularities are to be expected in the ‘past’ or the ‘future’. (In this respect our present theorem yields somewhat less information than I, II, or V.)

Our theorem will be directly applicable to *any one* of the following three situations. First, to the existence of a trapped surface; secondly, to the existence of a compact space-like hypersurface; thirdly, to the existence of a point whose null-cone begins to ‘converge again’ somewhere to the past of the point. We assume the energy condition and the non-existence of closed timelike curves. On the basis of this (and another very minor assumption which merely rules out some highly special models) we deduce that singularities will develop in fully general situations involving a collapsing star, *or* in a spatially closed universe, *or* (taking the point in question in the third case to be the earth at the present time) if the apparent solid angle subtended by an object of a given intrinsic size reaches some minimum when the object is at a certain distance from us. We show, in an appendix, that this last condition is indeed likely to be satisfied in our universe, assuming the correctness of the normal interpretation of the 2.7 K background radiation. A similar discussion was given earlier by Hawking & Ellis (1968) in connexion with theorem V. Since we now have a stronger theorem, we can use somewhat weaker physical assumptions concerning the radiation.

In §2 we give a number of lemmas and definitions that will be needed for our theorem. The precise statement of the theorem will be given in §3. This statement

† We must always bear in mind that a local ‘energy-condition’ (cf. (3.4)) is being assumed here, which might be violated not only in a modified Einstein theory (e.g. ‘C-field’), but also in the standard theory if we were allowed to have very ‘peculiar’ matter under extreme conditions. The quantum field-theoretic requirement of positive-definiteness of energy (in order that the vacuum remain stable) is of great relevance here, but its status is perhaps not completely clear (cf. Sexl & Urbantke 1967 for example).

is presented in a rather general form, which is somewhat removed from the actual applications. The main applications are given in a corollary to the theorem. One slight advantage of the form of statement that we have chosen will be that it enables a small amount of information to be extracted about the actual nature of the singularities. This is that (at least) one timelike or null geodesic must enter (or leave) the singularity not only in a finite proper (or affine) time, but also in such a way that none of the neighbouring initially parallel geodesics has time to be focused towards it before the singularity is encountered.

## 2. DEFINITIONS AND LEMMAS

A four-dimensional differentiable (Hausdorff and paracompact<sup>†</sup>) manifold  $M$  will be called a *space-time* if it possesses a pseudo-Riemannian metric of hyperbolic normal signature  $(+, -, -, -)$  and a time-orientation. (In fact the following arguments will apply equally well if  $M$  has any dimension  $\geq 3$ ; also, the time-orientability of  $M$  need not really be assumed if we are prepared to apply the arguments to a twofold covering of  $M$ .) There will be no real loss of generality in physical applications if we assume that  $M$  and its metric are both  $C^\infty$ . However, the arguments we use actually only require the metric to be  $C^2$ .

We shall be concerned with *timelike curves* and *causal curves* on  $M$ . (When we speak of a ‘curve’, we shall, according to context, mean either a continuous map into  $M$  of a connected closed portion of the real line, or else the image in  $M$  of such a map.) For definiteness we choose our timelike curves to be *smooth*, with future-directed tangent vectors everywhere strictly timelike, including at its end-points. A causal curve is a curve obtainable as a limiting case of timelike curves<sup>‡</sup> (cf. Siefert 1967; Carter 1967); it is continuous but not necessarily everywhere smooth; where smooth, its tangent vectors are either timelike or null. A timelike or causal curve will require end-points if it can be extended as a causal curve either into the past or the future (cf. P, p. 187). If it continues indefinitely into the past [*resp.* future] it will be called *past-inextendible* [*resp.* *future-inextendible*]. If both past- and future-inextendible it is called *inextendible*.

If  $p, q \in M$ , we write  $p \ll q$  if there is a timelike curve with past end-point  $p$  and future end-point  $q$ ; we write  $p \prec q$  if either  $p = q$  or there is a causal curve from  $p$  to  $q$  (cf. Kronheimer & Penrose 1967). If  $p \prec q$  but not  $p \ll q$ , then there is a null geodesic from  $p$  to  $q$ , or else  $p = q$ . If  $p \ll q$  and  $q \prec r$ , or if  $p \prec q$  and  $q \ll r$ , then  $p \ll r$ . We do not have  $p \ll p$  unless  $M$  contains *closed timelike curves*. A subset of  $M$  is called *achronal* if it contains no pair of points  $p, q$  with  $p \ll q$ .

<sup>†</sup> Geroch (1968b) has shown that the assumption of paracompactness is not actually necessary for a space-time, being a consequence of the other assumptions for a space-time manifold.

<sup>‡</sup> Except for very minor parts of our discussion, the fact that we are allowing our causal curves not to be smooth plays no significant role in this paper, but it is useful for the general theory. A continuous map of the connected closed interval  $\Gamma \subset \mathbb{R}$ , into  $M$ , can be characterized as a causal curve by the fact that if  $[a, b] \in \Gamma$  and if  $A, B$  and  $C$  are neighbourhoods in  $M$  of the images of  $a, b$  and  $[a, b]$ , respectively, then there exists a timelike curve lying in  $C$  with one end-point in  $A$  and another end-point in  $B$ .

We shall, for the most part, use terminology, definitions and some basic results as given in **P**. (However we use ‘causal’ for curves referred to in **P** as ‘nonspace-like’ and ‘achronal’ for sets referred to in **P** as ‘semispace-like’; cf. Carter 1967.) As in Kronheimer & Penrose (1967), we write  $I^+(p)$  for the open future of a point  $p \in M$ , i.e.  $I^+(p) = \{x: p \ll x\}$  and  $I^+[S]$  for the open future of a set  $S \subset M$ , i.e.  $I^+[S] = \bigcup_{p \in S} I^+(p)$ . (The sets  $I^+[S]$  are open in the manifold topology for  $M$ .) Similarly,  $J^+(p) = \{x: p \prec x\}$ ;  $J^+[S] = \bigcup_{p \in S} J^+(p)$ . These are not always closed sets.) We define

$$E^+(S) = J^+[S] - I^+[S]. \quad (2.1)$$

Then  $E^+(S)$  is part of the boundary  $\dot{I}^+[S]$  of  $I^+[S]$  but not necessarily all of it. The sets  $I^-(p)$ ,  $I^-[S]$ ,  $J^-(p)$ ,  $J^-[S]$  and  $E^-(S)$  are defined similarly, but with future and past interchanged.

