June 30, 1994

TOPOLOGY CHANGE IN CLASSICAL GENERAL RELATIVITY

Arvind Borde†*

Institute of Cosmology, Department of Physics and Astronomy,
Tufts University, Medford, MA 02155

Abstract

This paper clarifies some aspects of Lorentzian topology change, and it extends to a wider class of spacetimes previous results of Geroch and Tipler that show that topology change is only to be had at a price. The scenarios studied here are ones in which an initial spacelike surface is joined by a connected "interpolating spacetime" to a final spacelike surface, possibly of different topology. The interpolating spacetime is required to obey a condition called *causal compactness*, a condition satisfied in a very wide range of situations. No assumption is made about the dimension of spacetime. First, it is stressed that topology change is kinematically possible; i.e., if a field equation is not imposed, it is possible to construct topology-changing spacetimes with non-singular Lorentz metrics. Simple 2-dimensional examples of this are shown. Next, it is shown that there are problems in such spacetimes: Geroch's closed-universe argument is applied to causally compact spacetimes to show that even in this wider class of spacetimes there are causality violations associated with topology change. It follows from this result that there will be causality violations if the initial (or the final) surface is not connected, even when there is no topology change. Further, it is shown that in dimensions ≥ 3 causally compact topology-changing spacetimes cannot satisfy Einstein's equation (with a reasonable source); i.e., there are severe dynamical obstructions to topology change. This result extends a previous one due to Tipler. Like Tipler's result, it makes no assumptions about geodesic completeness; i.e., it does not

† *Permanent address*: Long Island University, Southampton, NY 11968, and High Energy Theory Group, Brookhaven National Laboratory, Upton, NY 11973.

∗ *Electronic mail*: borde@bnlcl6.bnl.gov

permit topology change even at the price of singularities (of the standard incomplete-geodesic variety). Brief discussions are also given of the restrictions that are placed on the source in this result, of the possibility of creating exotic differentiable manifolds, and of ways in which the results of this paper might be circumvented.

PACS numbers: 02.40.-k, 04.20.Gz, 04.20.-q

I. Introduction

If space has a certain topology at some initial time, can the topology be different at a later time? The question has, of course, animatedly been discussed before by a number of people [1–23], mostly in the context of quantum gravity. In a quantum theory of gravitation, it has been argued [1–3], we will necessarily have to consider fluctuations not only in geometry but also in topology. Topology change is also interesting for another reason. It has sometimes been suggested [2, 4, 12] that the particles of ordinary matter might possibly be viewed as kinks or knots in space. In this approach, nontrivial topological configurations of space (sometimes called *geons*) are shown to display such particle-like aspects as mass and charge [4] and even half-integral spin [10]. Such theories would describe the creation and annihilation of particles by allowing the spatial topology to change. And finally, purely as a question about the classical Einstein theory, we may ask: since geometry evolves in general relativity, might it not be possible that topology does as well?

Topology change is interesting, therefore, for a number of somewhat separate reasons. There is, however, a widespread notion that the process is intrinsically incompatible with a Lorentzian metric. This notion is quite incorrect. But, although Lorentzian topology change is possible, it is problematic: it has been known for some time that both in closed universes and in certain open ones there are problems associated with topology change. Geroch [5, 6] has shown that topology change may be obtained in these cases only at the price of causality violations, and Tipler [7, 8] has shown that Einstein's equation cannot hold (with a source with non-negative energy density) on such spacetimes if the spatial topology changes. Part of the point of this paper is to extend the results of Geroch and Tipler to a wider class of spacetimes, and part of the point is to address some of the more widespread misconceptions about topology change. (These misconceptions are listed in section IX.) Along the way a few minor novelties – such as an explicit example of non-singular 2-dimensional topology change – will be added to the general topology-change discourse.

A] A mechanism for topology change: classical evolution

To study topology change we first have to pick a mechanism through which the topology might change. The most obvious choice is "classical evolution;" i.e., a scheme in which the initial and final spatial configurations, represented respectively by hypersurfaces $\mathcal{S}_1$ and $\mathcal{S}_2$, are joined by a connected interpolating manifold $\mathcal{M}$ with a Lorentzian metric defined on it (with respect to which $\mathcal{S}_1$ and $\mathcal{S}_2$ are spacelike).

Now, it may be tempting to argue, especially with the keen hindsight provided by the Geroch and Tipler results, that classical evolution is a patently flawed mechanism – particularly if ones chief interest is quantum gravity. But, for at least one approach to quantum gravity – the path integral approach – the question of finding "classical paths" between the initial and final configurations is an important one. (The paths discussed here are, however, Lorentzian ones, and so the results of this paper are probably irrelevant to the Euclidean path integral approach to quantum gravity [3].) The existence of classical paths is also important for work on geons [13], although here, too, Euclidean paths have been considered [15]. And if we are interested in topology change in the classical Einstein theory, then classical evolution is (by definition) the mechanism to consider.

The appropriateness of classical evolution may be questioned at another level. If the interpolating spacetime $\mathcal{M}$ can be foliated (i.e., sliced) into a family of spacelike hypersurfaces (which corresponds to our immediate intuitive notion of an initial spatial configuration evolving into some final configuration), then it seems obvious that if the topology is to change, $\mathcal{M}$ must contain cuts or holes of some kind (or some type of singularity), and therefore be an unacceptable model for spacetime. This statement can indeed be made precise and proved (this is done in section IV, in a corollary to theorem 1), but it still does not rule out classical evolution as a mechanism.

There are two ways in which classical evolution may be rescued. One way is to allow the metric to become degenerate at isolated points [13, 20, 24], while still retaining the ability to foliate $\mathcal{M}$. This approach will be briefly mentioned again in section IX. For the bulk of this paper, however, another way out is taken. Here, the metric is required to behave itself everywhere, but the ability to foliate is given up. This approach is based on the point of view that there is no compelling reason why the spacetime that interpolates between the initial and final states (which are the ones of physical interest) must be restricted to be of the type that admits a foliation. Indeed, there are solutions of Einstein's equation that do not admit a foliation into spacelike hypersurfaces anywhere (e.g., the Gödel solution), or which admit foliations in some regions but not in others (e.g., the Taub-NUT solution) [25]. Once we allow these kinds of interpolating spacetimes, then at the "kinematical" level (i.e., before imposing Einstein's equation) it is easy to construct – as we shall see below – nonsingular topology-changing spacetimes.

B] An overview of the paper

After some preliminary comments in section II, the kinematics of topology change is discussed in sections III and IV. A class of spacetimes, called *causally compact* spacetimes, is defined in section III and is shown to cover many cases of physical interest. It is also argued that some restriction on the interpolating spacetime, along the lines of causal compactness, is indeed necessary. It is demonstrated that many topology-changing causally compact spacetimes exist, and some 2-dimensional examples are shown. Then, in section IV, a direct extension of Geroch's theorem [5, 6] is presented which shows that such topology-changing spacetimes must contain closed timelike curves. Further, if causality violations are excluded, then the initial and final surfaces must each be connected. In section V, the dynamics of topology change is discussed and an extension of Tipler's theorem [7, 8] is presented that shows that the process of topology change cannot be described by Einstein's equation (even if the spacetime has incomplete geodesics). In this result certain assumptions are made about the curvature of spacetime. The most stringent of these assumptions is discussed in section VI, along with some reasonable conditions on the source in Einstein's equation (the "energy conditions") that will justify it. Upto this point in the paper most of the discussion has assumed that the metric is time-orientable (i.e, that the future can be globally distinguished from the past). In section VII it is shown that dropping this assumption does not materially alter the conclusions of the main theorems (theorems 1 and 3) of this paper. A few mathematical comments are made in section VIII, and some concluding remarks in section IX.

II. Notation, Definitions and other Preliminaries

The conventions and notation that I use are those of Hawking and Ellis [25]; the proofs of all the assertions that I make below may be found there (unless otherwise indicated), as can further references to the original statements and proofs of these assertions.

A *Lorentzian metric* on an n-dimensional manifold is a metric with signature $(-, +, \ldots, +)$. Einstein's equation is

$$R_{ab} - \frac{1}{2} R g_{ab} + \Lambda g_{ab} = 8\pi T_{ab},$$

where R_{ab} is the Ricci tensor obtained from the metric g_{ab}, R the curvature scalar, Λ the cosmological constant, and T_{ab} the matter energy-momentum tensor. Einstein's equation is used in a very minor way in this paper, only to justify the

assumptions of theorem 3. It will not be necessary to make any assumptions about whether or not the cosmological constant Λ is zero.

The manifolds I consider are smooth, Hausdorff and paracompact. A smooth manifold is one that is, essentially, infinitely differentiable (i.e., C^∞; a brief discussion of this is given in section VIII). The other two restrictions are technical ones; they ensure that our models for spacetime are mathematically well-behaved. These two restrictions will not appear explicitly in the rest of this paper. On such manifolds it is always possible to define a smooth positive-definite metric; this facility will be used below. The manifolds are assumed to be without boundary, unless explicitly stated otherwise.

A] Interpolating spacetimes and time-orientation

When studying topology change, the region of interest is the spacetime strip between an initial surface and a final surface. This strip is usually part of some "full spacetime," but it will not be necessary at any stage for us to refer to this larger spacetime – i.e., we are only interested in what happens between the initial and final states, not in what has happened before or is to happen after. I consider in this paper situations that involve (roughly speaking) either a "fully spatially extended" strip of spacetime or a portion of it confined to a timelike tube. (See fig. 1.) This rough idea is captured in the following definitions:

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be disjoint $(n-1)$-dimensional manifolds, possibly with boundary. A connected n-dimensional manifold $\mathcal{M}$, with boundary $\partial\mathcal{M}$, is called an *interpolating manifold* between $\mathcal{S}_1$ and $\mathcal{S}_2$, if there is a further $(n-1)$-dimensional manifold $\mathcal{T}$ (possibly empty) such that $\mathcal{T}$ is compact and $\partial\mathcal{M} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{T}$. I.e., the boundary of $\mathcal{M}$ is allowed, in addition to $\mathcal{S}_1$ and $\mathcal{S}_2$, to have an extra (compact) component $\mathcal{T}$. This component may be empty, but if it is not, then it is required that (i) $\mathcal{T} \cap \mathcal{S}_i = \partial\mathcal{S}_i \neq \emptyset$ $(i = 1, 2)$, (ii) $\partial\mathcal{S}_1$ and $\partial\mathcal{S}_2$ have the same topology (more precisely, are diffeomorphic – see section VIII for a definition of the term), and (iii) $\mathcal{T}$ is diffeomorphic to $\partial\mathcal{S}_1 \times [0, 1]$. The role played by $\mathcal{T}$ will be seen shortly. Though $\mathcal{M}$ is required to be connected, $\mathcal{S}_1$ and $\mathcal{S}_2$ are not restricted in this way.

The interpolating manifold $\mathcal{M}$ is called an *interpolating spacetime* between $\mathcal{S}_1$ and $\mathcal{S}_2$, if, in addition, there is a smooth Lorentz metric on it with respect to which $\mathcal{S}_1$ and $\mathcal{S}_2$ are spacelike and $\mathcal{T}$, if non-empty, is timelike in this sense: it is possible to define a smooth, nowhere vanishing, timelike vector field everywhere tangent to $\mathcal{T}$ such that each integral curve of this field takes on one endpoint at $\mathcal{S}_1$ and another at $\mathcal{S}_2$. (Integral curves of a vector field are, roughly, curves to which the field is tangent; a more precise definition is given in the proof of Theorem 1.) If the surface $\mathcal{T}$ is non-empty, it represents a timelike tube within

which the topology change, if any, must occur. If $\mathcal{T}$ is empty, then it means that we are looking at a "fully spatially extended" strip of spacetime. The two scenarios are illustrated in fig. 1.

