Open and closed universes, initial singularities, and inflation

Arvind Borde*
Institute of Cosmology, Department of Physics and Astronomy, Tufts University, Medford, Massachusetts 02155

(Received 1 April 1994)

The existence of initial singularities in expanding universes is proved without assuming the timelike convergence condition. The assumptions made in the proof are ones likely to hold both in open universes and in many closed ones. (It is further argued that at least some of the expanding closed universes that do not obey a key assumption of the theorem will have initial singularities on other grounds.) The result is significant for two reasons: (a) previous closed-universe singularity theorems have assumed the timelike convergence condition, and (b) the timelike convergence condition is known to be violated in inflationary spacetimes. An immediate consequence of this theorem is that a recent result on initial singularities in open, future-eternal, inflating spacetimes may now be extended to include many closed universes. Also, as a fringe benefit, the time reverse of the theorem may be applied to gravitational collapse.

PACS number(s): 98.80.Cq, 04.20.Dw

I. INTRODUCTION

The singularity theorems of classical general relativity [1,2] may be divided into two categories: those that use the *timelike convergence condition* ($R_{ab}V^aV^b \geq 0$, for all timelike vectors V^a) and those that do not. The theorems that use this condition do so in order to make congruences of timelike geodesics focus. Those that do not, use instead the *null convergence condition* ($R_{ab}N^aN^b \geq 0$, for all null vectors N^a) in order to make congruences of null geodesics focus. In both cases the consequences of this focusing are then shown to be incompatible with the other assumptions of the theorem. Theorems in the second category include Penrose's pioneering 1965 theorem on singularities in gravitational collapse [3], Hawking's application of the time reverse of that theorem to cosmology [4], and a recent theorem on singularities in inflating spacetimes [5,6].

The timelike convergence condition implies, by continuity, the null convergence condition, but the reverse implication does not hold: there are spacetimes (de Sitter spacetime is an example) that violate the timelike convergence condition but honor the null convergence condition. The violation of the timelike convergence condition in de Sitter spacetime means that the condition is violated in the inflating regions of known inflationary spacetimes. In fact, it has been argued [6] that a violation of this condition is *necessary* in order that a region be considered "inflating." For these reasons it is important to prove singularity theorems without assuming the timelike convergence condition, especially if the theorems are meant to apply to cosmology. Such theorems exist (as mentioned above) but they have certain weaknesses. Theorems that are directly based on Penrose's 1965 theorem, for instance, make very strong additional assumptions about the global structure of spacetime. More significantly, cosmological singularity theorems that do not assume the timelike convergence condition have all (so far) been applicable only to open universes. Typical closed-universe singularity theorems, on the other hand, assume the timelike convergence condition [7–9], as do both the multipurpose 1970 theorem of Hawking and Penrose [10] and Galloway's theorems extending closed-universe singularity results [11]. Violations of the timelike convergence condition thus provide a basis for several apparently nonsingular closed cosmologies [12].

In this paper, a singularity theorem is proved without assuming the timelike convergence condition; the theorem applies to open universes and to many closed ones. It is further argued that some of the closed universes to which the theorem does not apply possess initial singularities for other reasons. The theorem provides (among other things) the extension to closed universes of a recent result that demonstrates the necessity of initial singularities in open, future-eternal, inflationary cosmologies [5,6].

The paper is organized as follows. Section II discusses notation and terminology, and it gives some background results. Section III analyzes the strategy used in some open-universe singularity theorems. Section IV discusses the recent singularity theorem that deals with open, future-eternal, inflationary spacetimes and it shows that this theorem, too, fits the pattern laid out in Sec. III. This analysis of open-universe theorems is important because it suggests how one might proceed in closed universes. Section V then discusses a feature of some closed universes that might prevent open-universe arguments from going through. It points out that this feature does not always occur, and it argues that some of the closed universes in which it does will have other properties that force them, too, to have initial singularities.

*Permanent address: Long Island University, Southampton, NY 11968, and High Energy Theory Group, Brookhaven National Laboratory, Upton, NY 11973. Electronic address: borde@bnlcl6.bnl.gov

Part of the discussion here revolves around an interesting singularity-free spacetime due to Bardeen. Section VI states and proves the main theorem of this paper and Sec. VII makes some comments on the significance of this theorem. The paper ends with three appendices. Appendix A briefly discusses how the standard convergence conditions follow from conditions on the energy-momentum tensor; it also discusses how these conditions may be weakened from point conditions to integral ones. Appendix B discusses certain features of Gödel's universe that make it a formidable obstacle when trying to prove simple singularity theorems. Appendix C discusses whether the singularity predicted by cosmological singularity theorems is indeed "cosmological," i.e., whether the theorems allow us to infer that the Universe as a whole had a single beginning.

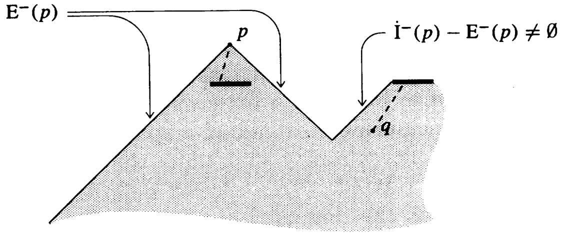

FIG. 1. An example of the causal complications that can arise in an unrestricted spacetime. Light rays travel along 45° lines in this diagram, and the two thick horizontal lines are identified. This allows the point q to send a signal to the point p along the dashed line, as shown, even though q lies outside what is usually considered the past light cone of p. The boundary of the past of p, $\dot{I}^-(p)$, then consists of the past light cone of p, $E^-(p)$, plus a further piece. Such a spacetime is not "causally simple."

II. NOTATION AND BACKGROUND RESULTS

Much of the discussion in this paper is based on the Penrose-Hawking-Geroch "global techniques" in general relativity. Everything that is needed is introduced and defined below. For further details, and for the proofs of all the assertions that are made in this section, see, for example, Hawking and Ellis [1].

Spacetime is represented by a manifold $\mathcal{M}$ with a Lorentz metric g_{ab} of signature $(-,+,+,+)$ defined on it. It is assumed that the metric permits a continuous global distinction between past and future (i.e., it is *time orientable*). It will not be necessary to assume a field equation in any of the arguments given below. Convergence conditions are imposed at certain points, and the conclusions of this paper will be valid in any theory of gravity (such as Einstein's) in which these conditions (or, as discussed in Appendix A, some suitable integral version) are reasonable impositions on the curvature.

A curve in spacetime is called *causal* if it is everywhere timelike or null (i.e., lightlike). Let p be a point in spacetime. The *causal and chronological pasts* of p, denoted, respectively, by $J^-(p)$ and $I^-(p)$, are defined as:

$$J^-(p) = \{q : \text{ there is a future-directed causal curve from } q \text{ to } p\}$$

and

$$I^-(p) = \{q : \text{ there is a future-directed timelike curve from } q \text{ to } p\} \ .$$

The *past light cone* of p may then be defined [13] as $E^-(p) = J^-(p) - I^-(p)$. It may be shown [1] that the boundaries of the two kinds of pasts of p are the same; i.e., $\dot{J}^-(p) = \dot{I}^-(p)$. Furthermore, it may be shown that $E^-(p) \subset \dot{I}^-(p)$. In general, however, $E^-(p) \neq \dot{I}^-(p)$; i.e., the past light cone of p (as it has been defined here) is a subset of the boundary of the past p, but is not necessarily the full boundary of this past. This is illustrated in Fig. 1. The sets $E^-(p)$ and $\dot{I}^-(p)$ [and thus also $\dot{J}^-(p)$] are *achronal*; i.e., no two points on any of them can be connected by a timelike curve. These definitions of pasts and of past light cones may all be extended in a straightforward way from single points to arbitrary sets.