For any set  $S \in M$  we can define the (*future*) domain of dependence  $D^+(S)$  and Cauchy horizon  $H^+(S)$  by

$$D^+(S) = \{x: \text{every past-inextendible timelike curve through } x \text{ meets } S\} \quad (2.2)$$

and

$$\begin{aligned} H^+(S) &= \{x: x \in D^+(S), I^+(x) \cap D^+(S) = \emptyset\} \\ &= D^+(S) - I^-[D^+(S)]. \end{aligned} \quad (2.3)$$

The sets  $D^-(S)$  and  $H^-(S)$  are correspondingly defined. (These definitions are chosen to agree with **P**; they differ somewhat from those of **H**.) We shall be concerned only with the cases when  $S$  is an *achronal closed set*. Then  $D^+(S)$  is a closed set and  $H^+(S)$  is an *achronal closed set*. One easily verifies:

$$I^+[H^+(S)] = I^+[S] - D^+(S). \quad (2.4)$$

Define the *edge* of an achronal closed set  $S$  to be the set of points  $p \in S$  such that† if  $r \ll p \ll q$ , with  $\gamma$  a timelike curve from  $r$  to  $q$ , containing  $p$ , then every neighbourhood of  $\gamma$  contains a timelike curve from  $r$  to  $q$  not meeting  $S$ . It follows that  $\text{edge}(S)$  is in fact the set of points in whose vicinity  $S$  fails to be a  $C^\circ$ -manifold ( $S$  achronal and closed). We have (cf. **P**, p. 191)  $\text{edge}(S) \subset H^+(S)$ . (In fact  $\text{edge}(S) = \text{edge}(H^+(S))$ .) Furthermore:

**LEMMA (2.5).** *Every point of  $H^+(S) - \text{edge}(S)$  is the future end-point of a null geodesic on  $H^+(S)$  which can be extended into the past on  $H^+(S)$  either indefinitely, or until it meets  $\text{edge}(S)$ .*

For the proof, see **P**, p. 217 (compare **H**).

A similar result (which follows at once from **P**, p. 216; **H**) is (with  $S$  closed and achronal).

**LEMMA (2.6).** *Every point  $p \in \dot{I}^+[S] - S$  is the future end-point of a null geodesic on  $\dot{I}^+[S]$  which can be extended into the past on  $\dot{I}^+[S]$  either indefinitely (if  $p \in \dot{I}^+[S] - E^+(S)$ ) or until it meets  $\text{edge}(S)$  (whence  $p \in E^+(S)$ ).*

We say that *strong causality* holds at  $p$  if arbitrarily small neighbourhoods of  $p$  exist, each intersecting no timelike curve in a disconnected set. (Roughly speaking,

† This replaces the definition of  $\text{edge}(S)$  given in **P**, which was not quite correctly stated.

this means that timelike curves cannot leave the vicinity of  $p$  and then return to it; i.e.  $M$  does not 'almost' contain closed timelike curves.) We must say 'arbitrarily small', rather than 'every', in the above definition because of the existence of 'hour-glass shaped' (or even 'ball shaped') neighbourhoods of any point in *any* space-time, which are left and re-entered by a timelike curve. To avoid this feature, let us call an open set  $Q$  *causally convex* (**P**, p. 224) if  $Q$  intersects no timelike curve in a disconnected set. Thus, strong causality holds at  $p$  if and only if  $p$  possesses arbitrarily small causally convex neighbourhoods (in which case, the 'Alexandrov neighbourhoods'  $I^+(q) \cap I^-(r)$  will suffice, with  $q \ll p \ll r$ ). A causally convex open set which lies inside a convex normal coordinate ball with compact closure† will be called a *local causality neighbourhood* (**H**, p. 192). Strong causality holds at every point of a local causality neighbourhood. The only properties of a local causality neighbourhood that we shall in fact use, are that it is open and causally convex, that it contains no past- (or future) -inextendible null geodesic and that any point at which strong causality holds possesses such a neighbourhood.

A property of  $D^+(S)$  we shall require is the following. Again,  $S$  is to be achronal and closed.

**LEMMA (2.7).** *If  $p \in \text{int } D^+(S)$ , then  $J^-(p) \cap J^+[S]$  is compact.*

This follows from **H**. (See also **P**, p. 227: if  $\text{edge}(S) = \emptyset$ , and strong causality holds at each point‡ of  $S$ , we have the stronger result that  $\text{int } D^+(S)$  is precisely the set of  $p \in I^+[S]$  for which  $J^-(p) \cap J^+[S]$  is both compact and contains no point at which strong causality fails. Lemma (2.7) follows by similar reasoning.)

We shall require the concept of *conjugate points* on a causal (i.e. timelike or null) geodesic. Two points  $p$  and  $q$  on a causal geodesic  $\gamma$  are said to be *conjugate* if a geodesic 'neighbouring' to  $\gamma$  'meets'  $\gamma$  at  $p$  and at  $q$ . Somewhat more precisely, the congruence of geodesics through  $p$  in the neighbourhood of  $\gamma$  has  $q$  as a *focal* point, that is, a point where the divergence of the congruence becomes infinite. (This focal point will in general be an 'astigmatic' focal point. It is a point of the 'caustic' of the congruence. Precise definitions of conjugate points will be found in Milnor (1963), Hicks (1965), Hawking (1966a).) The relation of conjugacy is symmetrical in  $p$  and  $q$ . The above definition still holds if the roles of  $p$  and  $q$  are reversed. The property of conjugate points that we shall require is the following (for the timelike case, see Boyer (1964), Hawking (1966a, c), cf. Milnor (1963); for the null case see Hawking (1966c) and also **P**, p. 215, for an equivalent result).

**LEMMA (2.8).** *If a causal geodesic  $\gamma$  from  $p$  to  $q$  contains a pair of conjugate points between  $p$  and  $q$ , then there exists a timelike curve from  $p$  to  $q$  whose length exceeds that of  $\gamma$ .*

We use the term 'length' for a causal curve to denote its proper time integral. A timelike geodesic is *locally* a curve of maximum length. As a corollary of lemma (2.8) we have:

† This condition was not explicitly included in the definition given in **H**.

‡ This condition should have been included in the conditions on  $\mathcal{H}$  in lemma  $V$  of **P**.

**LEMMA (2.9).** *If  $\gamma$  is a null geodesic lying on  $I^+[S]$  or on  $H^+(S)$  for some  $S \subset M$ , then  $\gamma$  cannot contain a pair of conjugate points except possibly at its end-points.*

Another consequence of lemma (2.8) is the following result:

**LEMMA (2.10).** *If  $M$  contains no closed timelike curves and if every inextendible null geodesic in  $M$  possesses a pair of conjugate points, then strong causality holds throughout  $M$ .*

*Proof.* The result has been given in Hawking (1966c). We repeat the argument here since this reference is not readily available. Suppose strong causality fails at  $p$ . Let  $B$  be a normal coordinate neighbourhood of  $p$  and  $Q_i$  a nested sequence of neighbourhoods of  $p$  converging on  $p$ . Now there is a timelike curve originating in  $Q_i$  which leaves  $B$  at a point  $q_i \in \dot{B}$ , re-enters  $B$  and returns to  $Q_i$ . As  $i \rightarrow \infty$  the  $q_i$  have an accumulation point  $q$  on  $\dot{B}$  ( $\dot{B}$  being compact). The geodesic  $pq$  in  $B$  cannot be timelike (since otherwise  $I^-(q)$  would contain some  $Q_i$ , so closed timelike curves would result), nor spacelike. It must therefore be null. Furthermore, strong causality must also fail at  $q$ . Repeating the argument with  $q$  in place of  $p$ , we obtain a new null geodesic  $qr$ . In fact this must be the continuation of  $pq$ , since otherwise closed timelike curves would result. Continuing the process indefinitely both into the future and into the past we get an inextendible null geodesic  $\gamma$  at every point of which strong causality must fail. By hypothesis  $\gamma$  contains a pair of conjugate points. Thus by lemma (2.8) two of its points can be connected by a timelike curve. It follows that each point of some neighbourhood of one of these points can be joined by a timelike curve to each point of some neighbourhood of the other. This leads at once to the existence of closed timelike curves (because of strong causality violation), contrary to hypothesis. This establishes the lemma.

An important consequence of strong causality is the following result.