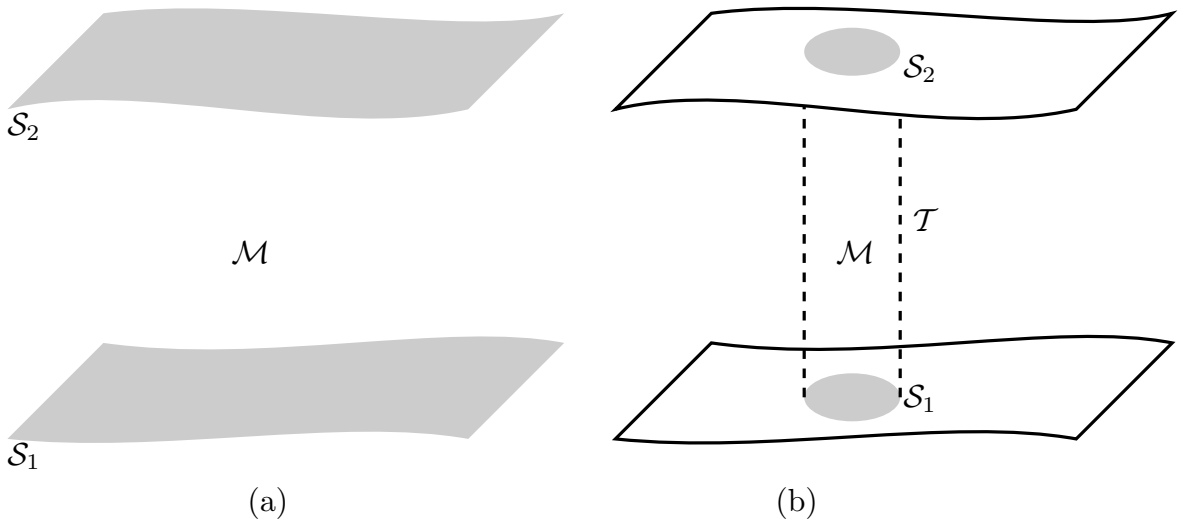

Figure 1: The two types of scenarios studied in this paper: (a) a fully spatially extended spacetime strip, and (b) a portion of such a strip confined to a timelike tube. In each case, the shaded regions $\mathcal{S}_1$ and $\mathcal{S}_2$ represent, respectively, the initial and final surfaces, and $\mathcal{M}$ the interpolating spacetime. In (b), $\mathcal{T}$ represents the timelike tube within which the topology change, if any, must occur.

The surface $\mathcal{S}_1$ will be called an *initial surface* for the interpolating spacetime $\mathcal{M}$ if no $p \in \mathcal{M}$ lies to the past of a point on $\mathcal{S}_1$, and $\mathcal{S}_2$ will be called a *final surface* if no $p \in \mathcal{M}$ lies to the future of a point on $\mathcal{S}_2$. It is possible that either of $\mathcal{S}_1$ or $\mathcal{S}_2$ may be empty: these cases will correspond, respectively, to the creation and the annihilation of the universe. (If they are both empty, $\mathcal{M}$ may be said to describe a situation in which nothing comes from nothing.)

These definitions of initial and final surfaces assume that it is possible to consistently distinguish future from past throughout $\mathcal{M}$. A metric that allows such a global choice of future and past is called *time-orientable* (locally any Lorentz metric permits such a choice: see ref. [25], p. 38–39). If such a choice of future direction has actually been made, the metric is called *time-oriented*. It is assumed that the metrics being considered are time-oriented. No assumptions are made directly about the orientability of the underlying manifold. (A non-

orientable manifold like a Möbius strip may admit a time-orientable metric, and an orientable manifold like an annulus may admit a metric that is not time-orientable [26]). If a result does depend on the orientability of the manifold, I will explicitly point this out. Questions of orientability are discussed again a little in section VII, especially the question of non-time-orientable metrics.

B] Lorentz metrics and vector fields

The existence of a time-orientable, smooth Lorentz metric g_{ab} on a manifold $\mathcal{M}$ (with or without boundary) is equivalent to the existence of a smooth, nowhere vanishing vector field, V^a, on $\mathcal{M}$ (ref. [25], p. 39 and p. 181). To see this, first choose a positive-definite metric h_{ab} on $\mathcal{M}$. Given V^a, define $g_{ab} = (h_{cd}V^cV^d)h_{ab} - 2h_{ac}V^ch_{bd}V^d$. This g_{ab} will be a Lorentz metric and V^a will be timelike with respect to it. The direction of V^a can be used to globally define a future (or a past) direction for time.

Conversely, let g_{ab} be a time-oriented Lorentz metric, and at any point $p \in \mathcal{M}$ consider the set $\{U^a\}$ of unit future-directed timelike vectors. A unique member of this set will minimize $h_{ab}U^aU^b$. For, suppose that there are two future-directed unit timelike vectors, V_1^a and V_2^a, such that $h_{ab}V_1^aV_1^b = K^2 = h_{ab}V_2^aV_2^b$. Let $M = h_{ab}V_1^aV_2^b$ and $L^2 = -g_{ab}V_1^aV_2^b$. Then, $K^2 > M$ and $L^2 > 1$. It follows from this that any unit timelike vector of the type $\hat{V}^a = \alpha V_1^a + \beta V_2^a$, where $\alpha, \beta > 0$, obeys $h_{ab}\hat{V}^a\hat{V}^b < K^2$. Thus the vector that minimizes $h_{ab}U^aU^b$ must be unique. This gives the vector field V^a.

Both constructions (g_{ab} and V^a) depend on an entirely arbitrary choice of positive-definite metric – different choices for h_{ab} will give different Lorentz metrics (and different vector fields). The constructions are standard and I go into them only because the existence of the vector field V^a is used repeatedly as a tool in this paper: either to show that some manifold admits a Lorentz metric, or, given a metric, to explore its properties.

If $\mathcal{M}$ has a boundary $\partial\mathcal{M}$, some components of which are spacelike, then V^a cannot be tangent to these components. Further, if $\partial\mathcal{M}$ has a timelike component $\mathcal{T}$, then V^a can be deformed so as to be tangent to this component: Let T^a be the timelike vector field tangent to $\mathcal{T}$ (chosen to be future-pointing) and let S_i^a ($i = 1, \ldots, n-1$) be a set of mutually orthogonal spacelike vectors, also orthogonal to T^a. Choose these vectors so that S_1^a is orthogonal to $\mathcal{T}$ and the other S_i^a are tangent to it. At each point $p \in \mathcal{T}$ let α_p be a small portion of the spacelike geodesic with initial tangent S_1^a. Let s be a parameter on these curves, chosen to be zero at $\mathcal{T}$ (and to vary smoothly near $\mathcal{T}$). The curves α_p will not intersect each other close to $\mathcal{T}$, i.e, in some parameter range $[0, \epsilon_p)$. Then V^a may be modified close to $\mathcal{T}$ by replacing it with $\hat{V}^a = (s/\epsilon_p)V^a + (1 - s/\epsilon_p)T^a$

along the curves α_p. This, along with a suitable smoothing at ϵ_p, is the required deformation of V^a.

I will assume for the rest of this paper that V^a is chosen to be tangent to $\mathcal{T}$ (at $\mathcal{T}$). This means that the integral curves of V^a must either lie in $\mathcal{T}$ throughout or not intersect it all.

C] Causal functions

We will also need to use some of the causal functions of global general relativity [25]. A C^1 curve is a curve $x^\mu(\tau)$ such that the components $dx^\mu/d\tau$ of the tangent to the curve exist and are continuous. A *timelike curve* is defined to be a non-degenerate (i.e., not just a single point) C^1 curve with a tangent that is everywhere timelike. A *null curve* is similarly defined, except that degenerate curves are permitted here. A set is called *achronal* if no two points in it can be connected by a timelike curve. Timelike and null curves are called *future-directed* if their tangents point somewhere (and therefore, since they are C^1, everywhere) in the future time direction. *Past-directed* curves are defined similarly. Given a point $p \in \mathcal{M}$, the *chronological future* of p is defined by

$$I^+(p) = \{q \mid \exists \text{ a future-directed timelike curve from } p \text{ to } q\}.$$

The *chronological past* of p, $I^-(p)$, is similarly defined. The sets $I^\pm(p)$ are open sets. (There is a slightly subtle point here: the "openness" of sets in the interpolating manifold $\mathcal{M}$ is with respect to the topology induced on it by the full spacetime manifold in which $\mathcal{M}$ lies. But the interpolating spacetimes that we are considering typically have boundaries; in such cases an open set around a boundary point of $\mathcal{M}$ will itself, roughly speaking, "have a partial boundary" at its intersection with $\partial\mathcal{M}$.) For $\mathcal{A} \subset \mathcal{M}$, the chronological futures and pasts are defined by:

$$I^\pm(\mathcal{A}) = \bigcup_{p\in\mathcal{A}} I^\pm(p).$$

Clearly, these sets can be defined only in a time-oriented spacetime. But an analogous set may be defined in general:

$$I(p) = \{q \mid \exists \text{ a timelike curve between } p \text{ and } q\}.$$

In a time-oriented spacetime we will have $I(p) = I^+(p) \cup I^-(p)$.

Another useful causal function is the *future domain of dependance* of a connected achronal spacelike hypersurface $\mathcal{S}$:

$$D^+(\mathcal{S}) = \{q \mid \text{every past-directed timelike or null curve from } q \text{ eventually meets } \mathcal{S} \text{ when extended into the past}\}.$$

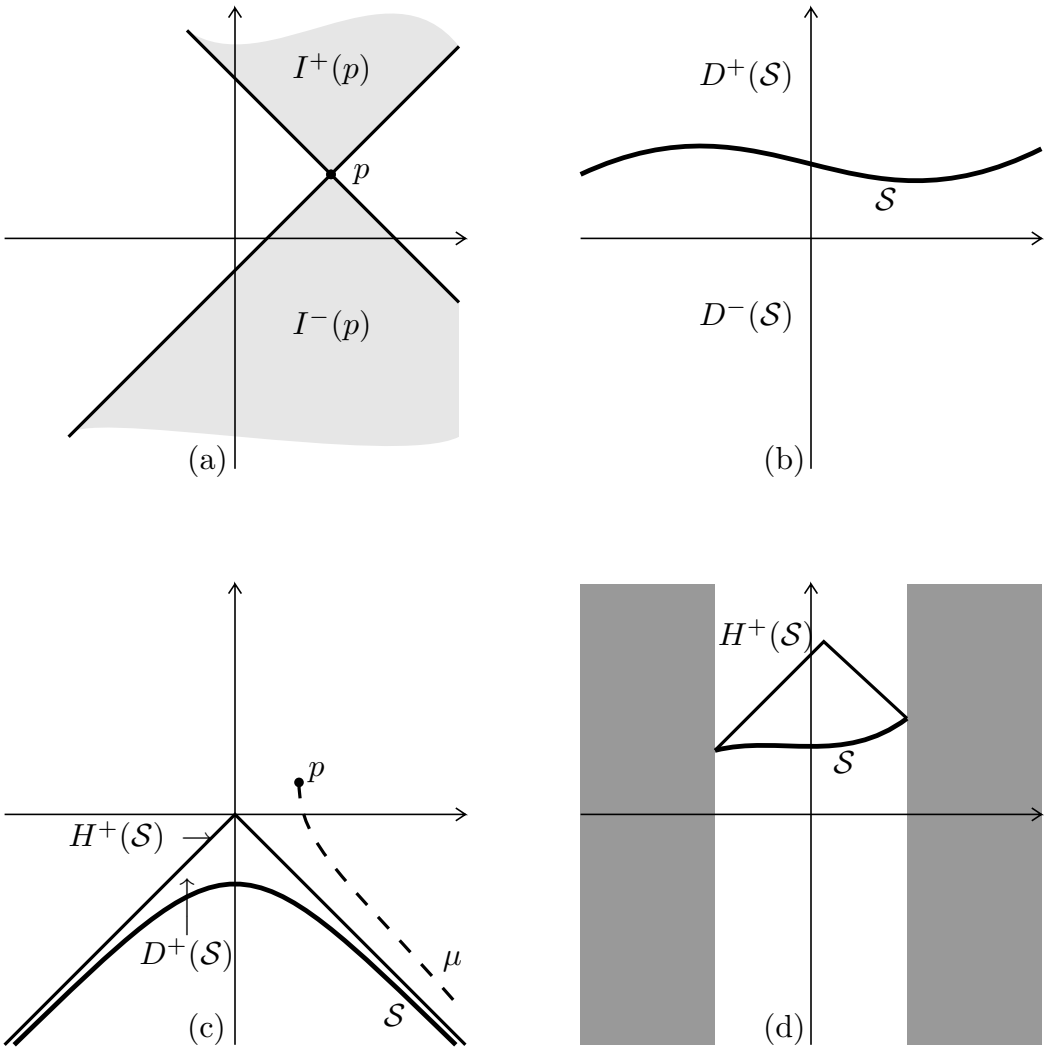

Figure 2: Some of the causal sets that are used in global general relativity. All figures are based on 2-dimensional Minkowski spacetime (with time pointing 'upward'). Figure (a) shows the future and past of a point p. Figure (b) shows the future and past domains of dependance of a spacelike surface $\mathcal{S}$. Figure (c) shows the future domain of dependance of another spacelike surface $\mathcal{S}$; here there is a Cauchy horizon because there are points p to the future of $\mathcal{S}$ from which there are past-directed timelike curves (such as the curve μ) that do not intersect $\mathcal{S}$. This Cauchy horizon exists because $\mathcal{S}$ is badly chosen. But in (d), if the shaded regions are deleted, then the existence of Cauchy horizons is intrinsic to the truncated spacetime. Though this is a contrived example, pretty much the same behavior occurs, for instance, in anti-de Sitter spacetime.