Spacetimes in which the type of behavior shown in Fig. 1 does not occur, i.e., in which $E^-(p) = \dot{I}^-(p)$ for all points p, are called *past causally simple.* The definition may also be tightened by further requiring that $\dot{I}^-(p) \neq \emptyset$ (this rules out certain causality violations), as was done in a previous theorem [5,6]. I will not use this tighter definition in the main theorem (theorem 6) since I will be making a separate causality assumption.

There are various kinds of causality conditions that may be imposed, depending on which of a hierarchy of causality violations are to be ruled out [1]. The most useful of the conditions is the *stable causality condition.* Roughly speaking, the condition says that spacetime does not contain closed timelike curves even when the metric is slightly perturbed (i.e., the spacetime neither violates causality nor is on the verge of doing so). A spacetime is stably causal if and only if it admits a *time function* [1]: i.e., it admits a function t whose gradient is timelike. We may assume that the gradient is future pointing; the function t must then strictly increase along every future-directed timelike curve.

Another concept that we will need is that of *global hyperbolicity.* Very roughly, a spacetime $\mathcal{M}$ is globally hyperbolic if there is a spacelike hypersurface $\mathcal{S}$ such that the entire future and past development of $\mathcal{M}$ can be predicted from data on $\mathcal{S}$. If a surface exists with this property, it is called a *global Cauchy surface* for $\mathcal{M}$. The existence of such a surface places very stringent constraints on the global structure of $\mathcal{M}$ [1].

At several points in this paper I compare closed universes with open ones; it is useful, therefore, to define precisely what is meant by these terms. Intuitively, a closed universe is one which "closes on itself spatially." This may be made precise by saying that a closed universe is one that contains a closed (i.e., compact, without boundary) spacelike hypersurface. An open universe may then be defined as one that contains no such surface. These definitions mean that an open universe is "open everywhere," but that a closed universe is just "closed somewhere." The definitions of open and closed universes may also be given a little differently by using an *achronal* hypersurface instead of a spacelike one. In spacetimes without causality violations the two definitions are closely related. This second definition of an open universe was the one used in the recent result on singularities in future-eternal, inflating spacetimes [5,6].

We also need some results from the theory of geodesic focusing. Consider a congruence of causal (i.e., null or timelike) geodesics. (A congruence is a set of curves in an open region of spacetime, one through each point of the region.) Let u be an affine parameter along the geodesics and let U^a be the tangent to the geodesics with respect to this parameter. The expansion of the geodesics may be defined as $\theta \equiv D_a U^a$, where D_a is the covariant derivative. Then the propagation equation for θ may be written in this form [1]:

$$\frac{d\theta}{du} \leq -\frac{1}{\alpha}\theta^2 - R_{ab}U^aU^b \ , \tag{1}$$

where $\alpha = 2$ for null geodesics and $\alpha = 3$ for timelike geodesics. This inequality leads to a key result on geodesic focusing:

Lemma 1. Let $\mathcal{M}$ be a spacetime in which $R_{ab}N^aN^b \geq 0$ for all null vectors N^a (i.e., the null convergence condition holds). Consider a congruence of null geodesics with affine parameter n. If γ is a member of this congruence, such that (i) the expansion θ of the congruence is negative on γ at some point $n = n_0$, and (ii) γ is complete in the direction of increasing n (i.e., γ is defined for all $n \geq n_0$), then $\theta \to -\infty$ along γ a finite affine parameter distance from n_0.

Proof. The proof is standard and is only sketched here; details may be found in Hawking and Ellis [1]. From formula (1) and the null convergence condition it follows that

$$\frac{d\theta}{dn} \leq -\frac{1}{2}\theta^2 \ .$$

The result is a consequence of this inequality. □

III. OPEN-UNIVERSE SINGULARITY THEOREMS

Penrose's 1965 theorem on singularities in gravitational collapse [3] is the mother of all singularity theorems in relativity. The theorem is based on the existence of a *future-trapped surface*: a compact (without boundary) spacelike two-surface such that both systems of future-directed null geodesics that emanate orthogonally from it (the "inward" system of light rays and the "outward" system) are converging (i.e., have negative expansion θ). A *marginally trapped surface* is defined similarly, but the expansion θ is just required to be nonpositive here. Although the concept of a trapped surface was originally invented in order to characterize a local collapsed system, it was soon realized by Hawking [4] that it could fruitfully be put to use in cosmology as well. Hawking pointed out that large enough two-surfaces on constant time slices (in terms of the usual time coordinate) of open Robertson–Walker spacetimes are past-trapped, allowing Penrose's argument to be applied here as well (in time-reversed form). de Sitter-like spacetimes also contain trapped surfaces [14,15]; this fact was exploited by Farhi and Guth [15] in arguing that it is impossible to create inflationary universes "in a laboratory."

Closely related to the concept of a trapped surface is the idea of a *reconverging light cone*: a point p is said to have a reconverging past light cone if the expansion θ of the past-directed null geodesics in the light cone becomes negative along every such geodesic [i.e., the null geodesics start to converge along every past-directed geodesic in $E^-(p)$]. The concept was used in the 1970 theorem proved by Hawking and Penrose [10]. It was further argued there (also see Hawking and Ellis [1]) that observations of the microwave background radiation allow us to infer how much this radiation must have been scattered, and that this in turn implies that there is sufficient matter along every line of sight from us to make our own past light cone reconverge.

The significance of trapped surfaces and reconverging light cones comes from this standard result:

Lemma 2. Let $\mathcal{M}$ be a spacetime in which $R_{ab}N^aN^b \geq 0$ for all null vectors N^a (i.e., the null convergence condition holds). Suppose that $\mathcal{M}$ contains either a point with a reconverging past light cone, or a past-trapped surface, both represented here by $\mathcal{X}$. If $\mathcal{M}$ is null-complete to the past, then $E^-(\mathcal{X})$ is compact.

Proof. The result is standard, and so the proof is only sketched here. Since θ becomes (or already is) negative along each of the null geodesics that initially lies in $E^-(\mathcal{X})$, it follows from lemma 1 (and the assumption of past null-completeness) that $\theta \to -\infty$ within a finite affine parameter distance on each geodesic. The divergence of θ to $-\infty$ is a signal that the geodesics have focused. It is a standard result in global general relativity that points on such null geodesics beyond the focal point enter the interior of the past light cone [i.e., enter $I^-(\mathcal{X})$] and no longer lie in $E^-(\mathcal{X})$ [1]. Thus each null geodesic that starts off in $E^-(\mathcal{X})$ leaves it within a finite affine parameter distance. Since $\mathcal{X}$ is compact, this implies that $E^-(\mathcal{X})$ is compact as well. □

The existence of such a compact set does not by itself lead to a singularity. To get a singularity from here, we need two additional ingredients. First, it appears that some sort of causality assumption is needed. Otherwise, it is possible for $E^-(\mathcal{X})$ to be empty (and thus trivially compact) or, if it is not empty, for it to be part of a complicated enough boundary, $\dot{I}^-(\mathcal{X})$, to make further analysis difficult. An interesting example along these lines is given by Gödel's universe [16]. Appendix B analyzes some features of this universe: the analysis shows that the spacetime obeys the strict null convergence condition and it contains both reconverging light cones and marginally trapped surfaces [17,18]. Yet it is nonsingular [1]. This escape from a singularity occurs because there are bad causality violations in the spacetime [17,19].