**LEMMA (2.11).** *Let  $p \ll q$  be such that the set  $J^+(p) \cap J^-(q)$  is compact and contains no points at which strong causality fails. Then there is a timelike geodesic from  $p$  to  $q$  which attains the maximum length for timelike curves connecting  $p$  to  $q$ .*

This result was proved by Siefert (1967). The result is, in effect, also contained in the earlier work of Avez (1963). (Unfortunately Avez's analysis contains some errors owing to the fact that the possibility of strong causality breakdown is not duly taken into account.) Lemma (2.11) follows also from lemma  $V$  in  $\mathbf{P}$  (p. 227) in conjunction with  $VI$  of  $\mathbf{P}$  (p. 228), as applied to the closed achronal set  $I^-(q)$ . In fact, lemma (2.11) can be generalized: if  $C$  is a compact subset of  $M$  containing no points at which strong causality fails, then the maximum length for all timelike curves contained in  $C$  is attained (though not necessarily by a geodesic). The essential feature of this situation is that the space of causal curves contained in  $C$  is compact, the length of a causal curve being an upper semi-continuous function of the curve. For this, we need the appropriate topology on the space of causal curves. (See Seifert (1967); cf. also Avez (1963)). But it will not be necessary to enter into the general discussion here, as lemma (2.11) is all we shall need.

We define a *future-trapped* [resp. *past-trapped*] set to be a non-empty achronal

closed† set  $S \subset M$  for which  $E^+(S)$  [resp.  $E^-(S)$ ] is compact. (Note that  $E^+(S)$  [resp.  $E^-(S)$ ] must then be a closed achronal set.) Any future-trapped set  $S$  must itself be compact, since  $S \subset E^+(S)$ .) An example of a future-trapped set is illustrated in figure 1. We now come to our main lemma.

**LEMMA (2.12).** *If  $S$  is a future-trapped set for which strong causality holds at every point of  $\bar{I}^+[S]$ , then there exists a future-inextendible timelike curve  $\gamma \subset \text{int } D^+(E^+(S))$ .*

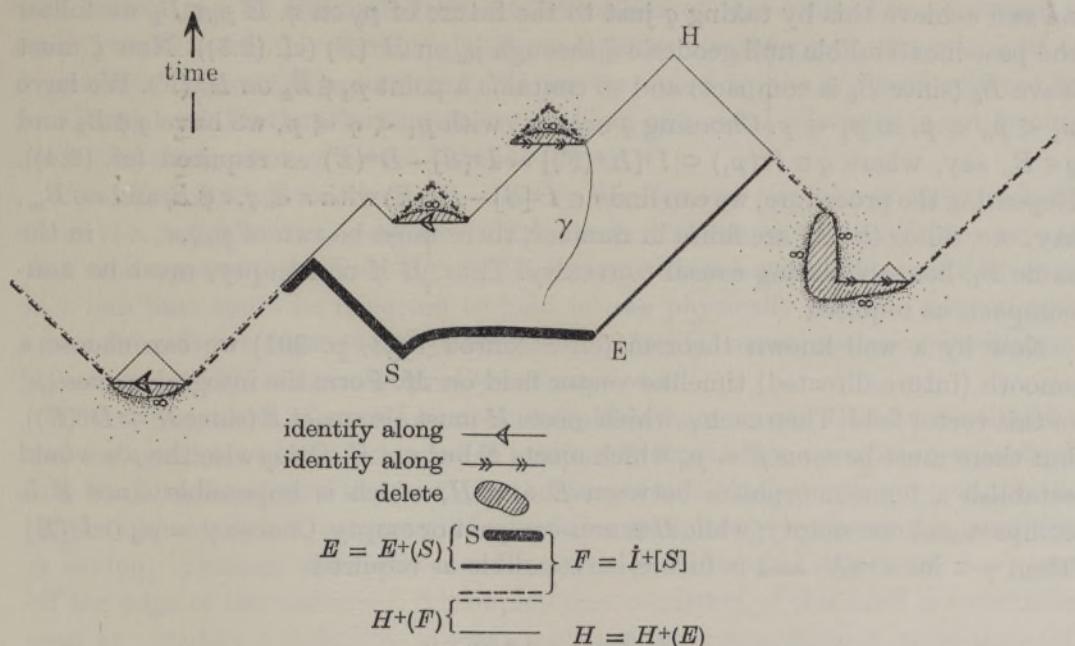


FIGURE 1. A future-trapped set  $S$ , together with the associated achronal sets  $E = E^+(S)$ ,  $F = \bar{I}^+[S]$ ,  $H^+(F)$ ,  $H = H^+(E)$ . (For the proof of lemma (2.12).) The figure is drawn according to the conventions whereby null lines are inclined at  $45^\circ$ . The diagonally shaded portions are excluded from the space-time and some identifications are made. The symbol  $\infty$  indicates regions 'at infinity' with respect to the metric. A future-inextendible timelike curve  $\gamma \in D^+(E)$  is depicted, in agreement with the conclusion of lemma (2.12).

*Proof.*‡ We first make some remarks concerning the relation between  $E = E^+(S)$  and  $F = \bar{I}^+[S] = \bar{I}^+[E]$ , and between their domains of dependence and their Cauchy horizons. We have  $E \subset F$ , whence  $D^+(E) \subset D^+(F)$ . We have  $\text{edge}(F) = \emptyset$ , so it follows from lemma (2.5) that each point of  $F - E$  lies on a past-inextendible null geodesic on  $F - E$ . (These null geodesics extend into the future, while remaining

† The condition that  $S$  be closed could be omitted from this definition if desired. For, if  $S$  is achronal with  $E^+(S)$  compact, then  $E^+(S) = E^+(\bar{S})$ . Another apparent weakening of the definition of 'future-trapped' for a closed achronal non-empty set  $S$  would be to say that  $E^+(S)$  has compact closure. ( $E^+(S)$  is not always a closed set, for general  $S$ .) This definition would be equivalent to the one we use, provided strong causality holds.

‡ This argument follows, to some extent, one given in **H** (pp. 198–9). It may also serve as a replacement for the final argument given in **P** (on p. 230) which was not stated correctly.

on  $F-E$ , perhaps reaching a future end-point on  $\text{edge}(E)$ . We readily obtain  $D^+(F) - D^+(E) = H^+(F) - H^+(E) = F - E$ , so  $\text{int } D^+(E) = \text{int } D^+(F)$ .)