The idea that this definition tries to capture is that the events that occur in $D^+(\mathcal{S})$ are determined entirely by initial data on $\mathcal{S}$. The *past domain of dependance*, $D^-(\mathcal{S})$, is defined similarly. Now, clearly $D^\pm(\mathcal{S}) \subset I^\pm(\mathcal{S})$. In many cases it happens that $\mathcal{S}$ is such (either because it is badly chosen, or because of some intrinsic property of the spacetime) that there are points to its future (i.e., in $I^+(\mathcal{S})$) that do not lie in $D^+(\mathcal{S})$. Then $D^+(\mathcal{S})$ has a future boundary, called the *future Cauchy horizon*, defined as

$$H^+(\mathcal{S}) = \overline{D^+(\mathcal{S})} - I^-(D^+(\mathcal{S})).$$

The *past Cauchy horizon*, $H^-(\mathcal{S})$, is similarly defined. Examples of most of these sets are shown in fig. 2. Various of their properties that are needed will be stated as the occasion arises.

III. The Kinematics of Topology Change I: Examples and Restrictions

Given two $(n-1)$-dimensional manifolds $\mathcal{S}_1$ and $\mathcal{S}_2$ (possibly of different topologies), under what conditions does there exist an interpolating spacetime $\mathcal{M}$ between them? If $\mathcal{M}$ exists, what properties must it have?

A] Connected sums

To discuss the first of these questions, it is useful to have on hand the following technique for constructing new manifolds out of old. Given any two n-dimensional manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ (with or without boundary), their *connected sum* is formed by deleting an open n-disc from the interior of each of $\mathcal{M}_1$ and $\mathcal{M}_2$ and identifying them along the resulting boundaries [12, 27]. The process is illustrated in fig. 3.

Using this construction it is possible to show that if no restrictions are placed on the interpolating spacetime, then it always exists. The existence, first, of an interpolating manifold is easily seen: given $(n-1)$-dimensional manifolds $\mathcal{S}_1$ and $\mathcal{S}_2$, let $\mathcal{M}_i = \mathcal{S}_i \times [0, \infty)$; then $\mathcal{M}' = \mathcal{M}_1 \# \mathcal{M}_2$ is the required manifold. This can now be modified to admit a Lorentz metric. Observe first that each of $\mathcal{M}_1$ and $\mathcal{M}_2$ admits a natural vector field along the $[0, \infty)$ lines. Pick the vector fields to point into $\mathcal{M}_1$ on $\mathcal{S}_1$ (i.e., in the direction of increasing $t \in [0, \infty)$ on $\mathcal{M}_1$) and out of $\mathcal{M}_2$ on $\mathcal{S}_2$ (i.e., in the direction of decreasing $t \in [0, \infty)$ on $\mathcal{M}_2$). On $\mathcal{M}_1$, let $\mathcal{D}_1$ be the disc that was removed for the purposes of constructing $\mathcal{M}'$. Cut out a shell almost completely surrounding $\mathcal{D}_1$ and modify the field near $\mathcal{D}_1$ so as to point into it (see fig. 4). Do the same with $\mathcal{M}_2$, but this time modify the vector field so that it points out of $\mathcal{D}_2$. The new connected sum, $\mathcal{M}$, will have a vector

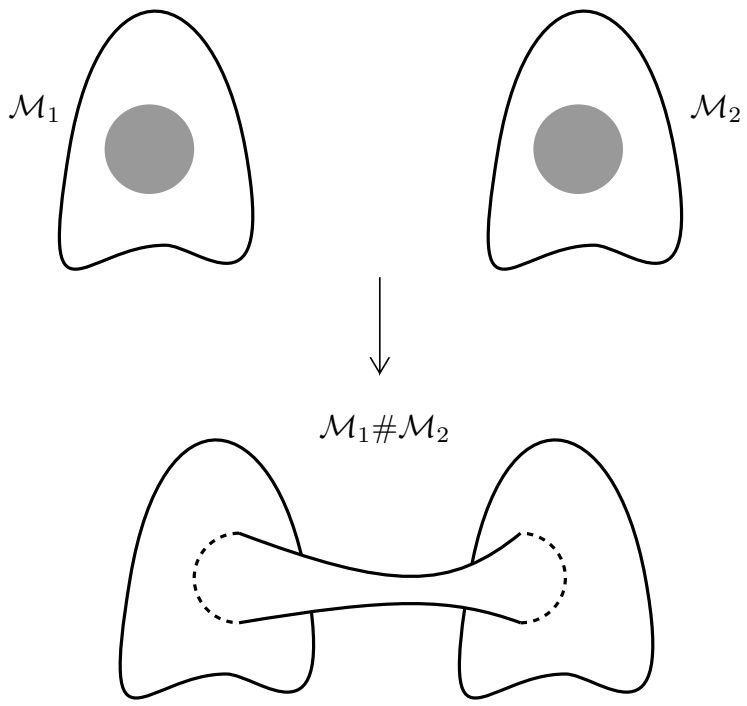

Figure 3: The connected sum of two manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$. An open disk – represented by a shaded region above – is removed from each of the manifolds and the edges of the two disks are identified.

field on it from which a Lorentz metric may be constructed. $\mathcal{S}_1$ will be an initial surface and $\mathcal{S}_2$ a final surface for this metric (if the direction of the vector field is used to define the future direction of time).

B] Causal compactness

The construction discussed above is very arbitrary. Since the interpolating spacetime $\mathcal{M}$ is not constrained in any way, we could continue making cuts and identifications in it as we choose. This is not a very satisfactory state of affairs and it is important, therefore, to restrict $\mathcal{M}$ in some reasonable manner. The sort of restriction that is desirable would be one that does not allow points of $\mathcal{M}$ to have causal access to holes or to "regions at infinity." Such a restriction will have to be made on the interpolating spacetime, not just on the underlying manifold. Consider, for example, the manifold $[0, 1] \times R^3$. A Minkowski metric can be put on it such that the two boundaries correspond to constant-t surfaces (where t is the usual time co-ordinate). There are no holes in this spacetime, and no point in

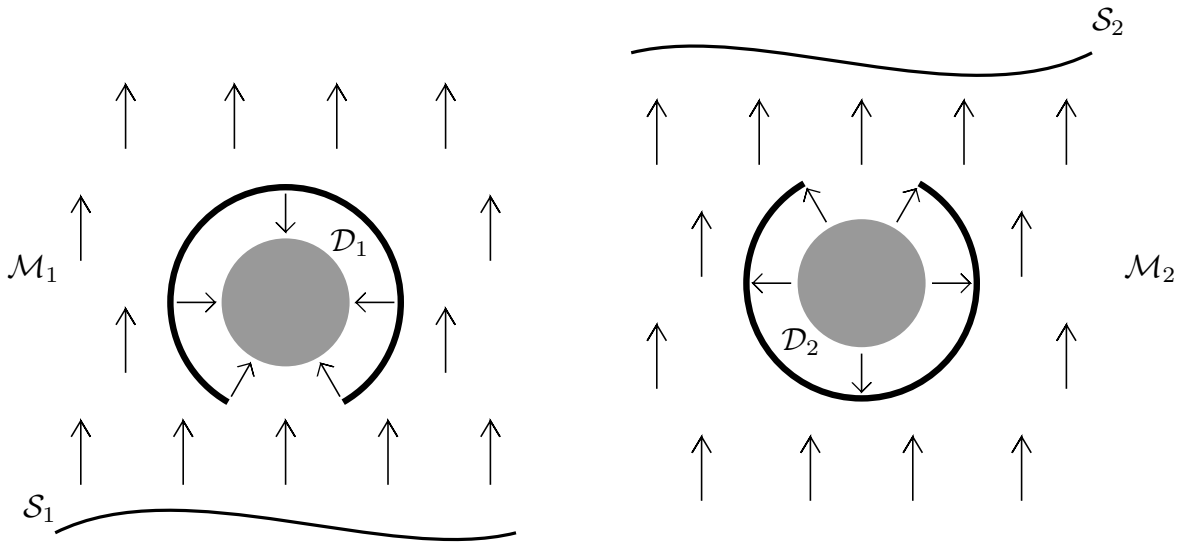

Figure 4: How the timelike vector field is set up on $\mathcal{M}_1$ and $\mathcal{M}_2$. The discs $\mathcal{D}_1$ and $\mathcal{D}_2$ (shown shaded), and the almost-complete shells surrounding them (shown as thick arcs), are deleted from the manifold. The edges of the two discs are identified in the standard connected-sum construction illustrated earlier.

it can be causally connected to a region at infinity. But, the same manifold also admits an anti-de Sitter metric. A time coordinate may be picked here such that the two boundaries again correspond to constant-t surfaces, but with points in $\mathcal{M}$ now able to receive signals from points at infinity (ref. [25], p. 131–134). (Fig. 2d illustrates such a spacetime. Between any two constant-t surfaces in it, however close they are to each other, there will be points that can receive signals from the boundary with the deleted region.) Now, this particular class of situations may be handled by the freedom that we have left ourselves in the definition of an interpolating spacetime: since the boundary ∂M may contain a further component, the timelike tube $\mathcal{T}$, the region near infinity can be put outside this tube. I will return to this when I discuss open universes. But the general problem of arbitrary cuts and holes in $\mathcal{M}$ is more serious: it cannot be handled, in general, by putting all the "bad parts" outside timelike tubes and concentrating only on what happens inside. Such problematic situations can all be excluded, however, if we require that the interpolating spacetime obey a condition called *causal compactness*. This condition is meant to take away the license to arbitrarily make cuts and holes in spacetime:

Definition: *A spacetime $\mathcal{M}$ is called causally compact if for any $p \in \mathcal{M}$, $\overline{I(p)}$ is compact.*

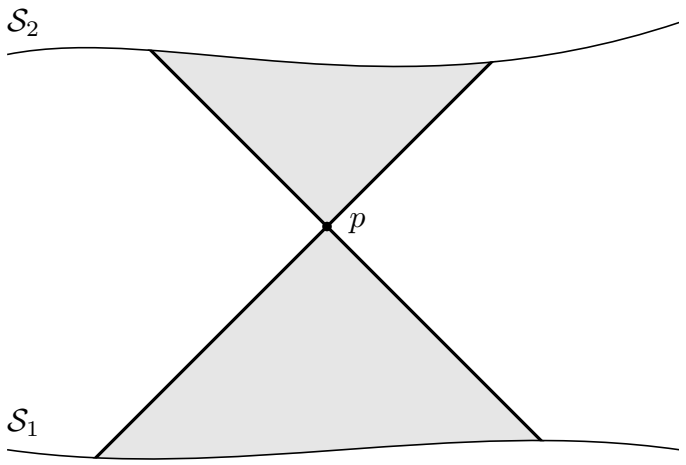

Figure 5: A causally compact spacetime. Points p between $\mathcal{S}_1$ and $\mathcal{S}_2$ cannot receive signals from, or send signals to, either regions at infinity or "holes" in the spacetime.

This definition has been phrased in such a way as to be applicable both to spacetimes that are not time-orientable, as well as to ones that are. If the spacetime is time-orientable we have $I(p) = I^+(p) \cup I^-(p)$. Thus $\overline{I^+(p)}$ and $\overline{I^-(p)}$, being closed subsets of the compact set $\overline{I(p)}$, must each be compact. The idea of causal compactness is illustrated in fig. 5.

There are many classes of situations in which $\mathcal{S}_1$ and $\mathcal{S}_2$ can be connected by a causally compact interpolating spacetime. They are studied systematically below.