Once causality violations are excluded [20], two different approaches have been taken in the past to obtain a singularity. One approach, taken, for instance, in the Hawking-Penrose theorem [10], uses the strong energy condition. This approach will not be discussed here. The other approach, which typically uses a stronger causality assumption, delivers a singularity immediately by ruling out the possibility of a compact topology for $E^-(\mathcal{X})$. For example, theorems based on Penrose's 1965 theorem impose a very simple causal structure on spacetime by requiring that it possess a global Cauchy surface $\mathcal{S}$. Such a spacetime is necessarily causally simple; i.e., $E^-(\mathcal{X}) = \dot{I}^-(\mathcal{X})$, and thus $E^-(\mathcal{X})$ has no edge [1]. The existence of such a compact, achronal, edgeless hypersurface is then ruled out by further requiring that $\mathcal{S}$ be noncompact (this means that the Universe is open).

Thus, the structure of a singularity theorem of this type is as follows:

(a) It is postulated that there is a point or a set, both represented here by $\mathcal{X}$, with properties that lead, *if the spacetime is past null complete*, to $E^-(\mathcal{X})$ being compact. The set $E^-(\mathcal{X})$ is also nonempty, either by fiat, or because of the absence of causality violations (i.e., of closed timelike curves). The absence of causality violations may, in turn, follow from some other postulate (such as an assumption that $\mathcal{M}$ contains a global Cauchy surface).

(b) The hypersurface $E^-(\mathcal{X})$ is edgeless. This follows from causal simplicity—the assumption may be made directly or it may follow from some other assumption (such as an assumption that $\mathcal{M}$ contains a global Cauchy surface).

(c) The existence of the compact, achronal, edgeless hypersurface $E^-(\mathcal{X})$ is either asserted to be inconsistent with the structure of an open universe, or is shown to be inconsistent with some other open-universe assumption (such as an assumption that $\mathcal{M}$ contains a noncompact global Cauchy surface).

The conclusion drawn from this argument is that a spacetime cannot be null complete to the past under the conditions of the theorem, i.e., it contains an initial singularity.

IV. INFLATION

A recent result [5,6] shows that open universes that eternally inflate to the future must contain initial singularities. A universe is said to eternally inflate to the future if the process of inflation, once started, never completely ends. Theoretical work as well as computer calculations [21–23] support the picture that inflation is indeed future-eternal: there is an inflationary background in which new post-inflationary regions (i.e., regions where inflation has ended) are continually formed, but these regions never fill the entire universe. The inflationary expansion is driven by the potential energy of a scalar field φ, while the field slowly "rolls down" its potential $V(\varphi)$. When φ reaches the minimum of the potential this vacuum energy thermalizes, and inflation is followed in this region by the usual radiation-dominated expansion. The evolution of the field φ is influenced by quantum fluctuations, and as a result thermalization occurs at different times in different parts of the Universe.

A cosmological model in which new "islands of thermalization" [24] are continually formed leads to this question: can such a model be extended in a nonsingular way into the infinite past? Assuming that some reasonable and rather general conditions are met, the recently proved result shows that in open universes the answer to this is "no:" such models must necessarily contain initial singularities. This is significant, because it forces us in inflationary cosmologies, as in the standard big-bang ones, to face the question of what, if anything, came before [25].

Here is the precise statement of the result:

Theorem 3. A spacetime $\mathcal{M}$ cannot be null-geodesically complete to the past if it satisfies the following conditions: (A) It is past causally simple, with $E^-(x) \neq \emptyset, \forall x \in \mathcal{M}$; (B) it is open (i.e., $\mathcal{M}$ contains no compact, achronal hypersurfaces without edge); (C) it obeys the null convergence condition; (D) it has at least one point p such that for some point q to the future of p the volume of the difference of the pasts of q and p is finite.

Assumptions (A)–(C) are conventional as far as work on singularity theorems goes. But assumption (D) is new and is inflation specific. It has been discussed in detail elsewhere [6,24], but here is a rough, short explanation: If inflation is to be future eternal, then for a point p in the inflating region there must be a nonzero probability for there to be a point q a given geodesic distance to the future of p such that q also belongs to the inflating region. Now, it may be shown that if a point r lies in a thermalized region, then all points in $I^+(r)$ also lie in that thermalized region [5]. Furthermore, it seems plausible that there is a zero probability for no thermalized regions to form in an infinite spacetime volume. Then assumption (D) follows.

Proof. The full proof of this result is available elsewhere [5,6], but here is a sketch: If $\mathcal{M}$ is null-complete to the past, then $E^-(p)$ must be compact. This is so, because the volume of a small wedge of the region $I^-(q) - I^-(p)$ around a geodesic γ that lies in $E^-(p)$ throughout may be expressed as

$$\Delta \int_0^\infty \mathcal{A}(v) dv \ ,$$

where Δ is a constant, $\mathcal{A}$ is the cross-sectional area of $E^-(p)$ around γ, and v is an affine parameter along the geodesic (chosen to increase in the past direction). From assumption (D) this volume must be finite. This can happen only if $\mathcal{A}$ decreases somewhere. But

$$\frac{d\mathcal{A}}{dv} = \theta \mathcal{A} \ .$$

This means that θ must become negative somewhere. From assumption (C) and the argument of lemma 2 it follows that γ must enter $I^-(p)$, and so leave $E^-(p)$, within a finite affine parameter distance. Thus $E^-(p)$ is compact. But assumption (A) implies that $E^-(p)$ has no edge. These two statements taken together contradict assumption (B). □

Therefore, this argument, too, follows the general open-universe pattern laid out at the end of the previous section.

V. CLOSED UNIVERSES

The open-universe pattern shows that the crucial contradiction in open-universe singularity theorems arises from the existence of a compact, edgeless past light cone. The reason why closed universes prove awkward for such theorems is that it is possible for light cones in at least some closed universes to "wrap around" the whole universe and thus be compact without causing any problems. This is illustrated in Fig. 2. (As a point of interest to theorem 3, the past-volume difference is finite in such a spacetime.) Such behavior occurs, for instance, in the Einstein universe [1].

This behavior also occurs in an interesting spacetime due to Bardeen [26]. Though this spacetime was originally constructed in the context of gravitational collapse, the lessons that it teaches are equally relevant to the existence of initial singularities. Bardeen's example uses the Reissner-Nordström spacetime as inspiration. The Reissner–Nordström metric represents the spacetime exterior to a spherically symmetric object of mass m and electric charge e. The global properties of the spacetime depend on the relative magnitudes of e and m; we will be interested here in the case when $e^2 < m^2$. The fully extended spacetime consists of infinitely many regions, in each of which the standard spherical coordinates (t, r, θ, ϕ) may be used. The metric in each region is

$$ds^2 = -f(r)dt^2 + \frac{1}{f(r)}dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \ , \quad (2)$$

where

$$f(r) = 1 - \frac{2m}{r} + \frac{e^2}{r^2} \ .$$

This spacetime obeys the null convergence condition. Furthermore, there are trapped surfaces in the region $r_- < r < r_+$ (where $r_\pm = m \pm \sqrt{m^2 - e^2}$), and a physical singularity at $r = 0$. But, as Bardeen pointed out [26], the function f may be replaced in (2) by a new function g, chosen to remove the singularity, while still retaining the trapped surfaces and preserving the null convergence condition. One such function displayed by Bardeen is

$$g(r) = 1 - \frac{2mr^2}{(r^2 + e^2)^{3/2}} \ , \quad r \geq 0 \ .$$

When $e^2 < \frac{16}{27}m^2$, once again there are values $r_\pm$ of r such that the region $r_- < r < r_+$ contains trapped surfaces. The spacetime obeys the null convergence condition, yet it contains no physical singularities. A similar example (i.e., possessing trapped surfaces and obeying the null convergence condition, yet nonsingular) may be constructed by directly modifying the Reissner-Nordström metric [27]. This is done by choosing a value $r_0 < r_-$ and by replacing the Reissner–Nordström function f by a function g that agrees with f for $r \geq r_0$ but not for $r < r_0$. One such function is

$$g(r) = \begin{cases} f(r) \quad (r \geq r_0) \ , \\ 1 - \left(\frac{25m^2}{16e^2}\right)\left[\left(\frac{r}{r_0}\right)^2 - \frac{2}{5}\left(\frac{r}{r_0}\right)^4\right] \quad (0 \leq r \leq r_0) \ . \end{cases}$$

If $r_0 = 4e^2/5m$ and if the parameters e and m are chosen such that $e^2 < m^2 < \frac{16}{15}e^2$, then the new metric will have all the desired properties.