We shall show that  $H = H^+(E)$  is non-compact or empty. For, suppose  $H$  is compact. Then we can cover  $H$  with a finite number of local causality neighbourhoods  $B_i$ . If  $H$  is non-empty, then  $D^+(E) \neq I^+[S]$ . Let  $p \in I^+[S] - D^+(E)$  with  $p$  near  $H$  and suppose  $p \in B_k$ . Since  $p \in I^+[S]$ , a timelike curve  $\eta$  exists connecting  $S$  to  $p$ . Since  $p \notin D^+(E)$ , it follows that  $\eta$  meets  $H$  at a point  $p_0$ , say. We wish to construct a point  $q \in I^+[S] - D^+(E)$  with  $q \ll p$ ,  $q \notin B_k$  and  $q \in B_l$ , say. If  $p_0 \notin \bar{B}_k$  we can achieve this by taking  $q$  just to the future of  $p_0$  on  $\eta$ . If  $p_0 \in \bar{B}_k$  we follow the past-inextendible null geodesic  $\zeta$  through  $p_0$  on  $H^+(F)$  (cf. (2.5)). Now  $\zeta$  must leave  $\bar{B}_k$  (since  $\bar{B}_k$  is compact) and so contains a point  $p_1 \notin \bar{B}_k$  on  $H^+(F)$ . We have  $p_1 < p_0 \ll p$ , so  $p_1 \ll p$ . Choosing  $q$  near  $p_1$ , with  $p_1 \ll q \ll p$ , we have  $q \notin B_k$  and  $q \in B_l$ , say, where  $q \in I^+(p_1) \subset I^+[H^+(F)] = I^+[S] - D^+(E)$  as required (cf. (2.4)). Repeating the procedure, we can find  $r \in I^+[S] - D^+(E)$  with  $r \ll q$ ,  $r \notin B_l$  and  $r \in B_m$ , say, etc. Since the  $B_i$  are finite in number, there must be two of  $p, q, r, \dots$ , in the same  $B_i$ , hence violating causal convexity. Thus,  $H$  if non-empty, must be non-compact, as required.

Now by a well known theorem (cf. Steenrod 1951, p. 201) we can choose a smooth (future-directed) timelike vector field on  $M$ . Form the integral curves  $\{\mu\}$  of this vector field. Then each  $\mu$  which meets  $H$  must also meet  $E$  (since  $H \subset D^+(E)$ ), but there must be some  $\mu = \mu_0$  which meets  $E$  but not  $H$ . Otherwise the  $\mu$ 's would establish a homeomorphism between  $E$  and  $H$ , which is impossible since  $E$  is compact and non-empty, while  $H$  is non-compact or empty. Choose  $\gamma = \mu_0 \cap I^+[E]$ . Then  $\gamma \subset \text{int } D^+(E)$  and is future-inextendible as required.

### 3. THE THEOREM

We shall begin by giving a precise statement of our theorem. The form of statement we adopt is made primarily for the sake of generality and for certain mathematical advantages. But in order that the theorem may be directly applied to physical situations, we single out the main special cases of interest in a corollary. This recasts our main result in a much more suggestive and immediately usable form. However, the generality of the statement given in the theorem will also yield some advantages as regards applications. It will enable a small amount of information to be extracted as to the actual nature of the space-time singularities. Also, it is by no means impossible that the theorem, as stated, may have relevance in physical situations other than precisely those which we have considered here. We shall follow the statement of the theorem with some explanations and interpretations.

**THEOREM.** *No space-time  $M$  can satisfy all of the following three requirements together:*

- (3.1)  *$M$  contains no closed timelike curves,*
- (3.2) *every inextendible causal geodesic in  $M$  contains a pair of conjugate points,*
- (3.3) *there exists a future- (or past-) trapped set  $S \subset M$ .*

Let us examine each of these three conditions in turn. With regard to (3.1), the existence of closed timelike curves in any space-time model leads to very severe interpretative difficulties. It might perhaps be argued that the presence of a closed timelike world-line could be admissible, provided the world-line entered a region of such extreme physical conditions, or involved such large accelerations, that no physical observer could 'survive' making this trip into his own past, so that any 'memory' of events would necessarily be destroyed in the course of the trip. However, it seems highly unlikely that the physical consequences of closed timelike curves can be eliminated by considerations of this kind. The existence of such curves can imply serious global consistency conditions on the solutions of hyperbolic differential equations.<sup>†</sup> We are reassured by the theorem referred to as II in § 1 (cf. H) that the singularity problem of general relativity is not forcing us into consideration of closed timelike curves.

Condition (3.2) of the theorem—namely that for any timelike or null geodesic, there is a 'neighbouring geodesic' which meets it at two distinct points—may, at *first sight* appear to be a strong one. However, this is not so. The condition is in fact one that could be expected to hold in *any* physically realistic non-singular space-time. It is a consequence of three requirements: *causal geodesic completeness*, the *energy condition* and a *generality assumption*.

The requirement of *causal geodesic completeness* is simply that every timelike and null geodesic can be extended to arbitrarily large affine parameter value both into the future and into the past. (In the case of timelike geodesics we can use the proper time as such a parameter.) In crude terms we could interpret this condition as saying: 'photons and freely moving particles cannot just appear or disappear off the edge of the universe'. A completeness condition of this kind is sometimes used as virtually a *definition* of what is meant by a non-singular space-time (cf. Geroch 1968a). Since one must normally 'delete' any actual singular points from consideration as part of the space-time manifold, it is by some criterion such as 'incompleteness' that the 'holes' left by the removal of the singularities may be detected.

The *energy condition* may be expressed as

$$t^a t_a = 1 \text{ implies } R_{ab} t^a t^b \leq 0. \quad (3.4)$$

(We use a +--- signature, with Riemann and Ricci tensor signs fixed by  $2\nabla_{[a}\nabla_{b]}k_c = k_d R_{aab}^d$ ,  $R_{ab} = R_{acb}^c$ .) With Einstein's equations

$$R_{ab} - \frac{1}{2}Rg_{ab} = -KT_{ab}, \quad (3.5)$$

(3.4) becomes

$$t^a t_a = 1 \text{ implies } T_{ab} t^a t^b \geq \frac{1}{2}T_c^c. \quad (3.6)$$

(We have  $K > 0$ . To incorporate a cosmological constant  $\lambda$ , we would have to replace  $T_{ab}$  in the above by  $T_{ab} + \lambda K^{-1}g_{ab}$ . Thus, (3.6), as it stands, would still

<sup>†</sup> For example,  $\phi = \text{const.}$  is the only solution of  $\partial^2\phi/\partial t^2 - \partial^2\phi/\partial x^2 = 0$ , on the  $(x, t)$ -torus, for which  $(t, x)$  is identified with  $(t+n, x+m\pi)$  for each pair of integers  $n, m$ .

imply (3.4) so long as  $\lambda \leq 0$ .) [If, in an eigentetrad of  $T_{ab}$ ,  $E$  denotes the *energy density* and  $p_1, p_2, p_3$  denote the three *principal pressures*, then (3.6) can be written as

$$E + \Sigma p_i \geq 0, \quad (3.7)$$

together with

$$E + p_i \geq 0, \quad (3.8)$$

where  $i = 1, 2, 3$ .

The *weak energy condition* is

$$l^a l_a = 0 \text{ implies } R_{ab} l^a l^b \leq 0, \quad (3.9)$$

which is a consequence of (3.4) (as follows by a limiting argument). This is equivalent, assuming Einstein's equations, to (3.8) (*without* (3.7)) and follows from the *positive-definiteness* of the energy expression  $T_{ab} t^a t^b$ , for  $t^a t_a = 1$ . (This is now irrespective of the value of  $\lambda$ .)

The assumption of *generality* we require (compare Hawking 1966*b*) is that every causal geodesic  $\gamma$  contains some point for which

$$k_{[a} R_{b]cd[e} k_{f]} k^c k^d \neq 0, \quad (3.10)$$

where  $k_a$  is tangent to  $\gamma$ . If  $\gamma$  is timelike, we can rewrite (3.10) as

$$R_{abcd} k^b k^c \neq 0. \quad (3.11)$$

(To see this, transvect (3.10) with  $k^a k^f$ .)