C] Closed universes

In a closed universe, $\mathcal{S}_1$ and $\mathcal{S}_2$ are (by definition) both closed surfaces (i.e., compact without boundary) and it is reasonable to require that $\mathcal{M}$ be compact as well. (Thus there is no timelike component to the boundary of $\mathcal{M}$ here; i.e., $\mathcal{T} = \emptyset$.) In this case, $\mathcal{M}$ will trivially be causally compact with respect to any Lorentz metric that it admits. To study such situations we can draw on results from cobordism theory [28]. If a compact manifold $\mathcal{M}$ interpolates

between closed manifolds $\mathcal{S}_1$ and $\mathcal{S}_2$, then $\mathcal{M}$ is called a *cobordism*, and $\mathcal{S}_1$ and $\mathcal{S}_2$ are called *cobordant*. The condition for $\mathcal{S}_1$ and $\mathcal{S}_2$ to be cobordant is that their Stiefel-Whitney numbers (these numbers characterize some of the properties of a manifold) be the same [28]. If $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{M}$ are required to be oriented, then their Pontryagin numbers must be equal as well [28]. These requirements are satisfied in a wide variety of cases: e.g., when $n = 2, 3, 4, 7$ or 8, any two $(n-1)$-dimensional closed oriented manifolds are cobordant through an oriented cobordism. In fact, when $n = 4$ (which, in the naive past, used to be considered the case of greatest physical interest), any two closed manifolds, oriented or not, are cobordant.

Given the existence of the cobordism $\mathcal{M}$, the next question is, is it *Lorentzian*? i.e., can an appropriate Lorentz metric be put on it? Equivalently, we can ask under what conditions it is possible to put a vector field on $\mathcal{M}$ that is nowhere tangent to $\mathcal{S}_1$ and $\mathcal{S}_2$. The condition for such a field to exist and to point outward everywhere (or inward everywhere) on $\partial\mathcal{M}$ is that $\chi(\mathcal{M}) = 0$, where $\chi(\mathcal{M})$ is the Euler characteristic of $\mathcal{M}$ [29]. We are, however, more interested in the case when the field points into $\mathcal{M}$ on $\mathcal{S}_1$ and out of $\mathcal{M}$ on $\mathcal{S}_2$. The conditions for the existence of such a field have also been obtained before, by Reinhart [30] and by Sorkin [13]: *Let n be the dimension of $\mathcal{M}$. If n is even, then $\mathcal{M}$ admits such a field if $\chi(\mathcal{M}) = 0$; if n is odd, then $\mathcal{M}$ admits the field if $\chi(\mathcal{S}_1) = \chi(\mathcal{S}_2)$.*

Observe that these conditions include the condition for the vector field to point outward everywhere (or inward everywhere) on $\partial\mathcal{M}$. For, that case can be regarded as representing the topological transition $\emptyset \to \partial\mathcal{M}(= \mathcal{S}_1 \cup \mathcal{S}_2)$. Then, both when n is even as well as when it is odd, the condition $\chi(\mathcal{M}) = 0$ can be recovered (using the result that $2\chi(\mathcal{M}) = \chi(\partial\mathcal{M})$ if $\mathcal{M}$ is odd-dimensional).

The condition when n is odd serves as a selection rule, which I will refer to as the Reinhart-Sorkin selection rule (or the RS rule, for short), and it forbids a number of topological transitions. To see this, consider the first few cases. (In the discussion below, S^n and T^n will stand for the n-dimensional sphere and torus respectively.)

• $n = 1$: This case is not very interesting, for there are topologically only four 1-manifolds [29]: S^1 (i.e., the circle); R^1; [0, ∞); and [0, 1].

• $n = 3$: The RS rule forbids all Lorentzian topological transitions between oriented closed 2-manifolds (except, of course, the identity transition). This is so, because any such manifold may be obtained by adding a certain number of handles to S^2. The number of handles is called the *genus*, g, and it completely characterizes the topology of the (oriented) manifold. (S^2 has g = 0; T^2 has g = 1; etc.) The Euler characteristic is related to the genus by $\chi = 2 - 2\mathrm{g}$ and is,

therefore, different for oriented closed 2-manifolds of different topologies.

• $n = 5$: Although some topological transitions may be possible here, Sorkin has shown [13] that the RS rule does not allow, for example, monopole pair creation in (5-dimensional) Kaluza-Klein theory.

• $n = 7$: Here is one example of topology change: S^6 and $(S^4 \times S^2)\#T^6$ are cobordant and both have $\chi = 2$. But the first of these is simply connected and the second is not, so the transition between the two indeed represents topology change.

In the even-dimensional case it is the interpolating spacetime that is restricted, not the initial and final surfaces. This suggests a further question: if a given cobordism does not have $\chi = 0$, can it be modified in some way so that χ now vanishes? Such a modification can be carried out by imitating a procedure due to Misner, and used by Geroch in 4 dimensions [5, 6]. The procedure is based on the following standard even-dimensional result:

$$\text{If } \mathcal{N} = \mathcal{M}\#\mathcal{V}, \text{ then } \chi(\mathcal{N}) = \chi(\mathcal{M}) + \chi(\mathcal{V}) - 2.$$

Now, if $\mathcal{V} = T^n$, then $\chi(\mathcal{V}) = 0$; and if $\mathcal{V} = S^2 \times S^{n-2}$ (for $n > 2$), then $\chi(\mathcal{V}) = 4$. Thus, if $\chi(\mathcal{M})$ is even to start with (in dimensions > 2), then by repeated application of $\#$ with the appropriate $\mathcal{V}$ it can be reduced to zero. If $\chi(\mathcal{M})$ is odd initially, and if $n = 4k$, we can use $\mathcal{V} = CP^2 \times CP^2 \times ...$ (k factors). Here CP^2 is complex projective 2-space, which is a real 4-dimensional closed manifold with $\chi = 3$. So $\chi(\mathcal{V}) = 3^k$; this $\mathcal{V}$ can be used to first make a manifold of even χ, to which the above procedure can then be applied. If $n = 4k + 2$, then there are no orientable closed manifolds of odd χ for us to use. But, if orientability is not a concern, then RP^n with $\chi = 1$ may be used in place of CP^2. And, finally, in the two dimensional case it is possible to enumerate all the possibilities explicitly [13]. They are listed separately below.

To summarize these results, suppose that $\mathcal{S}_1$ and $\mathcal{S}_2$ are cobordant and of dimension $(n - 1)$. Then, if n is odd, there is a spacetime that interpolates between them if and only if $\chi(\mathcal{S}_1) = \chi(\mathcal{S}_2)$ [13], and examples exist of such Lorentzian cobordisms. If $n = 2$, then again examples of topology change can be constructed, as shown below. And, if n is even and is greater than 2, then there is always an interpolating spacetime. (This result was found previously, with slightly different methods [30].) Thus, kinematically, though there are some constraints, a large variety of topology-changing closed universes exist. In particular, as Misner has pointed out (see [5]), since any two closed 3-manifolds are cobordant, there is at this stage no barrier whatsoever to four-dimensional topology change.

(It is worth noting here that Gibbons and Hawking [21] have recently found additional selection rules that further constrain the possible topological transitions that can occur.)

D] Two-dimensional topology change

The only Lorentzian cobordisms in two dimensions are the torus and the Klein bottle (both with no boundaries), the cylinder (with two S^1 boundaries), and the Möbius strip (with one S^1 boundary) [13]. The first two may be regarded as examples of a $\emptyset \leftrightarrow \emptyset$ transition. On the cylinder, the most obvious choice of metric makes it an example of the identity transition, $S^1 \leftrightarrow S^1$ (fig. 6a); but a metric may also be chosen on it (fig. 6b) so that it represents the transition $S^1 \cup S^1 \leftrightarrow \emptyset$. And on the Möbius strip, one choice of metric (fig. 6c) makes it an example of an $S^1 \leftrightarrow \emptyset$ transition [13]; another choice (non-time-orientable) is discussed in section VII. The spacetimes of fig. 6b and fig. 6c are explicit examples of Lorentzian topology change.

E] Open universes

Next, consider the case when $\mathcal{S}_1$ and $\mathcal{S}_2$ are non-compact. It seems reasonable to expect that many of these situations will also be causally compact, with $\mathcal{T} = \emptyset$, as they stand. For example, Minkowski spacetime is, and so are the open Friedmann cosmologies (if, in both cases, the boundaries $\mathcal{S}_1$ and $\mathcal{S}_2$ correspond to constant-t surfaces, where t is the usual time coordinate). But the freedom that comes from the extra component, $\mathcal{T}$, of the boundary of $\mathcal{M}$ allows us now to embrace even wider classes of situations.

One such situation of interest is the asymptotically flat spatial geometry of an isolated system. We expect to be able to compactify this situation by adding a point at infinity, or by imposing periodic boundary conditions ("putting it in a box"), and thus we might expect results similar to the compact case. For the 4-dimensional case Geroch has defined an *externally Euclidean* 3-manifold $\mathcal{S}$ to be one that contains a compact set $\mathcal{C}$ such that $\mathcal{S} - \mathcal{C}$ is diffeomorphic (see section VIII for a definition of the term) to $R \times S^2$. It is reasonable to assume that an isolated system can be represented thus. Then, given two externally Euclidean 3-manifolds $\mathcal{S}_1$ and $\mathcal{S}_2$, a spacetime $\mathcal{M}$ that interpolates between them is called *externally Lorentzian* if there is a compact set $\mathcal{K}$ such that $\mathcal{M} - \mathcal{K}$ is diffeomorphic to $S^2 \times R \times [1, 2]$ where $\mathcal{S}_t = S^2 \times R \times \{t\}$ is spacelike for each $t \in [1, 2]$, and $\gamma_p = \{p\} \times [1, 2]$ is timelike and future-directed for each $p \in S^2 \times R$. The idea behind this definition is that the topology change, if any, occurs within the compact set $\mathcal{K}$. Then, in 4 dimensions, in exact analogy with the compact

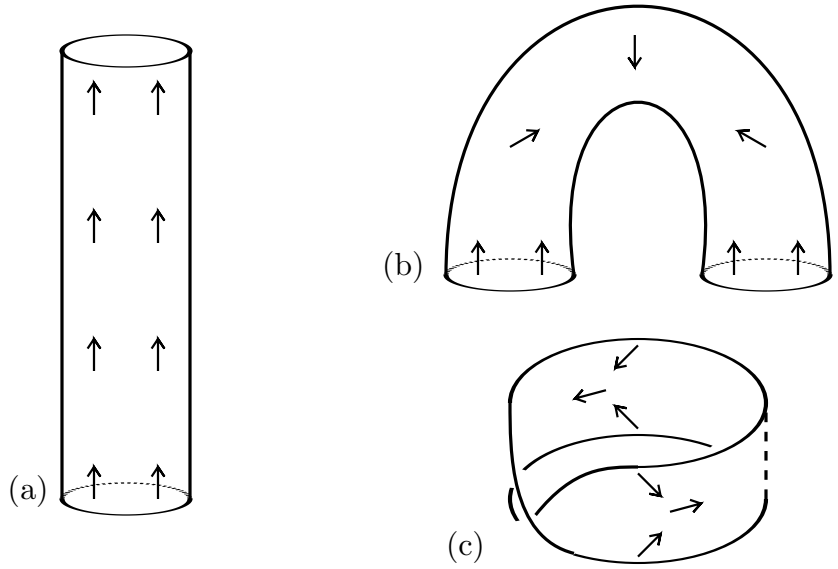

Figure 6: Examples of 2-dimensional Lorentzian cobordisms. The manifolds in (a) and (b) are cylinders, and the one in (c) a Möbius strip. (The spacetime of fig. (c) is due to Sorkin [13].) The arrows show the vector field V^a, from which the Lorentz metric may be constructed. The field V^a may be thought of as representing the future direction of time. The pictures in (b) and (c) are examples of topology change. They represent, respectively, the transitions $S^1 \cup S^1 \rightarrow \emptyset$, which may be thought of as universe pair-annihilation, and $S^1 \rightarrow \emptyset$, which may be thought of as universe self-annihilation. Reversing V^a gives the reverse transitions: universe pair-creation and universe self-generation.

case, any two externally Euclidean spaces can be connected by an interpolating externally Lorentzian spacetime [6].

Though this definition may be extended in a straightforward way to arbitrary numbers of dimensions, it is a little restrictive. The structure of an asymptotically flat space "close to infinity" in higher-dimensional theories would not, for instance, necessarily be $R \times S^{n-2}$. We might, in many cases, expect instead something like $R \times S^2 \times \mathcal{N}$ where $\mathcal{N}$ is the compact "internal manifold." In fact, examples of both kinds of behaviour at infinity are known [15]. It is also possible that we might have more complicated structures, not expressible as product manifolds globally. Another class of situations not necessarily covered above is the multi-geon class involving infinitely many topological kinks.