Now, the only condition of Penrose's theorem not obeyed by Bardeen's spacetime and its ilk (or, for that matter, by the Reissner–Nordström spacetime) is the global Cauchy surface condition. The lesson that is *conventionally* drawn from this is that the Cauchy surface assumption cannot be dropped lightly, if we still want to prove the existence of a singularity. If the assumption is dropped, it is usually argued, then another strong assumption must replace it, and that assumption is taken to be the timelike convergence condition [1].

But it is possible to draw a different lesson from Bardeen's example. This lesson is most clearly drawn if we compare the global structure of the Reissner–Nordström spacetime [1] with that of Bardeen's [26]. In the first case the topology of the spacelike sections is $S^2 \times R$ throughout. In the second it switches between $S^2 \times R$ and S^3. This is illustrated in Fig. 3. As the figure reveals, the escape from a singularity occurs in Bardeen's spacetime not because it fails to be globally hyperbolic, but because in the crucial region where trapped surfaces occur, it is possible for light rays to wrap around the universe. In fact, the trapped surface $\mathcal{T}$ and its future light cone $E^+(\mathcal{T})$ both lie in a globally hyperbolic region of the spacetime (more precisely, in the future Cauchy development of the surface $\mathcal{S}$). A singularity is avoided purely because $\mathcal{S}$ is compact.

The wrapping of light cones around the universe (such as occurs in Bardeen's spacetime or in the spacetime of Fig. 2) ought not, however, to be generic behavior, at least for the past cones of single points: the cosmological scale ought to be much larger than the scale on which light cones refocus [24] (or the scale on which light rays

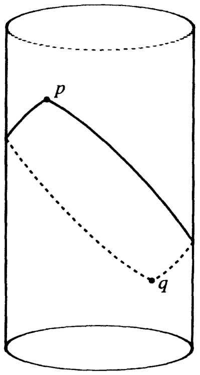

FIG. 2. A closed universe in which the past light cone of any point p is compact (and the volume of the difference of the pasts of any two points is finite). The past-directed null geodesics from p start off initially in $E^-(p)$; but, once they recross at q ("at the back") they enter $I^-(p)$ (because there are timelike curves between p and points on these null geodesics past q), and they thus leave $E^-(p)$.

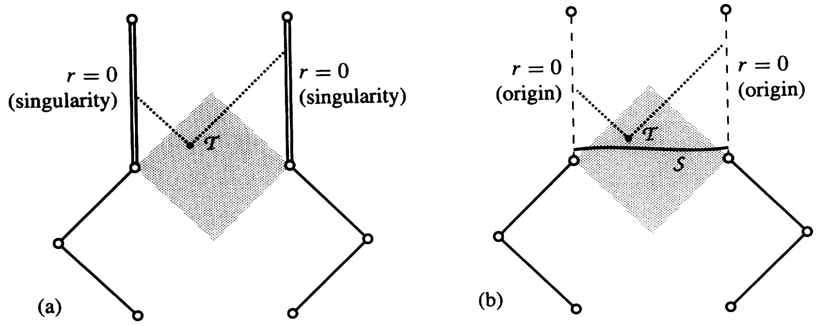

FIG. 3. The global structure of portions of (a) the Reissner-Nordström spacetime, and (b) Bardeen's spacetime. A point in the interior of both spacetimes represents a two-sphere. The boundaries of the diagram are drawn according to these conventions: single lines and hollow circles represent regions at infinity, double lines represent singularities, and dashed lines represent the origins ($r = 0$) of the coordinate systems. So, none of the boundaries in (a) are a part of the spacetime, whereas in (b) the $r = 0$ lines are. In both cases the $r = 0$ lines represent the origins of different coordinate patches. If one imagines a series of horizontal lines across each diagram, representing spacelike hypersurfaces, the topology of these surfaces will be $S^2 \times R$ throughout in (a), but in (b) the surfaces switch from $S^2 \times R$ to S^3 in the region between the $r = 0$ lines. (For instance, the surface $\mathcal{S}$ shown in (b) is a three-sphere.) There are trapped surfaces, $\mathcal{T}$, represented above by solid dots, in the shaded regions of both spacetimes. The dotted lines emanating from the trapped surfaces represent the two systems of future-directed null rays from $\mathcal{T}$: the "ingoing" and the "outgoing." Each system approaches (and in Bardeen's spacetime, reaches) a focal point at $r = 0$. Thus, in Bardeen's spacetime [i.e., in (b)], the future light cone of $\mathcal{T}$ "wraps around the universe." This light cone has topology S^3.

from "small" trapped surfaces, such as are ones likely to occur in gravitational collapse, focus). To state it another way, the behavior of light cones ought to depend only on (relatively) local effects, not on the behavior of the Universe as a whole.

There is an exception to this statement, one that is, unfortunately, of interest to this paper: an expanding closed universe might well have been small enough in the past for light rays to wrap around it easily. Since our chief interest is the existence of initial singularities, this scenario cannot lightly be dismissed. But, a slight adaptation of a theorem due to Hawking [8] may be used to show that in some cases an initial singularity will exist here as well.

Theorem 4. A spacetime $\mathcal{M}$ containing a compact, edgeless, spacelike hypersurface $\mathcal{S}$ cannot be timelike-geodesically complete to the past if there is a nonnegative number K such that (A) $R_{ab}T^aT^b \ge -\frac{1}{3}K^2$ for all unit timelike vectors T^a and (B) the past-directed timelike geodesics that emanate orthogonally from $\mathcal{S}$ have initial expansion $\theta_0 < -K$ at $\mathcal{S}$ (the past direction is the direction of increasing affine parameter).

Proof. Only the case when $K \ne 0$ is considered. (The $K = 0$ case is standard: see the comments immediately following the proof.) Suppose that $\mathcal{M}$ is timelike complete to the past. Let τ be the proper time along the past-directed timelike geodesics from $\mathcal{S}$. Choose τ to increase in the past direction and to have the value 0 at $\mathcal{S}$. Let T^a be the tangent to these geodesics with respect to τ (i.e., T^a is the four-velocity of the geodesics). From formula (1) and assumption (A) we get

$$\frac{d\theta}{d\tau} \le \frac{1}{3}(K^2 - \theta^2) .$$

This, along with assumption (B), means that

$$\theta \le K\coth[\tfrac{1}{3}K(\tau - \hat{\tau})] ,$$

where $\hat{\tau} \equiv -(3/K)\mathrm{arccoth}(\theta_0/K) > 0$ (because $\theta_0 < -K < 0$). Thus $\theta \to -\infty$ within a proper time $\hat{\tau}$ to the past of $\mathcal{S}$.