In any physically realistic 'generic' model, we would expect (3.10) to hold for each  $\gamma$ . For example, the condition can fail for a timelike geodesic  $\gamma$  only if  $R_{ab} k^a k^b$  vanishes at every point on  $\gamma$ , and then only if the Weyl tensor is related in a very particular way to  $\gamma$  (i.e.  $C_{abcd} k^b k^c = 0$ ) at *every* point on  $\gamma$ . (For a generic space-time this would not even occur at *any* point of any  $\gamma$ !) The condition can fail for a null geodesic  $\gamma$  only if  $R_{ab} k^a k^b$  vanishes at every point of  $\gamma$  and the Weyl tensor has the tangent direction to  $\gamma$  as a principal null direction at *every* point of  $\gamma$  (cf. P, p. 162). (In a generic space-time, there would not be *any* null geodesic  $\gamma$  which is directed along a principal null direction at *six* or more of its points. This is because null geodesics form a five-dimensional system. It is  $n$  conditions on a null geodesic that it be directed along a principal null direction at  $n$  of its points, so such null geodesics form a  $(5-n)$ -dimensional system in a generic space-time.) We can thus reasonably say that it is only in very 'special' (and therefore physically unrealistic) models that the condition will fail.

We must now show why these three conditions together imply (3.2). The fact that they do is essentially a consequence of the *Raychaudhuri effect* (1955, cf. also P, p. 169; compare also Myers 1941). The idea here is to proceed so far along the causal geodesic  $\gamma$  that we get beyond the focal length of the effective 'lens system' due to the curvature along  $\gamma$  (compare Penrose 1965*b*). Consider a causal geodesic  $\gamma$  belonging to a hypersurface orthogonal congruence  $\Gamma$  of causal geodesics. We are interested in the members of  $\Gamma$  only in the immediate neighbourhood of  $\gamma$ .

When  $\gamma$  is a null geodesic, we shall, for convenience, specify that all the other members of  $\Gamma$  shall also be null. In this case we shall, in fact, be interested only in those members of  $\Gamma$ , near  $\gamma$ , which generate a null hypersurface containing  $\gamma$ . When  $\gamma$  is time-like we define the vector field  $t^a$  to be the unit future-directed tangents to the curves of  $\Gamma$ . When  $\gamma$  is null, we choose a vector field  $l^a$  to be smoothly varying future-directed tangents to the curves of  $\Gamma$ , where  $l^a$  is parallelly propagated along each curve. We have

$$\nabla_a t_b = \nabla_b t_a, \quad t^a t_a = 1, \quad D t^a = 0, \quad \text{with} \quad D = t^a \nabla_a \quad (3.12)$$

and

$$l_{[c} \nabla_a l_{b]} = 0, \quad l^a l_a = 0, \quad D l^a = 0, \quad \text{with} \quad D = l^a \nabla_a \quad (3.13)$$

respectively.

Let us first consider the timelike case. Ricci identities give, with (3.12),

$$R_{abcd} t^b t^d = D(\nabla_c t_a) + (\nabla_c t^d) (\nabla_d t_a). \quad (3.14)$$

Now  $R_{abcd} t^b t^d$  and  $\nabla_c t_a$  each annihilate  $t^a$  when transvected with it on any free index. Introduce an orthonormal basis frame, with  $t^a$  as one of the basis elements. Let  $Q_{\alpha\beta}$  and  $U_{\alpha\beta}$  denote the symmetric  $(3 \times 3)$  matrices of spatial components of  $R_{abcd} t^b t^d$  and  $\nabla_a t_b$ , respectively. Then (3.14) becomes

$$Q_{\alpha\beta} = D U_{\alpha\beta} - U_{\alpha\gamma} U_{\gamma\beta}. \quad (3.15)$$

The matrix  $Q_{\alpha\beta}$  defines the geodesic deviation (relative acceleration) of  $\Gamma$ ; the trace-free part of  $U_{\alpha\beta}$  defines the shear of  $\Gamma$ . We define the divergence  $\theta$  of  $\Gamma$  to be

$$\theta = \nabla_a t^a = -U_{\alpha\alpha}. \quad (3.16)$$

Taking the trace of (3.15), we get

$$D\theta + \frac{1}{3}\theta^2 = \frac{1}{3}(U_{\alpha\beta} U_{\alpha\beta} \delta_{\rho\sigma} \delta_{\rho\sigma} - U_{\alpha\beta} \delta_{\alpha\beta} U_{\rho\sigma} \delta_{\rho\sigma}) - Q_{\gamma\gamma} \leq 0 \quad (3.17)$$

by Schwarz's inequality and the energy condition (3.4) (which asserts  $Q_{\gamma\gamma} \geq 0$ ). Equality holds only when  $Q_{\gamma\gamma} = 0$  and  $U_{\alpha\beta}$  is proportional to  $\delta_{\alpha\beta}$  (so that the shear would have to vanish).

Suppose  $R_{abcd} t^b t^d \neq 0$  at some point  $x$  of  $\gamma$ , in accordance with (3.11). Then  $Q_{\alpha\beta} \neq 0$  at  $x$ . We shall show, first, that this implies that the strict inequality holds in (3.17) at some point  $y$  on  $\gamma$  with  $x \prec y$ . For if it turns out that  $Q_{\alpha\beta} = \mu \delta_{\alpha\beta}$  at  $x$  (for some  $\mu$ ), then clearly  $Q_{\alpha\beta} \neq 0$  at  $x$  implies  $Q_{\gamma\gamma} \neq 0$  at  $x$ , so that strict inequality holds at  $y = x$ . On the other hand, suppose  $Q_{\alpha\beta}$  is not of this form at  $x$ . Then by (3.15)  $U_{\alpha\beta}$  cannot be proportional to  $\delta_{\alpha\beta}$  throughout any open segment of  $\gamma$  whose closure includes  $x$ . Thus, the expression in parentheses in (3.17) must fail to vanish at some point  $y \in \gamma$  with  $x \prec y$ , so the strict inequality in (3.17) must hold at  $y$ .

Let the real quantity  $W$  be defined along  $\gamma$  as a non-zero solution of

$$DW = \frac{1}{3}\theta W \quad (3.18)$$

(so that  $W^3$  measures a spacelike 3-volume element orthogonal to  $\gamma$  and Lie transported along the curves of  $\Gamma$ ). Then (3.17) gives†

$$D^2W \leq 0 \quad (3.19)$$

along  $\gamma$ , provided  $W$  remains positive. Furthermore the strict inequality holds at  $y$ . Choosing  $W > 0$  at  $x$ , we see from (3.18) and (3.19) that if  $\theta \leq 0$  at  $x$ , then  $W$  becomes zero at some point  $q$  on  $\gamma$  with  $x \ll q$ . Furthermore, if  $\theta > 0$  at  $x$ , then  $W$  becomes zero at some  $p \in \gamma$  with  $p \ll x$ . This is *provided* we assume that  $\gamma$  is a *complete* geodesic. (By (3.12), we can interpret the ' $D$ ' in (3.17), (3.18), (3.19) as  $d/ds$ , where  $s$  is a proper time parameter on  $\gamma$ . The completeness condition ensures that the range of  $s$  is unbounded.) When  $W$  becomes zero, we have a *focal point* of  $\Gamma$  (point of the caustic) at which  $\theta$  becomes infinite (since  $\theta = 3D \ln W$ ).