It is reasonable to suppose that all of these scenarios belong to a class of spacetimes that may be called *externally simple*:

Definition: *A spacetime $\hat{\mathcal{M}}$ that interpolates between two spacelike surfaces $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$ is called externally simple if*

a. $\hat{\mathcal{M}}$ contains a (possibly empty) causally compact region whose initial and final surfaces, $\mathcal{S}_1$ and $\mathcal{S}_2$, are subsets of $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$, respectively, and

b. $\hat{\mathcal{S}}_1 - \mathcal{S}_1$ is diffeomorphic to $\hat{\mathcal{S}}_2 - \mathcal{S}_2$.

The idea here is that nothing topologically interesting happens outside the causally compact part of an externally simple spacetime.

The results in the rest of this paper all deal with causally compact interpolating spacetimes. They are meant to apply to situations where the timelike component $\mathcal{T}$ of $\partial\mathcal{M}$ is empty, or, if non-empty, where $\mathcal{M}$ is a causally compact region of a larger, externally simple spacetime $\hat{\mathcal{M}}$.

IV. The Kinematics of Topology Change II: Causality Violations

We have seen so far that there appear to be few kinematical obstructions to topology change in general relativity. The question that next arises is, what are the properties of topology-changing spacetimes? It turns out that these properties are not pleasant. The first sign of this came with Geroch's discovery that causality violations are necessarily associated with topology change in the closed universe and externally Lorentzian cases [5, 6]. By "causality violation" it is meant that there is at least one closed timelike curve. This will happen, for instance, if there are points q_1 and q_2 such that $q_1 \in I^-(q_2)$ and $q_2 \in I^-(q_1)$.

The same result may be proved for causally compact interpolating spacetimes in general. The proof that I give below is slightly different from Geroch's.

Theorem 1: *Let $\mathcal{M}$ be a (time-oriented) causally compact spacetime that interpolates between an initial surface $\mathcal{S}_1$ and a final surface $\mathcal{S}_2$. Suppose that $\mathcal{M}$ has no closed timelike curves. Then, $\mathcal{M}$ is diffeomorphic to $\mathcal{S}_1 \times [0, 1]$ (and, in particular, $\mathcal{S}_1$ is diffeomorphic to $\mathcal{S}_2$).*

(A variant of this result, but now not assuming time-orientability, is discussed in section VII.)

Proof: Because $\mathcal{M}$ is time-oriented, $I(p) = I^+(p) \cup I^-(p)$ for any point $p \in \mathcal{M}$. So, by causal compactness, each of $\overline{I^+(p)}$ and $\overline{I^-(p)}$ must separately be compact. Let V^a be a smooth, future-directed timelike vector field on $\mathcal{M}$ (chosen to be tangent to the timelike component of $\partial\mathcal{M}$, $\mathcal{T}$, if such a timelike component exists).

If $p \in \mathcal{M}$, let V^a_p denote the element of this vector field at p. The field V^a will give rise to a set of unique, smooth integral curves (i.e., curves to which V^a is tangent – more precisely, solutions of $(dx^a/dv) = V^a$) on $\mathcal{M}$ [31]. The main thrust of the proof is to show that in the absence of closed timelike curves, each of the integral curves of V^a must have a past endpoint on $\mathcal{S}_1$ and a future endpoint on $\mathcal{S}_2$.

The curves that lie in $\mathcal{T}$ take on such endpoints by definition. Suppose that one of the other curves λ (i.e., $\lambda \cap \mathcal{T} = \emptyset$) does not have (say) a future endpoint on $\mathcal{S}_2$. A contradiction ensues, as shown below:

First, observe that λ cannot have a future endpoint anywhere else on $\mathcal{M}$ (since V^a is defined everywhere). Let $b \in \lambda$ and let μ be the portion of λ to the future of b. Then $I^+(\mu) \subset I^+(b)$. Now $\overline{I^+(b)}$ is compact by assumption; so, since $\overline{\mu} \subset \overline{I^+(\mu)} \subset \overline{I^+(b)}$, $\overline{\mu}$ must also be compact. Consider the set $\mathcal{N} = \bigcup_{p\in\mu} I^-(p)$. Clearly $\mu \subset \mathcal{N}$. I now show that, further, $\overline{\mu} \subset \mathcal{N}$.

Suppose that there is a sequence of points q_i on μ that converges to a point q not on μ. Then the vectors $V^a_{q_i}$ will converge to the vector V^a_q. Let ρ be the integral curve with initial tangent V^a_q at q and let q' lie to the future of q on ρ. Since $I^+(q)$ is open, all points sufficiently close to q' will also lie in $I^+(q)$. Some of these points will lie on integral curves of V^a through the q_i close to q, i.e., they will lie on μ. Thus, there will be some $p \in \mu$ such that $q \in I^-(p)$, i.e., $q \in \mathcal{N}$. Therefore, $\mathcal{N}$ provides an open covering of the compact set $\overline{\mu}$. Let $\mathcal{N}' = \bigcup_{i=1}^n I^-(p_i)$ be a finite subcovering and suppose that p_k is the futuremost of the points p_i on μ. But, $p_k \in \mu \subset \mathcal{N}'$, i.e., $p_k \in I^-(p_r)$ for some r. But we also have $p_r \in I^-(p_k)$. This gives a closed timelike curve.

So, every integral curve of the field V^a takes on an endpoint on $\mathcal{S}_1$ and on $\mathcal{S}_2$. We may choose a parameter on each curve that has value 0 at $\mathcal{S}_1$ and 1 at $\mathcal{S}_2$. Since V^a is smooth, this means that $\mathcal{M}$ is diffeomorphic to $\mathcal{S}_1 \times [0, 1]$. □

This theorem also makes it possible to make precise the result that was mentioned in the introduction: a spacetime cannot give topology change if it is without holes (i.e., it is causally compact) and can be foliated by spacelike hypersurfaces.

Corollary: *Let $\mathcal{M}$ be a (time-oriented) causally compact spacetime that interpolates between an initial surface $\mathcal{S}_1$ and a final surface $\mathcal{S}_2$. Suppose that $\mathcal{M}$ can be foliated by spacelike surfaces, with $\mathcal{S}_1$ as the first surface in the foliation and $\mathcal{S}_2$ as the last. Then, $\mathcal{M}$ is diffeomorphic to $\mathcal{S}_1 \times [0, 1]$ (and, in particular, $\mathcal{S}_1$ is diffeomorphic to $\mathcal{S}_2$).*

Proof: Since $\mathcal{M}$ can be foliated in the manner described above, it may be

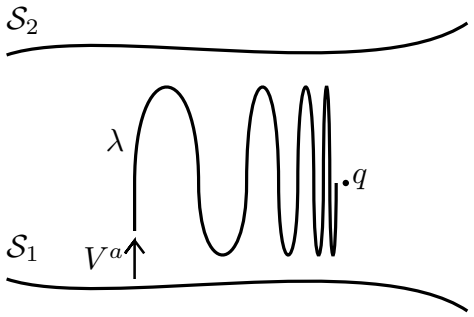

Figure 7: A sketch of how the proof of theorem 1 runs. (Though the surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$ are spacelike, time does not point upward everywhere in the region between the surfaces.) If $\mathcal{S}_1$ and $\mathcal{S}_2$ are not diffeomorphic, at least one integral curve of V^a, λ, must wind around trapped in the interpolating spacetime. By causal compactness, the curve accumulates at some point q, and this accumulation leads to a closed timelike curve.

expressed as

$$\bigcup_{t \in [1,\,2]} \mathcal{S}_t,$$

where each $\mathcal{S}_t$ is a spacelike hypersurface and $\mathcal{S}_{t_1} \cap \mathcal{S}_{t_2} = \emptyset$ for $t_1 \neq t_2$. Then $\mathcal{M}$ can contain no closed timelike curves (since t must strictly increase along every future-directed timelike curve), and the result follows from Theorem 1. □

Since, by assumption, $\mathcal{M}$ is connected, it follows from the conditions of Theorem 1 that $\mathcal{S}_1$ and $\mathcal{S}_2$ must also be connected. A similar restriction also follows from slightly weaker assumptions (the proof is fashioned after one originally given by Geroch [6]):

Theorem 2: *Let $\mathcal{M}$ be a manifold that interpolates between a surface $\mathcal{S}_1$ and a surface $\mathcal{S}_2$. Suppose that $\mathcal{M}$ admits a smooth vector field V^a, every integral curve of which has an endpoint on $\mathcal{S}_1$. Then $\mathcal{S}_1$ is connected. (And similarly for $\mathcal{S}_2$.)*

Proof: Suppose that $\mathcal{S}_1$ is not connected and suppose, further, that it has two disjoint components $\mathcal{A}_1$ and $\mathcal{A}_2$ (which may not themselves be connected). The manifold $\mathcal{M}$ can then be decomposed into two disjoint sets $\mathcal{N}_1$ and $\mathcal{N}_2$, where

$$\mathcal{N}_i = \{p \mid \text{the integral curve of } V^a \text{ through } p \text{ intersects } \mathcal{A}_i\}.$$

Both these sets will be open. This is not possible, since $\mathcal{M}$ is connected. □

V. The Dynamics of Topology Change

Theorem 1 does not completely rule out classical topology change. For, there are solutions of Einstein's equation (with reasonable sources) that have causality violations – the Gödel, Kerr and Taub-NUT solutions, among others (ref. [25], p. 161–178). It is true that such scenarios are difficult to interpret, or to reconcile with our experience. But, allowing topology change to occur opens up the possibility of such rich spacetime structures, that it is tempting to try anyway. (It might be possible, for example, to construct an interpretational framework in which only the initial and final states are viewed as physically "real," with the interpolating spacetime regarded as a device for carrying out calculations – as in the usual interpretation of the path integral approach to quantum theory.) The obvious question is, does Einstein's equation constrain topology change in any way, when the source is restricted in some reasonable manner? (If the source is not restricted, Einstein's equation provides no constraint at all, for then any metric is a solution.) The result that I discuss below suggests that Einstein's equation does place significant constraints: with certain restrictions on the curvature (which mainly follow from reasonable restrictions on the source), it appears that topology change is not allowed for causally compact spacetimes. This result, but with slightly stronger assumptions, was obtained previously by Tipler [7, 8] for the closed universe and externally Lorentzian cases.

Theorem 3: *Let $\mathcal{M}$ be a (time-oriented) causally compact interpolating spacetime of dimension ≥ 3 with initial surface $\mathcal{S}_1$ and final surface $\mathcal{S}_2$. Suppose that*

i) every full (i.e., no endpoint on $\mathcal{S}_1$ or on $\mathcal{S}_2$) null geodesic has a point on it at which $F_{abcd} \equiv U_{[a}R_{b]ef[c}U_{d]}U^eU^f \neq 0$, where U^a is the tangent to the geodesic; and

ii) for any point u_0 on a past-complete, affinely parametrized null geodesic $\lambda(u)$, the half-integral null convergence condition (explained below) holds along λ to the past of u_0.

Then $\mathcal{M}$ is diffeomorphic to $\mathcal{S}_1 \times [0, 1]$, and $\mathcal{S}_1$ and $\mathcal{S}_2$ are connected.

It is important to note that this result is not a "singularity theorem," at least of the standard type, i.e., *it makes no assumptions about geodesic completeness.* As it turns out, the geodesics that we are interested in here are guaranteed to be complete (in the direction of interest) under the conditions of the theorem (this is a standard result – see lemma A below). Thus the theorem does not allow topology change to be had, even at the price of a (standard) singularity. The question of possible singularities associated with topology change – when one of the conditions imposed above is relaxed – is discussed again in section 1X.

Two restrictions are made on the curvature in the theorem. I discuss each in turn:

A] Assumption i

This assumption is essentially a generality condition; i.e., it will fail to hold only in situations that are, in a precise sense, highly special. There are various ways in which such generality assumptions may be made. The condition that I am using is the null form of the "generic condition" that was used by Hawking and Penrose in their 1970 singularity theorem [32, 25]. Their original argument in support of the reasonableness of the condition was made in four dimensions, but similar arguments may be made in higher dimensions as well [33].

First, consider non-vacuum spacetimes (i.e., $T_{ab} \neq 0$). For null vectors U^a, Einstein's equation implies that $R_{ab}U^aU^b = kT_{ab}U^aU^b$, where k is a constant (the relationship holds even if there is a cosmological constant). For matter energy-momentum tensors of known types, $T_{ab}U^aU^b$ will vanish only if T_{ab} represents pure radiation travelling in the U^a direction (ref. [25], p. 101). In any realistic model with sources it is reasonable to suppose that there are other types of matter in addition to pure radiation, or that a null geodesic does not align itself with the flow of radiation throughout, i.e., it is reasonable to require that $R_{ab}U^aU^b \neq 0$ somewhere on each null geodesic. At that point we must also have $F_{abcd} \neq 0$.