Once the existence of these focal points is established, the rest of the argument is identical to the one given by Hawking [1,8]. □

In the original statement and proof of Hawking's result, K is zero (the focusing is then shown slightly differently than is done here). This makes assumption (A) the standard timelike convergence condition, and assumption (B) a statement that the universe is contracting in the past direction (or, equivalently, expanding in the future direction). The slightly different formulation given here is meant to apply to situations where the timelike convergence condition might not hold. Many inflationary models, for example, assume the form $R_{ab} = 3\nu^2 g_{ab}$ for the Ricci tensor. This form satisfies assumption (A), with $K = 3\nu$ [28]. Thus Hawking's theorem adapted to such situations says that there will be an initial singularity, provided that the surface $\mathcal{S}$ is expanding sufficiently fast in the future direction. This theorem will cover at least some cases where past light cones wrap around the universe.

For the rest of this paper I will concentrate on situations where light cones do not wrap around the universe. A light cone such as the one in Fig. 2 that wraps around also swallows the Universe entirely [14]. One feature of this "swallowing" may be seen if we examine a point close to the light cone, but to its future: it appears impossible for *any* past-directed signal from the point to avoid intersecting the cone [29]. Actually, appearances are a little deceptive here: these statements are not true, as Fig. 4 shows, in causality-violating spacetimes of the

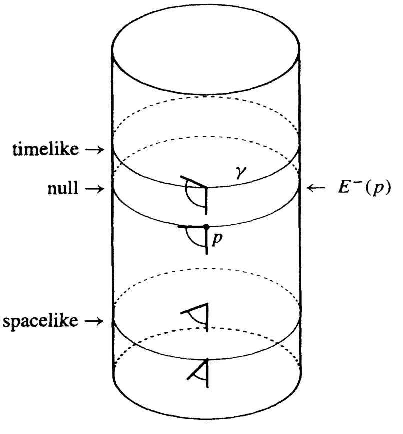

FIG. 4. A two-dimensional Taub-NUT-like closed universe due to Misner [1]. The causal behavior here is very different from that in Fig. 2. Local light cones are shown at several points (only the past cones): the two straight lines represent the two past-directed null vectors at each point, and the arc between them the interior of the local cone. The (global) past light cone of p, $E^-(p)$, consists of a single closed null geodesic through p. [The other null geodesic from p enters $I^-(p)$ immediately, since past-directed timelike curves from p can wind around the cylinder and return arbitrarily close to p.] The curve γ to the future of p is a closed timelike curve, and so it does not intersect the past of p. This means that even though in a certain sense $E^-(p)$ wraps around the universe, points to the future of this set can avoid sending signals that intersect it no matter how close they lie to $E^-(p)$.

Taub-NUT (Newman-Unti-Tamburino) variety [1]. But, if we exclude causality violations, we may take the statements made above as characterizing the behavior that we want to exclude. Therefore, I will assume in the theorem that past light cones are localized in the following sense:

Definition. A past light cone in a stably causal spacetime is called localized if from every spacetime point p not on the cone, there is at least one fully extended, past-directed timelike curve that does not intersect the cone.

A "fully extended, past-directed" curve is often called "past inextensible." Localized past light cones are illustrated in Fig. 5. Spacetimes in which past light cones are localized include all the standard open universes (Minkowski, Schwarzchild, the open Robertson-Walker cosmologies, etc.), as well as closed ones such as de Sitter and some of the closed Robertson-Walker cosmologies (here, the localization or not of light cones depends on the time scale for the recollapse of the universe).

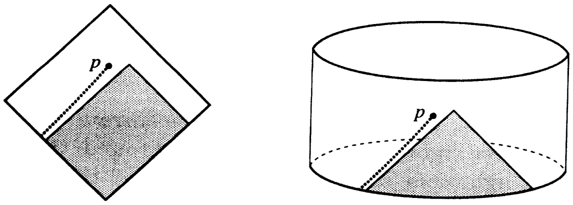

FIG. 5. Two examples of spacetimes with localized past light cones. The boundaries of both figures are boundaries at infinity. A typical past light cone is shown in each spacetime. The dotted line from each point p represents a past-directed timelike curve that avoids intersecting the past light cone.

VI. THE SINGULARITY THEOREM

If a past light cone in a stably causal spacetime is compact and without edge, it is not "localized." In fact, a slightly more general result holds.

Lemma 5. Let $\mathcal{M}$ be a spacetime that obeys the stable causality condition. Suppose that $\mathcal{M}$ contains a compact, achronal hypersurface, $\mathcal{S}$, without edge. There are then points p in $I^+\mathcal{S}$ such that every past-directed timelike curve from p intersects $\mathcal{S}$.

Proof. Let t be a time function on $\mathcal{M}$. Vary $\mathcal{S}$ forward a small amount in the future t direction. For a sufficiently small variation this gives a compact spacetime region $\mathcal{N}$ with two compact components to its boundary: $\mathcal{S}$, and a second component denoted by $\mathcal{S}'$ (see Fig. 6). Though $\mathcal{S}$ is achronal by assumption, $\mathcal{S}'$ does not have to be. Let t_0 be the minimum value of t on $\mathcal{S}'$, attained at some point p. Every past-directed timelike curve from p must enter $\mathcal{N}$ (because t_0 is the minimum value of t on the edgeless hypersurface $\mathcal{S}'$). Each such curve γ must also eventually leave $\mathcal{N}$: if it does not, it must accumulate at some point in the compact set $\mathcal{N}$, and examination of constant t surfaces in a small neighborhood of the accumulation point shows that this cannot happen (because t decreases along γ). The curve must leave through $\mathcal{S}$ and not through some other point of $\mathcal{S}'$ (again, because t decreases along γ). Thus, every past-directed timelike curve from p intersects $\mathcal{S}$. □

The main result of this paper follows trivially from the preceding discussion.

Theorem 6. A stably causal spacetime cannot simultaneously satisfy the following two conditions: (A) It is past causally simple; (B) it contains a compact, localized past light cone.

Proof. If $\mathcal{E}$ is the past light cone given by assumption (B), then assumption (A) implies that $\mathcal{E}$ has no edge (being the full boundary of the past). Also, $\mathcal{E}$ is achronal. Since $\mathcal{E}$ is localized, this contradicts lemma 5. □

This theorem (backed by lemma 2 and theorem 3) shows that spacetimes cannot be past null complete in a variety of circumstances:

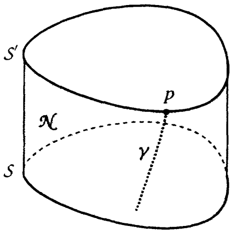

FIG. 6. An illustration of the strategy used in lemma 5.

Corollary. A stably causal spacetime $\mathcal{M}$ cannot be null-geodesically complete to the past if it satisfies the following conditions: (A) It is past causally simple; (B) all past light cones in $\mathcal{M}$ are localized; (C) it obeys the null convergence condition; (D) it contains either (i) a point with a reconverging past light cone, or (ii) a past-trapped surface, or (iii) a point p such that for some point q to the future of p the volume of the difference of the pasts of q and p is finite.

This result offers (among other consequences) a modification of theorem 3 by replacing its assumption (B) by the assumptions that the spacetime is stably causal and that all past light cones in the spacetime are localized. The modified theorem will not exclude closed universes, as the original one had.