Now fix the causal geodesic  $\gamma$  and fix a point  $x$  on it at which (3.11) holds: then allow the congruence  $\Gamma$  to vary. Thus, we consider solutions of (3.15), where the matrix  $Q_{\alpha\beta}$  is a given function of  $s$ . We shall be interested, in the first instance, in solutions for which  $\theta \geq 0$  at  $x$ . Then by the above discussion there will be a first focal point  $q_\Gamma$  on  $\gamma$ , for each  $\Gamma$  (with  $x \ll q_\Gamma$ ). Each solution of (3.15) is fixed once the value of  $U_{\alpha\beta} = \dot{U}_{\alpha\beta}$  is fixed at  $x$  (with  $\dot{U}_{\alpha\alpha} \geq 0$ ). Thus,  $q_\Gamma$  is a function of the nine  $\dot{U}_{\alpha\beta}$ . Furthermore, it must be a continuous function. We note that if any component of  $\dot{U}_{\alpha\beta}$  is very large, then  $q_\Gamma$  is very near  $x$  (since, in the limit  $Q_{\alpha\beta}$  becomes irrelevant and the solution resembles the flat space-time case). It follows that the  $q_\Gamma$ 's must lie in a bounded portion  $\zeta$  of  $\gamma$ . (The one-point compactification of the space of  $\dot{U}_{\alpha\beta}$ , with  $\dot{U}_{\alpha\alpha} \geq 0$  is mapped continuously into  $\gamma$ , with the point at infinity being mapped to  $x$  itself. Thus, the image must be compact.) Choose a point  $q \in \gamma$ , to the future of  $\zeta$  and let  $\Gamma$  consist of the timelike geodesics (near  $\gamma$ ) through  $q$ . If there were no conjugate point to  $q$  on  $\gamma$ , then the  $\Gamma$  congruence would be non-singular to the past of  $q$ . We cannot have  $\theta \leq 0$  at  $x$ , since this would imply  $q \in \zeta$ . But we have seen that  $\theta > 0$  implies another focal point to the past of  $x$ . This establishes the existence of a pair of conjugate points on  $\gamma$  in the timelike case.

When  $\gamma$  is null, the argument is essentially similar. In place of (3.14) we can use the Sachs equations (cf. P, p. 167) which have a matrix form similar to (3.15). The components of the curvature tensor which enter into these equations are just the four independent real (or two independent complex) components of  $l_{[a}R_{b]cd[e}l_{f]}l^cl^d$ . The analogue of  $\theta$  is  $-2\rho = \nabla_a l^a$ . In place of  $W$  we have a 'luminosity parameter'  $L$ , satisfying  $DL = -\rho L$  and  $D^2L \geq 0$ . The conclusion is the same: If (3.10) holds at some point on  $\gamma$ , if  $\gamma$  is complete and if the energy condition holds (in this case the weak energy condition (3.9) will suffice), then  $\gamma$  contains a pair of conjugate points.

† Equation (3.19), which follows from  $R_{ab}k^a k^b \leq 0$ , is essentially the statement that 'gravitation is always attractive' (cf. §1). It tells us that the geodesics of  $\Gamma$ , neighbouring to  $\gamma$ , have a tendency to accelerate towards  $\gamma$ —in the sense that freely falling 3-volumes accelerate inwards.

We now come to (3.3), the final condition of the theorem. A drawback of this condition, when it comes to applications, is that we may require considerable information of a global character concerning the space-time  $M$ , in order to decide whether or not a given set  $S$  is future-trapped. However, in certain special cases, we can invoke the weak energy condition and null-completeness, to enable us to infer, on the *basis of these two properties*, that a certain set should be future-trapped. An example of such a set  $S$  is a *trapped surface* (Penrose 1965a; P, p. 211), defined as a compact spacelike 2-surface with the property that *both* systems of null geodesics which intersect  $S$  orthogonally *converge* at  $S$ , as we proceed into the future. (For simplicity, suppose  $S$  to be achronal.) We expect trapped surfaces to arise when a gravitational collapse of a localized body (e.g. a star) to within its Schwarzschild radius takes place, which does not deviate too much from spherical symmetry. The significant feature of a trapped surface arises from the fact that the null geodesics meeting it orthogonally are the generators of  $E^+(S)$ . If these null geodesics start out by converging ( $\rho > 0$ ) then by the earlier discussion (Raychaudhuri effect in the null case—weak energy condition and null completeness assumed), they must continue to converge until they encounter a focal point. Either then, or before then, they must leave  $E^+(S)$  (cf. P, p. 218). Since  $S$  is compact and since the focal points must move continuously with the geodesic (being obtainable via integration of curvature), it follows that the geodesic segments joining  $S$  to the focal points must sweep out a compact set. Thus  $E^+(S)$ , being the intersection of this compact set with the closed set  $I^+[S]$ , must also be compact—so  $S$  is future-trapped and the theorem applies.

Precisely the same argument will apply in more general situations. For example, if  $S$  is any compact achronal set whose edge is smooth and at which the null geodesics which form the local boundary of its future (these will be orthogonal to  $\text{edge}(S)$ ) *converge* at  $\text{edge}(S)$  as we proceed into the future, then (again *assuming* null completeness and the weak energy condition)  $S$  will be future-trapped. More generally still, we need not require that the null geodesics which form the local boundary of the future of  $S$  actually converge at  $\text{edge}(S)$ . It is only necessary that we should have some reason for believing that they converge *somewhere* to the future of  $S$ . In particular,  $S$  might contain but a single point  $p$ , located somewhere near the centre of a collapsing body, but at a time before the collapse has drastically affected the geometry at  $p$ . Then, under suitable circumstances the future null cone of  $p$  can encounter sufficient collapsing matter that it (locally) starts converging again. Thus every null geodesic through  $p$  will encounter a point conjugate to  $p$  in the future (assuming null completeness and the weak energy condition), so again these null geodesic segments sweep out a compact set. Its intersection with  $I^+(p)$  is  $E^+(\{p\})$ , implying that  $E^+(\{p\})$  is compact, so  $\{p\}$  is future-trapped and the theorem applies.

In its time-reversed form, this last example has relevance to cosmology. If the point  $p$  refers to the earth at the present epoch, the null geodesics into the past, through  $p$  sweep out a region which can be taken to represent that portion of the

universe which is visible to us now. If sufficient matter (or curvature in general) encounters these null geodesics, then the divergence ( $-\rho$ ) of the geodesics may be expected to change sign somewhere to the past of  $p$ . This sign change occurs where an object of given size intercepting the null ray subtends its *maximum solid angle* at  $p$ . Thus, the *existence* of such a maximum solid angle for objects in each direction, may be taken as the physical interpretation of this type of past-trapped set  $\{p\}$ . Again the theorem applies. In an appendix we give an argument to show that the required condition on  $p$  seems indeed to be satisfied in our universe.

Another example of a future- (or past-) trapped set is any achronal set which is a *compact spacelike hypersurface*. (If we do not assume that the hypersurface is achronal, we can produce a ‘copy’ of it which is achronal by taking a suitable covering manifold of the entire space-time, cf. H. Thus, we actually lose no generality by assuming that  $S$  is achronal.) In this case, since  $\text{edge}(S) = \emptyset$ , we have  $E^+(S) = S$ , so  $E^+(S)$  is compact. Hence the theorem applies to ‘closed universe’ models. It is possible that still other situations of physical interest might arise in which a future- (or past-) trapped set  $S$  would be inferred as existing (perhaps on the basis of completeness or energy assumptions).