In the vacuum case (i.e., "pure gravity") in spacetimes of dimension greater than three, the null generic condition is equivalent to requiring that U^a not point in a principal null direction of the Weyl tensor throughout the length of the null geodesic. Since at any point there are only a finite number of such directions [33], it is reasonable to require that such an alignment not occur.

In three dimensions there is no vacuum case, for $C_{abcd} \equiv 0$ and the Riemann tensor is determined by the Ricci tensor. But there is further reason here to believe that $R_{ab}U^aU^b$ will not generically vanish along the entire length of a geodesic. Consider the expression $R_{ab}U^aU^b$, for all null vectors U^a at a point. Pick a basis $\{P^a,\ M^a,\ S^a\}$ at that point such that P^a and M^a are both null, S^a is a unit spacelike vector orthogonal to P^a and M^a, and $P^aM_a = -1$. Then an arbitrary null vector is either proportional to M^a or is proportional to a vector of the type $P^a + (\beta^2/2)M^a + \beta S^a$. Thus, $R_{ab}U^aU^b = 0$ will yield a quartic equation for β, unless $R_{ab}M^aM^b$ is already zero whereupon the equation is cubic; i.e., there are at most 4 null directions, U^a, at a point that can satisfy $R_{ab}U^aU^b = 0$. And, as above, it seems reasonable to suppose that in a generic spacetime a null geodesic will not align itself throughout along one of these directions; i.e., that $R_{ab}U^aU^b$, and therefore F_{abcd}, is non-zero somewhere.

A stronger form of the generic condition that we might consider using requires that $F_{abcd} \neq 0$ somewhere on one side of a given point on every geodesic. This form is useful in situations where we are interested in what happens to the future (or has happened to the past) of some initial (or final) state. In gravitational collapse, for example, we are interested in what happens to the future of the initial state. And, in discussions of the initial singularity, we are interested in the past. In such situations it is often necessary to impose conditions on the half-geodesics that have initial (final) endpoints at the initial (final) state. If this is not done, we often lose important information. For example, the Hawking-Penrose singularity theorem does not make assumptions about half-geodesics and, as a consequence, it does not yield information about the location (in time) of the singularity. In fact, it leaves open the possibility that the singularity of gravitational collapse might have occurred to the past of the (non-singular) initial state, or that the 'initial' cosmological singularity might occur in the future. On the other hand, with a condition on half-geodesics it is possible to ensure, for instance, that the initial singularity does indeed lie to the past [34].

The problem with assumptions on half geodesics, however, is that they are considerably stronger than than conditions on full geodesics. Consider a situation where a black hole forms out of some realistic (and therefore asymmetrical) initial state. It will settle down to a highly symmetrical final state [25]. In such a case it is much weaker to require that $F_{abcd} \neq 0$ somewhere on a geodesic than to require, say, that this happens at some point after the black hole forms.

Topology change is a problem where we are chiefly interested in that portion of a full spacetime that lies between the initial and final surfaces, the interpolating spacetime $\mathcal{M}$. It might seem that to extract information we must require that every geodesic in $\mathcal{M}$ have a point at which $F_{abcd} \neq 0$, even if the geodesic has endpoints on $\partial\mathcal{M}$. Such a condition was imposed in Tipler's theorem [7, 8]. But, as we shall see, the weaker condition that $F_{abcd} \neq 0$ somewhere just on full geodesics is enough.

B] Assumption ii

Let $\lambda(u)$ be an affinely parametrized null geodesic, where u increases in the future direction. The *half-integral null convergence condition* is said to hold along λ to the past of some point u_0 (more precisely, to the past of $\lambda(u_0)$) if for any $\delta > 0$, $\exists b > 0$ such that for any $u_1 < u_0$ there is an interval I of length $\geq b$, with $u_1 > \sup I$ and with

$$\int_u^{u_0} R_{ab} U^a U^b \, du \geq -\delta, \qquad \forall u \in I.$$

This condition limits how negative the "matter term" $R_{ab}U^aU^b$ can get [35]. (The quantity $R_{ab}U^aU^b$ is called the matter term because it may be related, via Einstein's equation, to $T_{ab}U^aU^b$.) The precise statement of the condition is, unfortunately, a little involved, but the idea that it tries to express is simple. Suppose that $\lambda(u)$ is a null geodesic with affine parameter u and tangent U^a. Let $p = \lambda(u_0)$ be some point on the geodesic. Now, suppose that in some places along λ before p the term $R_{ab}U^aU^b$ is negative. The condition then requires that there always be other regions where this term is positive enough to make $\int_u^{u_0} R_{ab}U^aU^b du$ come out as close to zero as we want, for all u in an interval of some finite length b. (Roughly speaking, it asks that $\int R_{ab}U^aU^b du$ at least come repeatedly close to zero as one looks along λ in the direction of decreasing u.) This condition is a more stringent restriction on the curvature than the one in assumption (i). It will be discussed in greater detail in the next section.

C] The proof

The proof of Theorem 3 depends on observing that there must be a non-empty future Cauchy horizon $H^+(\mathcal{S}_1)$, if $\mathcal{S}_1$ and $\mathcal{S}_2$ are not diffeomorphic. It is a standard result in global general relativity that such horizons are "generated by null geodesics" (in a sense that is defined below). The main thrust of the proof involves first showing that $H^+(\mathcal{S}_1)$ contains a full geodesic, and then observing that focusing effects that occur on this null geodesic, as a consequence of assumptions (i) and (ii), are incompatible with certain consequences of the assumption of causal compactness.

The detailed proof relies on some standard results from global general relativity. Let $\mathcal{H} = H^+(\mathcal{S})$ be a future Cauchy horizon, where $\mathcal{S}$ is a spacelike surface, and let $\mathcal{A} = \dot{I}^-(p)$ be the boundary of the past of p, where p is any point. Then the following hold (ref. [25], chapter 6):

- *(a)* $\mathcal{H}$ and $\mathcal{A}$ are achronal, i.e., no two points on either of these sets can be connected by a timelike curve.

- *(b)* Through each point of $\mathcal{H}$ there passes a past-directed null geodesic that either does not leave $\mathcal{H}$ when followed into the past or leaves it at the edge of the surface $\mathcal{S}$. This geodesic may leave $\mathcal{H}$ at any point when followed into the future. (See fig. 2c and fig. 2d for an illustration of this behavior.)

- *(c)* Through each point of $\mathcal{A}$ there passes a future-directed null geodesic that can leave $\mathcal{A}$, when followed into the future, only through p, but may leave $\mathcal{A}$ anywhere when followed into the past.

• *(d)* let μ and ρ be future-directed null geodesics that intersect at some point q; if q' lies to the future of q on μ, or if $q' \in I^+(q)$, then the portion of ρ to the past of q lies in $I^-(q')$. Similarly, if the future endpoint of a timelike or null curve ρ coincides with the past endpoint of another timelike or null curve μ, with a discontinuity in their tangents at the intersection point (i.e., they "meet at an angle"), then $\rho \subset I^-(\mu)$.

The null geodesics in (b) and (c) are called the *null generators* of $\mathcal{H}$ and $\mathcal{A}$. Before embarking on the main proof, it is useful to state two standard results that we will need:

Lemma A: *Let $\mathcal{H}$ be a future Cauchy horizon and λ a generator of this horizon with no past endpoint on $\mathcal{H}$. If λ is confined within a compact set when followed into the past, then λ is past complete* (ref. [25], p. 295–297, and ref. [7], p. 98–103).

It is this result that makes it unnecessary for us to assume geodesic completeness in Theorem 3.

The second result describes the focusing of null geodesics in an n-dimensional spacetime ($n > 2$). The standard four-dimensional procedure is simply imitated: Consider a congruence of null geodesics; let u be an affine parameter along the null geodesics, chosen to increase in the past direction, and let $U^a = (\partial/\partial u)^a$. Let D_a be the covariant derivative, and let $D = U^a D_a$. At each point set up a basis of vectors consisting of U^a, a null vector M^a obeying $M^a U_a = -1$ and $DM^a = 0$, and $(n-2)$ spacelike vectors S_i^a orthogonal to each other and to U^a and M^a and also obeying $DS_i^a = 0$. Then $h_{ab} = g_{ab} + 2U_{(a}M_{b)}$ will be a positive-definite metric on the space orthogonal to U^a and M^a. Define the expansion of the congruence by $\theta = D_a U^a$ $(= h^{ab} D_a U_b)$ and the shear by $\sigma_{ab} = D_{(a}U_{b)} - \frac{1}{n-2} h_{ab}\theta$. Let $\sigma_{ij} = S_i^a S_j^b \sigma_{ab}$. Then, $2\sigma^2 \equiv \sigma_{ab}\sigma^{ab} = \sigma_{ij}\sigma^{ij} \geq 0$, with equality holding iff $\sigma_{ij} = 0$. When the spacetime dimension n is greater than 3, the behavior of θ and σ_{ij} along the null geodesics of the congruence is given by:

$$\frac{d\theta}{du} = -\frac{1}{n-2}\theta^2 - R_{ab}U^a U^b - 2\sigma^2 \tag{1}$$

$$\frac{d\sigma_{ij}}{du} = -C_{iajb}U^a U^b - \frac{2}{n-2}\theta\sigma_{ij} - \sigma_{ik}\sigma_{jl}h^{kl} + \frac{2\sigma^2}{n-2}h_{ij} \tag{2}$$

where C_{abcd} is the Weyl tensor. When $n = 3$, $\sigma_{ij} \equiv 0$ and we have

$$\frac{d\theta}{du} = -\theta^2 - R_{ab}U^a U^b. \tag{1'}$$

In both cases we are interested in focusing, i.e., in the conditions under which θ diverges.

Lemma B: *Suppose that the expansion θ is positive at some point $\lambda(u_0)$ of a member λ of a congruence of past-directed null geodesics (with affine parameter u chosen to decrease in the past direction). Further, suppose that λ is past-complete and that the half-integral null convergence condition holds to the past of u_0. Then, $\theta \to \infty$ within a finite affine parameter distance to the past of u_0* [35].

Here, now, is the proof of Theorem 3:

Proof: The main argument is best given in steps:

1. There is a non-empty future Cauchy horizon $\mathcal{H}$.

Set up a timelike vector field V^a on $\mathcal{M}$, and suppose that one of its integral curves does not intersect either $\mathcal{S}_1$ or $\mathcal{S}_2$. Then, as we have seen in Theorem 1, there will be a closed timelike curve in $\mathcal{M}$. This curve cannot intersect $\mathcal{S}_1$, and so points on it cannot lie in $D^+(\mathcal{S}_1)$. Therefore $\mathcal{H} = H^+(\mathcal{S}_1) \neq \emptyset$. (It is assumed here that $\mathcal{S}_1$ is connected. If it is not, an argument identical to the one given below may be applied to a connected component of $\mathcal{S}_1$, with an identical ensuing contradiction.)

Since I am allowing interpolating spacetimes that are confined within a timelike tube $\mathcal{T}$, there is a slight subtlety here. The interpolating spacetime $\mathcal{M}$ – bounded by $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{T}$ – is the region of interest. Thus a point p will lie in $D^+(\mathcal{S}_1)$ as long as every past-directed timelike or null curve from it *that is confined to* $\mathcal{M}$ eventually intersects $\mathcal{S}_1$. Now, let γ be a past-directed timelike or null curve and suppose that γ intersects $\mathcal{T}$. As long as γ can be continued – either along $\mathcal{T}$ or within the interior of $\mathcal{M}$ – to $\mathcal{S}_1$, it is considered a curve that meets $\mathcal{S}_1$.

2. The generators of $\mathcal{H}$ are past complete.

It follows from the comment in the preceding paragraph that no null generator of $\mathcal{H}$ can intersect $\mathcal{T}$ when followed into the past. For, suppose there is a generator ν of $\mathcal{H}$ that intersects $\mathcal{T}$ at a past endpoint x. Let μ be any timelike curve on $\mathcal{T}$ with x as its past endpoint, and let x_i be a sequence of points on μ converging to x. Each point x_i will lie to the future of some point on ν, and thus none of the x_i will lie in $D^+(\mathcal{S}_1)$. So from each x_i there will be a past-directed timelike curve α_i that never intersects $\mathcal{S}_1$ when followed into the past within $\mathcal{M}$. The curves α_i are thus trapped in $\mathcal{M}$. The sequence $\{\alpha_i\}$ of these curves will have a past-directed limit curve, α, through x (ref. [25], p. 185). The curve α must also lie in $\mathcal{M}$. Points on α close to x must lie to the chronological past of points on ν (since α will meet ν "at an angle" at x) and thus they will lie in the interior of $D^+(\mathcal{S}_1)$. But this will make points on some of the curves α_i also lie in $D^+(\mathcal{S}_1)$, contradicting the fact that the α_i have been chosen not to intersect $\mathcal{S}_1$.