VII. DISCUSSION

The result presented above has many of the strengths of the other singularity theorems, and it also shares many of their weaknesses. It is based on very modest assumptions, and it may thus be considered a strong result, but the conclusions that it arrives at are also somewhat limited. The theorem demonstrates the existence of a past-incomplete null geodesic, but it yields no information on where in the past the singularity lies. Nor does it demonstrate that the universe had a "single beginning" in the sense that Robertson-Walker models might be said to have one. The question of a single beginning is addressed further in Appendix C.

Still, the theorem closes several gaps in our understanding of the conditions that are likely to lead to, or to prevent, singularities. It shows, for instance, that nonsingular cosmologies must either violate the null convergence condition (in addition to the violations of the timelike convergence condition that many are already known to possess), or they must not have reconverging past light cones, or if light cones do reconverge then they must swallow the universe. (If there is a violation of the null convergence condition, it must be severe enough to also violate the integral conditions discussed in Appendix A.) The theorem also places the initial singularity in the past, where it rightfully belongs, unlike, for instance, the Hawking-Penrose theorem [10] which is silent on the location of the singularity.

Apart from its cosmological applications, the theorem may also be applied to gravitationally collapsing systems in which future-trapped surfaces occur. A future singularity is predicted here by the time reverse of the theorem, thus covering spacetimes like Reissner-Nordström (which, oddly, had not hitherto met the conditions of any singularity theorem [30]).

The theorem here is, of course, only as good as its assumptions: the weaker we make the assumptions, the stronger and more physically reliable the result. One condition that can probably be somewhat weakened is the causal simplicity assumption. This assumption was made solely to prove the theorem in the most direct way possible and with a minimum of mathematical fuss. But it would be preferable to have stable causality as the only restriction on causal structure. This issue will be discussed elsewhere.

It would also be nice if the assumption on localized null cones could be replaced by something likely to hold in all closed universes. But the existence of nonsingular closed spacetimes that obey the null convergence condition makes it far from obvious (at least to me) how to proceed without such an assumption—although it is possible that some kind of genericity condition might help here. In this connection, it is curious that closed-universe singularity theorems have tended to need stronger convergence conditions than open-universe theorems, despite the fact that a closed universe is presumably denser than an open one and so ought to have a greater natural proclivity for singularities.

ACKNOWLEDGMENTS

This paper arose out of discussions with Alex Vilenkin. It is a pleasure to thank him for stimulating my interest in the questions discussed here, for offering very many helpful comments and suggestions, and for reading the manuscript. I have also benefited from discussions with Rosanne Di Stefano, Alan Guth, and Tom Roman, and I thank them as well. I thank the Institute of Cosmology at Tufts University for its warm hospitality over the period when this work was done, and the High Energy Theory Group at Brookhaven National Laboratory for its continued support.

APPENDIX A: CONVERGENCE CONDITIONS AND ENERGY

Both of the standard convergence conditions, timelike and null, follow, via Einstein's equation, from certain inequalities, known as energy conditions, on the matter energy-momentum tensor [1]. Einstein's equation is

$$R_{ab} - \frac{1}{2}Rg_{ab} + \Lambda g_{ab} = 8\pi T_{ab} \ , \qquad \text{(A1)}$$

where R_{ab} is the Ricci tensor obtained from g_{ab}, R is the curvature scalar, Λ is the cosmological constant, and T_{ab} is the matter energy-momentum tensor. The inequalities that are useful for our purposes are the *strong energy condition* (which says that $T_{ab}V^aV^b - \frac{1}{2}T^a_aV^bV_b \geq 0$ for all timelike vectors V^a) and the *weak energy condition* (which says that $T_{ab}V^aV^b \geq 0$ for all timelike vectors V^a, from which it follows by continuity that $T_{ab}N^aN^b \geq 0$ for all null vectors N^a). The timelike convergence condition follows from the strong energy condition when the cosmological constant is zero, and the null convergence condition follows from the weak energy condition, even if there is a cosmological constant.

All these conditions are point conditions, and there has been discussion [31–37] of scenarios in which the energy conditions might be violated in a limited way (for example, at some points but not at others). It is known from some of this work [28,31,32], that what is important in order to ensure focusing is that $R_{ab}U^aU^b$ (where

U^a is the tangent to a null or a timelike geodesic) obey an integral inequality, not necessarily one that holds at every point. Such integral (or, as they have come to be called, averaged) convergence conditions will do just as well for the purposes of this paper. For instance, a condition that (roughly speaking) requires that $\int R_{ab}N^aN^b dn$ be repeatedly non-negative (along a geodesic with affine parameter n) is known to be sufficient to give focusing [28].

Despite the availability of these weaker conditions, the central arguments of this paper are phrased in terms of point conditions. This is done in order to keep the line of the argument clean and uncluttered with extraneous detail. It should be kept in mind, though, that any convergence condition, point or integral, that is sufficient to guarantee that a congruence of initially converging, complete geodesics actually comes to a focus is adequate for our purposes.

APPENDIX B: GÖDEL'S UNIVERSE

In the course of an investigation of the idealistic conception of time (i.e., of whether or not "reality consists of an infinity of layers of 'now' "), Gödel discovered in 1949 a very interesting solution to Einstein's equation [16]. The solution has closed timelike curves. It is also simply connected. It follows from this that Gödel's universe contains no edgeless spacelike hypersurfaces: "reality" here does not contain even a single layer of "now." Another interesting feature of this universe, though less dramatic, is that it has no singularities, despite (as we shall see below) possessing reconverging light cones and marginally trapped surfaces, and obeying the strict null convergence condition [38]. This makes it a tantalizing obstacle when trying to develop singularity theorems [39].

The manifold on which Gödel's metric is defined is $\mathbb{R}^4$. A set of coordinates (t', x, y, z) may be chosen such that each coordinate has range $(-\infty, \infty)$, with the metric given by

$$ds^2 = -dt'^2 + dx^2 - \frac{1}{2}\exp(2\sqrt{2}\omega x)dy^2$$

$$+dz^2 - 2\exp(\sqrt{2}\omega x)dt'dy \ ,$$

where ω is a positive constant. The metric satisfies Einstein's equation, with a cosmological constant [see Eq. (A1)], if $T_{ab} = \rho U_a U_b$, where $U^a = (\partial/\partial t')^a$, and $\omega^2 = -\Lambda = 4\pi\rho$. If N^a is a null vector, we have $R_{ab}N^aN^b = 2\omega^2(U_aN^a)^2 > 0$; i.e., the strict null convergence condition holds here.

The coordinates (t', x, y, z) are not the best ones in which to investigate light cones. New coordinates (t, r, ϕ, z) may be defined [1], with $-\infty < t < \infty, 0 \leq r < \infty, 0 \leq \phi \leq 2\pi$, and $-\infty < z < \infty$, by the following transformations:

$$\exp(\sqrt{2}\omega x) = \cosh 2r + \cos\phi \sinh 2r \ ,$$

$$\omega y \exp(\sqrt{2}\omega x) = \sin\phi \sinh 2r \ ,$$

$$\tan\frac{1}{2}(\phi + \omega t - \sqrt{2}t') = \exp(-2r)\tan\frac{1}{2}\phi \ ,$$

and

$$z = z \ .$$

In these coordinates, the metric is given by

$$ds^2 = \frac{2}{\omega^2}[-dt^2 + dr^2 - \sinh^2 r(\sinh^2 r - 1)d\phi^2$$

$$+2\sqrt{2}\sinh^2 r d\phi dt] + dz^2 \ .$$

The coordinate z plays no role in the behavior of the spacetime, and it will be ignored from now on. It may be checked that the closed curve given by

$$t = \text{const} \ , \quad r = \ln(1 + \sqrt{2}) \ , \quad \phi = \phi(p)$$

is a nongeodesic null curve [18] (where p is an arbitrary parameter). But, in the region $r < \ln(1 + \sqrt{2})$, the surfaces of constant t and r are spacelike. (They are also compact and edgeless, if we ignore z, or if we compactify the z direction by identifying, for instance, $z = 0$ with $z = 1$.) The tangents to the two systems of past-directed null geodesics that emanate orthogonally from these surfaces have these non-zero components:

$$N^t_\pm = \frac{\sinh^2 r - 1}{\cosh^2 r} \ ,$$

$$N^r_\pm = \pm\frac{(1 - \sinh^2 r)^{1/2}}{\cosh r} \ ,$$

and

$$N^\phi_\pm = \frac{\sqrt{2}}{\cosh^2 r} \ ,$$

where N^a_+ is tangent to the outgoing null geodesics and N^a_- to the ingoing ones (both vectors point in the past direction). The expansion of the two systems may be computed to be

$$\theta_\pm \equiv \nabla_a N^a_\pm = \pm\left[\frac{1 - 2\sinh^2 r}{\sinh r\sqrt{1 - \sinh^2 r}}\right] \ .$$