We are now in a position to state the corollary to our theorem.

**COROLLARY.** *A space-time  $M$  cannot satisfy causal geodesic completeness if, together with Einstein’s equations (3.5), the following four conditions hold:*

- (3.20)  *$M$  contains no closed timelike curves.*
- (3.21) *the energy condition (3.6) is satisfied at every point,*
- (3.22) *the generality condition (3.10) is satisfied for every causal geodesic,*
- (3.23)  *$M$  contains either*
  - (i) *a trapped surface,*
  - or      (ii) *a point  $p$  for which the convergence of all the null geodesics through  $p$  changes sign somewhere to the past of  $p$ ,*
  - or      (iii) *a compact spacelike hypersurface.*

We may interpret failure of the causal geodesic completeness condition in our corollary as virtually a statement that any space-time satisfying (3.20)–(3.23) ‘possesses a singularity’ (cf. Geroch 1968a and our earlier remarks). However, one cannot conclude, on the basis of the corollary, that such a singularity need necessarily be of the ‘infinite curvature’ type. Although one might infer that in some sense a ‘maximally extended’ space-time satisfying (3.20)–(3.23) should obtain arbitrarily large curvatures, there are, nevertheless, other possibilities to consider (cf. H). In fact, very little is known about the nature of the space-time singularities arising in general relativity other than in highly symmetrical situations. For this reason, it is worth pointing out the minor inference that can be made about the nature of these singularities if we revert back to our original statement of the theorem. The implication is, virtually, that a space-time satisfying

(3.20)–(3.23) must contain a causal geodesic which possesses no pair of conjugate points. At a first guess, one might have imagined that causal geodesics entering very large curvature regions would be inclined to possess many pairs of conjugate points. Instead, we see that our theorem implies that *some* causal geodesic ‘enters a singularity’ (i.e. is compelled to be geodesically incomplete) before any repeated focusing has time to take place.

*Proof of the theorem:*

Take  $S$  as future-trapped. Then, by lemma (2.12), there is a future-inextendible timelike curve  $\gamma \subset \text{int } D^+(E^+(S))$ . (That strong causality holds for  $M$  follows from lemma (2.10).) Define  $T = \bar{I}^-[ \gamma ] \cap E^+(S)$ . We shall show that  $T$  is past-trapped. (That  $T$  is closed and achronal follows at once since  $\bar{I}^-[ \gamma ]$  is closed and  $E^+(S)$  is closed and achronal.) Now, since  $\gamma \subset D^+(E(S))$ , every past-inextendible timelike curve with future end-point on  $\gamma$  must cross  $E^+(S)$ . More particularly, it must cross  $T$ . Also,  $\bar{I}^-[T] \subset I^-[ \gamma ]$ . Thus  $I^-[T]$  is simply a portion of  $I^-[ \gamma ]$  ‘cut off’ by  $T$ . Examining the boundaries of these sets, we see  $\dot{I}^-[T] \subset T \cup \dot{I}^-[ \gamma ]$ . We are interested in  $E^-(T) - T$ . This is generated by null geodesics  $\{\beta\}$  on  $\dot{I}^-[T]$  with future end-point on  $T$  (at  $\text{edge}(T)$ ). These null geodesics can be continued on  $\dot{I}^-[ \gamma ]$  inextendibly into the future. (For, by lemma (2.6), each point of  $\dot{I}^-[ \gamma ]$  is the past end-point of a null geodesic on  $\dot{I}^-[ \gamma ]$  which continues future-inextendibly unless it meets  $\gamma$ . But it clearly cannot meet  $\gamma$ , since  $\gamma$  is timelike and future-inextendible.) But, by (3.2), every generator  $\beta$  of  $\dot{I}^-[T]$  must, when maximally extended, contain a pair of conjugate points  $p, q$ , with  $p \prec q$ , say. By lemma (2.9),  $p$  cannot lie on  $\dot{I}^-[ \gamma ]$  (so  $p \in I^-[ \gamma ]$ ). Thus  $\beta$  must contain a past end-point either at  $p$ , or to the future of  $p$ . Now  $T$  and  $\text{edge}(T)$  are compact (being closed subsets of the compact set  $E^+(S)$ ). Since  $\beta$  meets  $\text{edge}(T)$  and since conjugate points vary continuously, (being obtainable as integrals of curvature, cf. Hicks 1964, H) we can choose  $p$  and  $q$ , for each  $\beta$ , so that the segment of the extension of  $\beta$  from  $p$  to  $q$  sweeps out a compact region. Thus, the segment of the extension of  $\beta$  from  $p$  to  $\text{edge}(T)$  also sweeps out some compact region  $C$  of  $M$ . We have  $E^-(T) = \dot{I}^-[T] \cap (C \cup T)$ , showing that  $E^-(T)$  is a closed subset of the compact set  $C \cup T$  and is therefore itself compact. Thus,  $T$  is past-trapped, as required.

By lemma (2.12) there exists a past-inextendible timelike curve  $\alpha \in \text{int } D^-(E^-(T))$ . Choose a point  $a_0 \in \alpha$ . We have  $a_0 \in I^-[ \gamma ]$ , so we find  $c_0 \in \gamma$  with  $a_0 \ll c_0$ . Choose the sequence  $a_0, a_1, a_2, \dots \in \gamma$ , receding into the past indefinitely (i.e. with no limit point). Similarly choose  $c_0, c_1, c_2, \dots \in \gamma$  proceeding into the future indefinitely. We have  $a_i \ll c_i$  for all  $i$ . Now  $a_i \in \text{int } D^-(E^-(T))$  and  $c_i \in \text{int } D^+(E^+(S))$ . Thus by lemma (2.7)  $J^+(a_i) \cap J^-[T]$  is compact (with strong causality holding throughout) and so is  $J^-(c_i) \cap J^+[S]$ . It is easily seen that  $J^+(a_i) \cap J^-(c_i)$  is a closed subset of  $\{J^-(c_i) \cap J^+[S]\} \cup \{J^+(a_i) \cap J^-[T]\}$  and so is also compact with strong causality holding throughout. Thus, by lemma (2.11) there is a maximal causal geodesic  $\mu_i$  from  $a_i$  to  $c_i$ . Now  $\mu_i$  must meet  $T$ , which is compact, at  $q_i$ , say. As  $i \rightarrow \infty$ , there will be an accumulation point  $q$  in  $T$  and an accumulation causal direction at  $q$ .

Choose the causal geodesic  $\mu$ , through  $T$ , in this direction, so  $\mu$  is approached by  $\mu_i$ . By (3.2),  $\mu$  contains a pair of conjugate points,  $u$  and  $v$ , say, with  $u \prec v$ . Since conjugate points vary continuously, we must have  $u$  as a limit point of some  $\{u_j\}$  and  $v$  as a limit point of some  $\{v_j\}$  where  $u_j$  and  $v_j$  are conjugate points on the maximal extension of  $\mu_j$ , the  $\{\mu_j\}$  being chosen to converge on  $\mu$ . But  $\{a_i\}$  and  $\{c_i\}$  cannot accumulate at any point of the segment  $uv$  of  $\mu$ . Hence, for some large enough  $j$ ,  $a_j$  will lie to the past of  $u_j$  in  $\mu_j$  and  $c_j$  to the future of  $v_j$  on  $\mu_j$ . This contradicts lemma (2.8) and the maximality of  $\mu_j$ . The theorem is thus established.