Now, let p be any point in the interior of $\mathcal{M}$ that lies on a null generator of $\mathcal{H}$ and is not its future endpoint, and let λ be the portion of this null geodesic that

lies to the past of p. Let p' be some point in $I^+(p)$. Then $\lambda \subset \overline{I^-(p')}$, i.e., it is a subset of a compact set (by causal compactness of $\mathcal{M}$). It follows from lemma A that λ is necessarily past complete. Clearly, this applies to any null generator of $\mathcal{H}$.

3. $\mathcal{H}$ contains a full geodesic.

Let p, p' and λ be, respectively, the points and the curve from step 2, and let $\mathcal{B} = \mathcal{H} \cap \overline{I^-(p')}$. When λ is extended into the future it may or may not leave $\mathcal{B}$. Suppose it does leave $\mathcal{B}$ at some future endpoint. Then there will be some other null generator of $\mathcal{H}$ that does not leave $\mathcal{B}$ to the future. To see this, let $\{q_i\}$ be a sequence of points on λ such that each q_{i+1} lies to the past of q_i, and with no finite segment of λ containing an infinite number of these points. These points will have a limit point q. Now, $\lambda \subset \mathcal{B}$. Also $\lambda \subset \dot{I}^-(p)$. Both these sets are closed and so the point q belongs to both. The null generator of $\mathcal{B}$ through q when followed into the future will also lie on $\dot{I}^-(p)$. Call this generator μ. I first show that μ cannot represent the same geodesic as λ itself.

Suppose, to the contrary, that it does. Choose a point x on the future extension of λ, after it leaves the set $\mathcal{B}$. Let $\mathcal{O}$ be a neighborhood of the segment of λ between x and q such that there is some point $\hat{q}$ on λ to the past of q that does not lie in $\mathcal{O}$. From now on let the sequence $\{q_i\}$ be restricted to those points that lie to the past of $\hat{q}$. Let N_y represent the tangent to λ at any point $y \in \lambda$. Since $q_i \to q$ we must have $N_{q_i} \to N_q$ (otherwise there will be some other generator of $\mathcal{B}$ that intersects λ at q, and that is not possible). This means that from all the q_i sufficiently close to q the null geodesics with initial tangents N_{q_i} must reach an arbitrarily small neighborhood of x without ever leaving $\mathcal{O}$. This is not possible since (i) all of λ to the past of q lies in $\mathcal{B}$, (ii) $x \notin \mathcal{B}$, and (iii) the segment of λ between each q_i and q contains the point $\hat{q}$ that does not lie in $\mathcal{O}$.

So μ cannot represent the same geodesic as λ. It (i.e., μ) can leave $\dot{I}^-(p)$ to the future only through p. Suppose it does. But then for any point b in $\mathcal{H}$ to the future of p along λ, we have $\mu \subset I^-(b)$, which is not possible since $q \in \mu$, and q and b, being points on $\mathcal{H}$, cannot be connected by a timelike curve. So, μ does not leave $\dot{I}^-(p)$ to the future, and therefore it does not leave $\overline{I^-(p')}$ to the future either. Also, it does not leave $\mathcal{H}$ to the future. For, suppose it does. Let c be a point on μ such that $c \in \dot{I}^-(p)$ but $c \notin \mathcal{H}$. Points close to c will not lie in $D^+(\mathcal{S}_1)$ but some will lie in $I^-(p)$. Let ρ be a future-directed timelike curve from one of these points to p. Points on ρ close to p will lie in $D^+(\mathcal{S}_1)$, i.e., ρ must enter $D^+(\mathcal{S}_1)$ at some point on $\mathcal{H}$ before it reaches p. This violates the achronal nature of $\mathcal{H}$. So, μ does not leave $\mathcal{H}$ to the future. Thus, it does not leave $\mathcal{B}$ to the future or to the past (since it cannot leave either $\mathcal{H}$ or $I^-(p')$ to the past). (A similar result was obtained previously in a different way [36].)

4. Focusing occurs on $\mathcal{H}$, leading to a contradiction.

Next, we consider the focusing of null geodesics in $\mathcal{B}$. To do this, we need to define a congruence of null geodesics in a neighborhood of $\mathcal{B}$ (so that quantities like derivatives may be defined). This can be done by varying points on $\mathcal{B}$ in a direction not contained in $\mathcal{B}$ [35]. All the quantities that we are interested in (θ, etc.) turn out to be independent of the particular variation that is used. Choose an affine parameter on the geodesics of this congruence that increases in the future direction.

Since $\mathcal{B}$ is compact, it may be shown that we cannot have $\theta \leq 0$ throughout and $\theta < 0$ somewhere (ref. [25], p. 297–298). Therefore, we either have $\theta = 0$ throughout, or $\theta > 0$ somewhere on $\mathcal{B}$. If the spacetime dimension n is 3, assumption (i) of the theorem means that $R_{ab}U^aU^b$ – and hence from equation (1′), θ – cannot be zero everywhere along μ. If $n > 3$ and if θ is zero throughout on μ, then equation (1) will yield $R_{ab}U^aU^b = -2\sigma^2 \leq 0$. From assumption (ii) of the theorem, it follows that $R_{ab}U^aU^b = 0$ throughout; therefore $\sigma_{ij} = 0$ throughout as well. But, from assumption (i) and equation (2) applied to μ, it follows that $\sigma_{ij} \neq 0$ somewhere on μ. So, we must have $\theta > 0$ somewhere on μ. From lemma B, $\theta \to \infty$ within a finite affine parameter distance. This will violate the achronality of the horizon (ref. [25], p. 115–116). (Just as the crossing of two separate generators of $\mathcal{H}$ will violate the achronality of $\mathcal{H}$, by the time reverse of (d) above, so also will the crossing of 'infinitesimally close' generators, the diverging of θ being taken to indicate such a crossing.)

Therefore, there can be no Cauchy horizon under the assumptions of the theorem, i.e., every integral curve of V^a must take on endpoints on $\mathcal{S}_1$ and on $\mathcal{S}_2$. Hence the first part of the result.

The second part follows immediately from these facts: (a) $\mathcal{M}$ is connected, and (b) $\mathcal{M} = [0, 1] \times \mathcal{S}_1$ (as shown above). □

VI. Energy conditions

Assume that Einstein's equation holds (possibly with cosmological constant) on $\mathcal{M}$. What sorts of conditions on the matter energy-momentum tensor, T_{ab}, will yield the half-integral null convergence condition of the previous theorem?

Obviously, any condition on T_{ab} that implies that $R_{ab}U^aU^b \geq 0$ (this is called the *null convergence condition*) will be sufficient. An important case is the vacuum: $T_{ab} = 0$. This covers those situations where we are interested in the behaviour of "pure gravity." Such situations include models where "matter" is built out of gravity either by using non-trivial topologies [2, 4, 10], or higher dimensions, or both [13, 37].

Another case of interest is when $T_{ab} \neq 0$ but is bounded below in some suitable sense. For example, suppose at some $p \in \mathcal{M}$ that $T_{ab}V^aV^b \geq K$ for all unit timelike V^a at p, where K is some number (possibly negative). Physically this means that the energy density as seen by any observer passing through p will be $\geq K$. It follows that $R_{ab}U^aU^b \geq 0$ for all null U^a at p. For, let U^a be any null vector at p and let V_i^a be a sequence of unit timelike vectors at p that approaches the direction of U^a as a limit direction. Let $T_i^a = b_i V_i^a$ (no sum on i) such that the vectors T_i^a converge to U^a. Then we must have $b_i \to 0$. Therefore, $T_{ab}T_i^aT_i^b = T_{ab}V_i^aV_i^b(b_i)^2 \geq K(b_i)^2 \to 0$. Thus $T_{ab}U^aU^b \geq 0$ and hence, by Einstein's equation, $R_{ab}U^aU^b \geq 0$. (This result was proved previously by Tipler for a special class of energy-momentum tensors [7, 38].)

Though in this result K can be negative, and can vary from point to point on $\mathcal{M}$, the physically interesting case appears to be $K = 0$. In this case we say that T_{ab} obeys the *weak energy condition*, i.e., all observers measure a non-negative energy density. This condition is obeyed by all known forms of classical matter (ref. [25], p. 89–91).

The energy-momentum tensor associated with quantum fields can, however, have expectation values that violate the weak energy condition [38, 39, 17]. Can such fields be used to drive topology change so that the process, though forbidden classically, can occur semi-classically? When dealing with such situations, the freedom that we have in allowing K to be negative is not likely to be of much use. Suppose that there is a lump of good (i.e., positive energy) matter somewhere: then all observers will measure the energy to be positive. The value of this measured energy will not be bounded above, for by moving past the lump at speeds approaching the speed of light the energy can be made to appear as large as we wish. Similarly for bad matter then, if the energy as measured by some observer at a point p is negative, it seems unreasonable to expect in general that there be a lower bound for the value of the measured energy for all observers passing through p. (Even though, formally, it is possible to construct examples where this happens: a simple one is $T_{ab} = Cg_{ab}$.) This is born out in calculations of the energy-momentum tensor that have been done for quantum fields. For example, for the Casimir effect or for Hawking radiation the violations of the weak energy condition appear not to be bounded below [39].

In such cases, $R_{ab}U^aU^b$ can be negative. If the regions where this happens do not dominate the spacetime then we might expect that on the average geodesics see non-negative $R_{ab}U^aU^b$, i.e., that integral conditions such as the half-integral convergence condition might perhaps hold. This is so far only a piece of wishful thinking. Integral convergence conditions were introduced by Tipler [38] to replace the standard pointwise convergence conditions in studies of geodesic focusing. Though

much further work has been done on such conditions [35, 39, 40, 41, 42, 43, 44] their ultimate usefulness remains uncertain. The question needs further study.

An indication that we might be able to make some reasonably reliable statement about the extent of the violations of the weak energy condition when quantum fields are present comes from a study of black hole evaporation [45]. The qualitative picture of some aspects of this process that we are led to believe (from using energy conservation to relate the energy flux at large radial distances to the energy flux near the black hole) [46] is this: Consider a Schwarzschild black hole of initial mass m. If it were not evaporating, the null geodesic generators of its event horizon would sit at $r = 2m$ throughout (assuming that no new matter falls in). During evaporation, however, these null geodesics that were initially at $r = 2m$ will start diverging and will escape to infinity (because, crudely, the black hole mass decreases and the forces holding them in place get weaker). Slightly inside $r = 2m$ (at $r \sim 2m(1 - (m_p/m)^2)$, where m_p is the Planck mass), there will be a null hypersurface whose generators were initially converging slightly. This convergence will slowly go to zero and the surface will be, by definition, the event horizon. Inside it there will be null geodesics that were also initially converging and which continue to do so with the convergence not approaching zero. On all these geodesics the violations of the weak energy condition cannot be uncontrollably large – for, in that case, one would expect the initially converging geodesics slightly inside $r = 2m$ to also all quickly diverge to infinity. All the preliminary calculations suggest that this does not happen – offering the hope that some sensible estimate can be made, both of the extent of the violations of the weak energy condition in this quantum process (and in others), as well as of the effect that such violations might have. It might be possible, for example, that a singularity theorem might be provable for evaporating black holes [47, 39, 48] using the idea that violations of the weak energy condition are limited. And, it might be possible to make some statement about topology change, even in the presence of quantum fields. The statement of assumption (ii) as an integral convergence condition is an attempt to provide a framework for doing this – but, clearly, a lot more remains to be done.