Thus, the expansion of the outgoing null geodesics is positive for small r, becomes zero at $r = r_0$ (where $\sinh^2 r_0 = 1/2$), and negative thereafter; i.e., the past light cone of a point on the $r = 0$ axis reconverges for $r > r_0$. Meanwhile, the ingoing null geodesics start off with negative expansion for small r. The expansion for this system, too, becomes zero at $r = r_0$. Therefore, at any time t, the surface given by $r = r_0$ is a marginally trapped surface. It is interesting to note that although the outgoing null geodesics are converging (i.e., $\theta < 0$) for $r > r_0$, r continues to increase for these geodesics. And the "focus" that the geodesics come to, occurs on the surface given by $r = \ln(1 + \sqrt{2})$.

APPENDIX C: A SINGLE BEGINNING?

Singularity theorems in general relativity often confine themselves to proving the existence of one incomplete causal geodesic. In what sense, then, can a cosmological singularity theorem be said to show that the universe *as a whole* [23] had a single singular beginning?

Although it does not appear possible to conclusively prove the necessity of a single beginning without resorting to model-dependent calculations, there are pieces of evidence that suggest that the existence of a single beginning is a plausible consequence of the singularity theorems. For instance, theorem 3 (and its extension in Sec. VI) shows that the past light cone of any point p satisfying assumption (D) must be incomplete. Arguments given elsewhere [6,24], show that almost all points in the inflating region of an inflationary spacetime will satisfy assumption (D). It follows that each of these points must have an initial singularity (i.e., a past-incomplete causal geodesic) somewhere to the past. This does not prove that all these singularities lie on one spacelike hypersurface, but it does make such a scenario possible.

Stronger evidence for this scenario may be obtained by adding to theorem 4 the assumption that $\mathcal{S}$ is a global Cauchy surface and requiring that K be nonzero (it is not necessary then to assume that $\mathcal{S}$ is compact, making the result applicable to both open and closed universes). It follows by a standard argument from the altered assumptions that no timelike geodesic that emanates orthogonally from $\mathcal{S}$ can exist for a proper time greater than $\tau_{\max} \equiv -(3/K)\text{arccoth}(\theta_{\max}/K)$ to the past of $\mathcal{S}$, where $\theta_{\max}$ is the largest value of the divergence of the geodesics at $\mathcal{S}$ (i.e., it represents the least past-convergent of these geodesics). For, suppose that there is a point r on one of these geodesics a proper time $\tau > \tau_{\max}$ to the past of $\mathcal{S}$. Since $\mathcal{S}$ is a global Cauchy surface there must be a maximal timelike curve γ between r and $\mathcal{S}$ [1]. This curve must have length greater than $\tau_{\max}$ and it must intersect $\mathcal{S}$ orthogonally [1]. But then θ must diverge to $-\infty$ on γ, between $\mathcal{S}$ and r. This means that γ cannot be maximal [1].

The result shows that in models that (a) possess a global Cauchy surface, and (b) are expanding sufficiently fast, there is an upper bound on the lengths (i.e., proper times) of timelike geodesics when they are followed into the past from the Cauchy surface. This is a further piece of evidence (though again not conclusive [40]) that it seems reasonable to infer from singularity theorems that the classical universe did indeed have a single beginning.

[1] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Spacetime* (Cambridge University Press, Cambridge, England, 1973).

[2] For reviews of singularity theorems other than the one in Ref. [1], see, for example, F. J. Tipler, C. J. S. Clarke, and G. F. R. Ellis, in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980).

[3] R. Penrose, Phys. Rev. Lett. **14**, 57 (1965).

[4] S. W. Hawking, Phys. Rev. Lett. **15**, 689 (1965).

[5] A. Borde and A. Vilenkin, Phys. Rev. Lett. **72**, 3305 (1994).

[6] A. Borde and A. Vilenkin, in the Proceedings of the Eighth Yukawa Symposium on Relativistic Astrophysics, Japan, 1993 (Universal Academic, Japan, in press).

[7] R. Geroch, Phys. Rev. Lett. **17**, 445 (1966).

[8] S. W. Hawking, Proc. R. Soc. London **A300**, 187 (1967).

[9] A. Borde, Class. Quantum. Grav. **2**, 589 (1985).

[10] S. W. Hawking and R. Penrose, Proc. R. Soc. London **A314**, 529 (1970).

[11] G. Galloway, Math. Proc. Camb. Philos. Soc. **99**, 367 (1986).

[12] L. Parker and S. A. Fulling, Phys. Rev. D **7**, 2357 (1973); G. L. Murphy, *ibid.* **8**, 4231 (1973); J. D. Bekenstein, Phys. Rev. D **11**, 2072 (1975); M. Markov and V. Mukhanov, Nuovo Cimento **86B**, 97 (1985); V. Mukhanov and R. Brandenberger, Phys. Rev. Lett. **68**, 1969 (1992); R. Brandenberger, V. Mukhanov, and A. Sornborger, Phys. Rev. D **48**, 1629 (1993).

[13] The expression "light cone" is used differently in different places in the literature. For instance, it is often used to refer to the collection of *all* points on null geodesics that emanate from a given point. This use of the expression is related to the definition of this paper, but it is not identical to it. As far as we are concerned, a point q does not lie on the light cone of a point p if there is a timelike curve connecting p and q, even if q also lies on a null geodesic from p. (Such connections cannot occur in Minkowski spacetime, but they can, and do, occur as soon as there is a little curvature.) The expression "light cone" is also often used in a more local sense, to describe just the null vectors at a point p; such a structure will be called a "local light cone" in this paper.

[14] R. Penrose, in *Battelle Rencontres*, edited by C. M. DeWitt and J. A. Wheeler (Benjamin, New York, 1968).

[15] E. Farhi and A. H. Guth, Phys. Lett. B **183**, 149 (1987).

[16] K. Gödel, in *Albert Einstein: Philosopher-Scientist*, edited by P. A. Schilpp (Open Court, Chicago, 1949); Rev. Mod. Phys. **21**, 447 (1949).

[17] A. Borde, Syracuse University Report, 1987 (unpublished).

[18] A. Borde, *Example V in TeX by Example* (Academic, Cambridge, Massachusetts, 1992).

[19] R. P. A. C. Newman, Gen. Relativ. Gravit. **21**, 981 (1989).

[20] A causality assumption appears necessary only in theorems that use trapped surfaces or reconverging light cones as their starting point. There are other theorems, which start differently and do not need causality assumptions [8,9]. These theorems are, however, closed-universe theorems that assume the strong energy condition.