The authors are grateful to C. W. Misner and to R. P. Geroch for valuable discussions.

#### REFERENCES

- Avez, A. 1963 *Inst. Fourier* **105**, 1.  
 Boyer, R. H. 1964 *Nuovo Cim.* **33**, 345.  
 Brans, C. & Dicke, R. H. 1961 *Phys. Rev.* **124**, 925.  
 Carter, B. 1967 Stationary axi-symmetric systems in general relativity (Ph.D. Dissertation, Cambridge University).  
 Chandrasekhar, S. 1935 *M.N.* **95**, 207.  
 Chandrasekhar, S. 1964 *Phys. Rev. Lett.* **12**, 114, 437.  
 Dicke, R. H. 1962 *Phys. Rev.* **125**, 2163.  
 Geroch, R. P. 1966 *Phys. Rev. Lett.* **17**, 446.  
 Geroch, R. P. 1968a *Ann. Phys.* **48**, 526.  
 Geroch, R. P. 1968b *J. Math. Phys.* **9**, 1739.  
 Hawking, S. W. 1966a *Proc. Roy. Soc. Lond. A* **294**, 511.  
 Hawking, S. W. 1966b *Proc. Roy. Soc. Lond. A* **295**, 490.  
 Hawking, S. W. 1966c Singularities and the Geometry of space-time (Adams Prize Essay, Cambridge University).  
 Hawking, S. W. 1967 *Proc. Roy. Soc. Lond. A* **300**, 187.  
 Hawking, S. W. & Ellis, G. F. R. 1968 *Astrophys J.* **152**, 25.  
 Hicks, N. J. 1965 *Notes on differential geometry*. Princeton: D. van Nostrand Inc.  
 Hoyle, F. & Narlikar, J. V. 1963 *Proc. Roy. Soc. Lond. A* **273**, 1.  
 Kronheimer, E. H. & Penrose, R. 1967 *Proc. Camb. Phil. Soc. Lond.* **63**, 481.  
 Lifshitz, E. M. & Khalatnikov, I. M. 1963 *Adv. Phys.* **12**, 185.  
 Lindquist, R. W. & Wheeler, J. A. 1957 *Rev. Mod. Phys.* **29**, 432.  
 Milnor, J. 1963 *Morse theory*. Princeton University Press, Princeton.  
 Myers, S. B. 1941 *Duke Math. J.* **8**, 401.  
 Penrose, R. 1965a *Phys. Rev. Lett.* **14**, 57.  
 Penrose, R. 1965b *Revs. Mod. Phys.* **37**, 215.  
 Penrose, R. 1966 An analysis of the structure of space-time (Adams Prize Essay, Cambridge University).  
 Penrose, R. 1968 in *Battelle Rencontres, 1967 Lectures in Mathematics and Physics* (Ed. De Witt, C. M. & Wheeler, J. A.) New York: W. A. Benjamin Inc.)  
 Penrose, R. 1969 in *Contemporary physics: Trieste Symposium* 1968 (paper SMR/63).  
 Raychaudhuri, A. 1955 *Phys. Rev.* **98**, 1123.  
 Seifert, H. J. 1967 *Z. Naturforsch.* **22a**, 1356.  
 Sexl, R. U. & Urbantke, H. 1967 *Acta Phys. Austriaca* **26**, 339.  
 Steenrod, N. 1951 *The topology of fibre bundles* (Princeton University Press).

## APPENDIX

We wish to show that there is enough matter on the past light-cone of our present location  $p$  to imply that the divergence of this cone changes sign somewhere to the past of  $p$ . A sufficient condition for this to be so is that there should be (affine) distances  $R_1$  and  $R_2$  such that along every past-directed null geodesic from  $p$ ,

$$\frac{1}{2}KR_1 \int_{R_1}^{R_2} T_{ab} l^a l^b dr > 1. \quad (\text{A } 1)$$

(This formula can be obtained by using a variational approach similar to that used in Hawking (1966a).) As in (3.5),  $K = 8\pi G$ , where  $G$  ( $= 7.41 \times 10^{-29}$  cm g $^{-1}$ ) is the gravitational constant. (Length and time units are related via  $c = 1$ , i.e.  $3 \times 10^{10}$  cm = 1 s.)

In this integral, the vector  $l^a$  is a future-directed tangent to the null geodesic and  $r$  is a corresponding affine parameter ( $l^a \nabla_a r = -1$ ). Here  $l^a$  is parallelly propagated along the null geodesic and is such that  $r = 0$  at  $p$  and  $l^a U_a = 1$ , where  $U^a$  is the future-directed unit timelike vector representing the local standard of rest at  $p$ .

In a recent paper (Hawking & Ellis 1968) it was shown that, with certain assumptions, observations of the microwave background radiation indicate that not only do the past directed null geodesics from us start 'converging again' but so also do the timelike ones. As we are concerned only with the null geodesics, the assumptions we shall need will be weaker.

The observations show that between the wavelengths of 20 cm and 2 mm the background radiation is isotropic to within 1% and has a spectrum close to that of a black body at 2.7 K. We shall assume that this spectrum and its isotropy indicate not that the radiation was necessarily created with this form, but that it has undergone repeated scattering. (We do not assume that the radiation is necessarily primeval.) Thus there must be sufficient matter on each past directed null geodesic from  $p$  to make the optical depth large in that direction. We shall show that this matter will be sufficient to cause the inequality (A 1) to be satisfied.

The smallest ratio of density to opacity at these wavelengths will be obtained if the matter consists of ionised hydrogen in which case there would be scattering by free electrons. The optical depth to distance  $R$  would be

$$\int_0^R \frac{\sigma}{m} \rho l^a V_a dr,$$

where  $\sigma$  is the Thomson scattering cross-section,  $m$  the mass of a hydrogen atom,  $\rho$  the density, measured in g cm $^{-3}$ , of the ionised gas and  $V^a$  the local velocity of the gas. The red-shift  $Z$  of the gas is given by ( $l^a V_a - 1$ ). We assume that this increases down our past-light cone. As galaxies are observed with red-shifts of 0.46 most of the scattering must occur at red-shifts greater than this (in fact if the quasars really are at cosmological distances, the scattering must occur at red-shifts

of greater than 2). With a Hubble constant of  $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$ , a red-shift of 0.4 corresponds to a distance of about  $3 \times 10^{27} \text{ cm}$ . Taking  $R_1$  to be this distance, the contribution of the gas density to the integral in (A1) is

$$2.6 \int_{R_1}^{R_2} \rho (l^a V_a)^2 dr$$

while the optical depth of gas at red-shifts greater than 0.4 is

$$0.4 \int_{R_1}^{R_2} \rho l^a V_a dr.$$

As  $l^a V_a$  will be greater than 1.4 for  $r > R_1$  it can be seen that the inequality (A1) will be satisfied at an optical depth of about 0.1. If the optical depth of the Universe were less than this, one would not expect either a black body spectrum or a high degree of isotropy, as the photons would not suffer sufficient collisions. Even if the radiation arose from an isotropic distribution of black-body emitters at a higher temperature but covering less than  $\frac{1}{10}$  of the sky, what one would see would then be a dilute 'grey' body spectrum which could agree with the observations between 20 and 2 cm but which would not fit those at 9 and 2 mm. Thus we can be fairly certain that the required condition is satisfied in the observed Universe.