[21] P. J. Steinhardt, in *The Very Early Universe*, edited by G. W. Gibbons, S. W. Hawking, and S. T. C. Siklos (Cambridge University Press, Cambridge, England, 1983); A. Vilenkin, Phys. Rev. D **27**, 2848 (1983); A. D. Linde, Phys. Lett. B **175**, 395 (1986).

[22] M. Aryal and A. Vilenkin, Phys. Lett. B **199**, 351 (1987); A. S. Goncharov, A. D. Linde, and V. F. Mukhanov, Int. J. Mod. Phys. A **2**, 561 (1987); K. Nakao, Y. Nambu,

and M. Sasaki, Prog. Theor. Phys. **80**, 1041 (1988).

[23] A. Linde, D. Linde, and A. Mezhlumian, Phys. Rev. D **49**, 1783 (1994).

[24] A. Vilenkin, Phys. Rev. D **46**, 2355 (1992).

[25] Inflation-related calculations are usually carried out by assuming that the metric is exactly de Sitter. In these cases, the existence of past-incomplete geodesics is well known, since the "exponentially expanding" part of de Sitter spacetime represents only part of the full, geodesically complete, spacetime. See, for example, Ref. [1] for a discussion of this point. The significance of the result in Refs. [5] and [6] is that geodesic incompleteness is shown to follow from very general assumptions, without assuming a particular form for the metric.

[26] J. Bardeen, in Proceedings of GR5, Tiflis, USSR, 1968 (unpublished).

[27] A. Borde, Ph.D. dissertation, SUNY at Stony Brook, 1982.

[28] A. Borde, Class. Quantum. Grav. **4**, 343 (1987).

[29] In fact, in Fig. 2, a point far from the light cone, but still to its future, will also have this property. But the statement here is phrased in terms of a point that is "close" so as to cover other situations where there might be topological interconnections or singularities far from the light cone that divert or block signals from reaching it.

[30] The multipurpose Hawking-Penrose theorem [10] comes close to embracing the Reissner-Nordström spacetime, but its "generic condition" does not hold there. (Though the original impetus for the development of singularity theorems came from the need to show that singularities were not just a feature of special solutions, but were generic, the adoption of the "generic condition" in the Hawking-Penrose theorem has had the effect of only covering generic situations, and excluding special ones.)

[31] F. J. Tipler, J. Diff. Eq. **30**, 165 (1978); Phys. Rev. D **17**, 2521 (1978).

[32] C. Chicone and P. Ehrlich, Manuscripta Math. **31** 297 (1980).

[33] T. Roman, Phys. Rev. D **33**, 3526 (1986); **37**, 546 (1988).

[34] M. S. Morris, K. S. Thorne, and U. Yurtsever, Phys. Rev. Lett. **61**, 1446 (1988).

[35] U. Yurtsever, Class. Quantum Grav. **7**, L251 (1990).

[36] G. Klinkhammer, Phys. Rev. D **43**, 2542 (1991).

[37] R. Wald and U. Yurtsever, Phys. Rev. D **44**, 403 (1991).

[38] The strict condition is that $R_{ab}N^aN^b$ is strictly positive for all null vectors N^a. Under this condition, the light rays from even a marginally trapped surface will focus.

[39] My attention was drawn to this aspect of Gödel's solution in 1982 by R. Penrose (private communication).

[40] The expansion rate needed for the proof to go through will be too high to hold in many cases of interest.

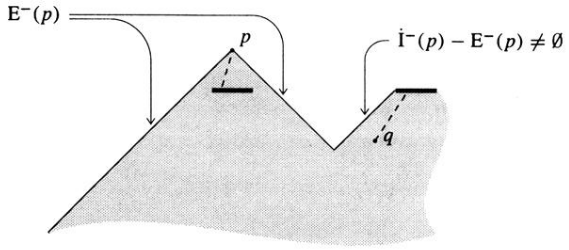

FIG. 1. An example of the causal complications that can arise in an unrestricted spacetime. Light rays travel along 45° lines in this diagram, and the two thick horizontal lines are identified. This allows the point q to send a signal to the point p along the dashed line, as shown, even though q lies outside what is usually considered the past light cone of p. The boundary of the past of p, $\dot{I}^{-}(p)$, then consists of the past light cone of p, $E^{-}(p)$, plus a further piece. Such a spacetime is not "causally simple."

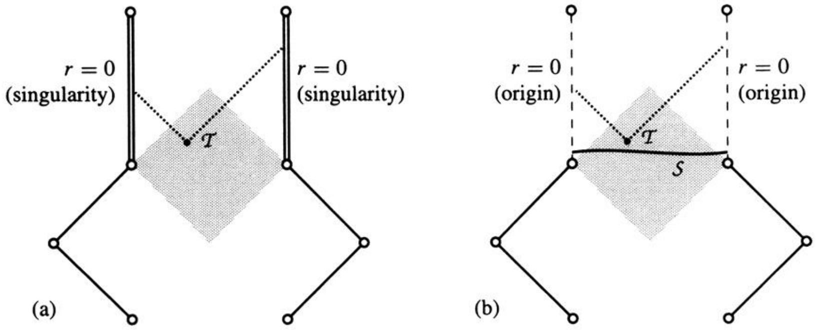

FIG. 3. The global structure of portions of (a) the Reissner-Nordström spacetime, and (b) Bardeen's spacetime. A point in the interior of both spacetimes represents a two-sphere. The boundaries of the diagram are drawn according to these conventions: single lines and hollow circles represent regions at infinity, double lines represent singularities, and dashed lines represent the origins ($r = 0$) of the coordinate systems. So, none of the boundaries in (a) are a part of the spacetime, whereas in (b) the $r = 0$ lines are. In both cases the $r = 0$ lines represent the origins of different coordinate patches. If one imagines a series of horizontal lines across each diagram, representing spacelike hypersurfaces, the topology of these surfaces will be $S^2 \times R$ throughout in (a), but in (b) the surfaces switch from $S^2 \times R$ to S^3 in the region between the $r = 0$ lines. (For instance, the surface $\mathcal{S}$ shown in (b) is a three-sphere.) There are trapped surfaces, $\mathcal{T}$, represented above by solid dots, in the shaded regions of both spacetimes. The dotted lines emanating from the trapped surfaces represent the two systems of future-directed null rays from $\mathcal{T}$: the "ingoing" and the "outgoing." Each system approaches (and in Bardeen's spacetime, reaches) a focal point at $r = 0$. Thus, in Bardeen's spacetime [i.e., in (b)], the future light cone of $\mathcal{T}$ "wraps around the universe." This light cone has topology S^3.

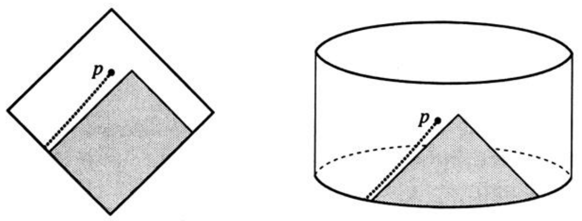

FIG. 5. Two examples of spacetimes with localized past light cones. The boundaries of both figures are boundaries at infinity. A typical past light cone is shown in each spacetime. The dotted line from each point p represents a past-directed timelike curve that avoids intersecting the past light cone.

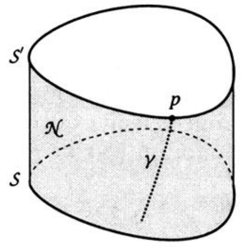

FIG. 6. An illustration of the strategy used in lemma 5.