VII. Time-orientability

Would the conclusions of this paper be significantly different were we to drop our insistence on time-orientable metrics? If the metric on $\mathcal{M}$ is not required to be time-orientable, then some changes do show up at the kinematical level. For example, Sorkin has shown that the Reinhart-Sorkin selection rule, $\Delta\chi \equiv \chi(\mathcal{S}_1) - \chi(\mathcal{S}_2) = 0$, is no longer a condition for the existence of a Lorentzian

cobordism in odd dimensions [49]. The results of this paper are, however, still essentially true. To see this, note first that the existence of a not-necessarily-time-orientable metric is equivalent to the existence of a (timelike) direction field (i.e., at each point a vector is defined up to sign) [25]. Integral curves to this field may still be constructed (if necessary, first in the time-orientable double covering manifold (ref. [25], p. 181) and then mapped back to the original manifold). If the interpolating spacetime is assumed to be causally compact, it follows that there will be a closed timelike curve unless each of these integral curves takes on two endpoints on $\partial\mathcal{M}$. (The existence of the closed timelike curve may first be proved, if necessary, in the time-orientable double covering manifold; this curve may then be mapped into a closed timelike curve in $\mathcal{M}$.) Therefore, if causality violations are to be excluded, $\mathcal{M}$ may be divided into three disjoint sets, $\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{N}$, where

$$\mathcal{N} = \{p \mid \textit{the integral curve through } p \textit{ takes on one endpoint each on } \mathcal{S}_1 \textit{ and } \mathcal{S}_2\},$$

and

$$\mathcal{N}_i = \{p \mid \textit{the integral curve through } p \textit{ takes on both endpoints on } \mathcal{S}_i\}.$$

These are all open sets, and so, since $\mathcal{M}$ is connected, two of them must be empty. Therefore, $\mathcal{M}$ has either a single boundary $\mathcal{S}_1$ (or $\mathcal{S}_2$), which is mapped onto itself by the integral curves of the direction field, or it has two separate, but diffeomorphic, boundaries. In neither case can topology change be said to occur. (A proof along these lines was given by Geroch for the case of closed and externally Lorentzian spacetimes [6]. Geroch's result is sometimes quoted as having assumed that $\mathcal{S}_1$ and $\mathcal{S}_2$ are connected. In fact, the proof works even if they are not, as long as $\mathcal{M}$ is connected.) An example of the first case is the Möbius strip with time along the non-circular direction and the space axis along the circular direction. Here, the single S^1 boundary is mapped to itself. See fig. 8. In the second case, when $\mathcal{S}_1$ and $\mathcal{S}_2$ are diffeomorphic (and non-empty), a time-orientation may in fact be introduced on $\mathcal{M}$ as follows: choose a parameter t along the integral curves that increases from $\mathcal{S}_1$ to $\mathcal{S}_2$, then pick $(\partial/\partial t)^a$ to be the future direction.

Thus, we have the following result: a causally compact spacetime that interpolates between non-empty boundaries $\mathcal{S}_1$ and $\mathcal{S}_2$ and which does not have closed timelike curves must (a) be diffeomorphic to $\mathcal{S}_1 \times [0, 1]$, and (b) hence be time-orientable.

Therefore, topology change will have causality violations associated with it, even in the non-time-orientable case. Further, theorem 3 may be applied in the time-orientable double covering manifold (all the other assumptions of the theorem will continue to hold) to get a contradiction.

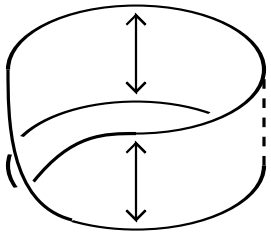

Figure 8: A non-time-orientable "transformation" of a circle onto itself in a Möbius strip spacetime. Compare this case with the Möbius strip example of fig. 6c; that example may be viewed as a $\emptyset \leftrightarrow S^1$ transition whereas the non-time-orientability of this one makes it difficult to interpret.

No other orientability requirements were placed on the interpolating spacetime. If the manifold itself is orientable, but the metric on it is not time-orientable, then it cannot be space-orientable either [26]. It might be that violations of P and of CP in particle interactions force us to only consider orientable manifolds as models for space. But, if this is true, and if we wish to consider non-time-orientable metrics (despite the implication of the CPT theorem that T must be violated if CP is), we can do so by looking at non-orientable interpolating manifolds. Thus there appears to be no overwhelming reason against formally considering non-time-orientable metrics – but, as we have seen above, nothing of overwhelming significance seems to emerge either from doing so.

VIII. A Few Words on Differentiability

I have assumed in the discussion so far that the manifolds under discussion are smooth and that the various fields defined on them are smooth as well. In fact, we usually need a much lower degree of differentiability. In order to discuss this, the following standard notation (already used in this paper) is helpful: C^n means n times differentiable, with continuous n-th derivatives; *smooth* means C^∞, i.e., all derivatives exist. Now, integral curves of the vector field V^a will exist if V^a is continuous, and the curves will be unique if V^a is C^1 (this is a sufficient condition; the necessary condition is weaker [31]). And Theorem 3 will go through if the metric is C^2. So, it would have been sufficient to have assumed the manifold to be C^3 and the basic fields on it to be C^2. (The manifold needs one higher order of differentiability than the tensor fields on it because the transformation formula for such fields when co-ordinates x are changed to co-ordinates x' involves $\frac{\partial x'}{\partial x}$.)

The precise degree of differentiability, however, ought not to matter in classical physics. Since the matter and the geometrical fields, as well as the spacetime manifold itself, are probably only approximations to more fundamental underlying structures, we ought to be able to consider smooth enough such approximations. Mathematically this can be made precise, for it may be shown that C^r structures may indeed be approximated by C^∞ ones [50]. For these reasons, it seems justified to work always with smooth quantities (and not have to keep track at each stage of the number of derivatives).

This justification for assuming that the quantities that we are considering are smooth, works if they were differentiable to begin with. What would happen if they were only continuous? If these quantities are fields defined on a given manifold, then they too can be smoothed. But, if we are considering mappings from one manifold to another (as we are, between the manifolds $\mathcal{S}_1$ and $\mathcal{S}_2$, in this paper), then continuous mappings cannot always be smoothed. This occurs in situations involving the so-called 'exotic' differentiable manifolds [28, 51, 52]. There are arguments by Witten [52] that such structures might be interesting to consider. To explain what is involved, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two smooth manifolds of the same dimension and let f be a 1-1 mapping of $\mathcal{M}_1$ onto $\mathcal{M}_2$. f is called a *homeomorphism* if it and f^{-1} are continuous, and it is called a *diffeomorphism* if it and f^{-1} are differentiable. If a diffeomorphism exists and is C^r, then a C^∞ diffeomorphism also exists [50], and so it is not important for many purposes to specify the degree of differentiability of f. It is important to note that when we talk about two manifolds being diffeomorphic or not, we are not making a statement about the manifolds only as point sets or as topological spaces. Our statement refers also to the *differential structure* on the manifolds, i.e., the maximal class of coordinate systems on each manifold that are compatible with each other (in the sense that when two coordinate systems overlap, the transformations from one to the other are differentiable (smooth, if the manifold is to be called smooth)). It turns out to be possible to define on the same topological space two different differential structures, i.e., to have manifolds that are homeomorphic, but not diffeomorphic. A differentiable manifold that is homeomorphic to some standard differentiable manifold like R^n or S^n, but not diffeomorphic to it, is called an *exotic differentiable manifold*. Examples of this exist for S^n, $n \geq 7$ [28, 53] and for R^4 [51].

How is all this related to "topology change"? If we are to consider exotic differentiable manifolds, then a natural question to ask is if they can be created from ordinary ones, i.e., is it possible to find an interpolating spacetime with boundaries that are homeomorphic but not diffeomorphic? Would the theorems of this paper apply to such a situation?

First consider the question at the manifold level. Here, examples of interpolating manifolds exist. For example, there is an 8-dimensional manifold (which is constructed by looking at vectors of length ≤ 1 in a certain type of R^4 bundle over S^4) whose boundary may be shown to be an exotic S^7 [28]. In this manifold remove a ball of radius ϵ around some point p. The resulting manifold will be a cobordism between an ordinary and an exotic S^7. Next, can a Lorentz metric be put on this cobordism? Since it is 8-dimensional, it can be modified so that its Euler characteristic vanishes. On this cobordism the required vector field V^a will exist, and so the cobordism will be Lorentzian. The theorems of this paper should then apply. Thus, though the creation of exotic manifolds in relativity appears to be kinematically possible, this process, too, appears to be dynamically forbidden within a classical framework of the type being considered here.

It is worth observing that this conclusion depends on V^a being differentiable. If we were satisfied with metrics that are only continuous (as we might be, if we were interested purely in causal structure and not in dynamics), then the associated V^a would also only be continuous. In this case, $\mathcal{S}_1$ and $\mathcal{S}_2$ need not be diffeomorphic, even if causality violations are forbidden. Indeed, $\mathcal{S}_1$ and $\mathcal{S}_2$ need not even be homeomorphic here, since the integral curves of V^a need not be unique [31].

IX. Concluding Comments

Since part of the point of this paper is to address (and, with luck, dispel) certain misconceptions about Lorentzian topology change, here is a list of a few of the more common ones, along with some comments:

• *Topology change is intrinsically incompatible with a Lorentzian metric.* Much of section III addresses this and shows that this perception is not true.

• *Two-dimensional topology change is necessarily singular.* This perception appears to be based on the studies that have been made [14, 16] of the $S^1 \cup S^1 \to S^1$ transition (the so-called "trousers topology"). In this case there is a singularity, but the examples of section III show that there are also non-singular topology-changing spacetimes in two dimensions (albeit with closed timelike curves).

• *Closed-universe topology change leads to closed timelike curves only when the metric is time-orientable.* This is addressed in section VII. As shown there, it is possible to use non-time-orientable metrics to avoid closed timelike curves only when one of the boundaries is empty.

• *Closed-universe topology change leads to closed timelike curves only if some suitable energy condition holds.* Neither Geroch's original theorem, nor its mild generalization in section IV, assume anything about the energy-momentum tensor, or indeed about a field equation – the results are purely kinematical.

• *Closed-universe topology change leads either to closed timelike curves or to a singularity.* The truth of this depends on the definition of a singularity. If the standard incomplete-geodesic definition is used, then this statement is not true: as long as the causal compactness condition is met, causality violations have to occur when the topology changes, even if incomplete geodesics are admitted. But, other definitions of a singularity may make the statement true: this is briefly discussed at the very end of this paper.

• *Topology change may be dynamically had in closed universes if the metric is allowed to be singular.* The comments under the previous misconception apply here as well. The situation here is complicated, however, by the existence of a further theorem due to Tipler [8] that was originally presented – and has been quoted – as a singularity theorem. Compactness assumptions are not made in this theorem, and topology change is then shown to lead to singularities, *but only under a significant additional assumption.* The result is discussed further below.

Of course, all of these comments are valid only under the assumptions of this paper. It is best not to view them too dogmatically: it is always possible that different conclusions may be drawn if different assumptions are made.

The same cautionary note applies to the main theorems of this paper which, following on the earlier work of Geroch and Tipler, appear to forbid topology change. Their true value is not so much that they actually rule out topology change, but rather that they allow us to pinpoint what modifications we have to make in our general framework so as to allow it. A popular modification (for a number of reasons, not all directly related to the specific problem of topology change) is to abandon the Lorentzian framework altogether and to use a Euclidean path integral formalism. But even within the general Lorentzian framework there are still several interesting possibilities.

A] Dropping causal compactness

One possibility is to drop the causal compactness assumption. But, it is hard to see that it could be replaced by a weaker assumption that still restricts the candidates for interpolating manifold in some way. And without some restriction, as we have seen, these manifolds can be cut and truncated in an entirely arbitrary manner. Also, there is the theorem of Tipler (ref. [8], Theorem 5) that was mentioned above: this result shows (in the closed universe case) that if topology change

occurs via a non-compact interpolating spacetime, then it contains (under some mild additional assumptions) a singularity (in the sense of an incomplete timelike geodesic) *or a point at infinity*. That it is a singularity that must occur may be inferred only under the significant additional assumption that there is an upper bound on the lengths of certain timelike curves in the region of interest. It has been shown by Yodzis [27] that we have a certain amount of choice in the matter: a conformal transformation may be found in some cases to make the interpolating spacetime future complete (i.e., the singularity may be pushed to infinity). But the presence of points at infinity is still a highly undesirable feature. It seems reasonable, therefore, to retain some type of compactness assumption.

B] Weakening the curvature constraints

We might also consider weakening the constraints on the curvature in Theorem 3. This would not affect the presence of causality violations, but it might permit topology change as a dynamical process. A drastic step in this direction would be to alter Einstein's equation so that assumptions (i) and (ii) can no longer be justified from reasonable restrictions on the source. But such an alteration would have to be fairly severe – for Einstein's equation was used only in a very weak way in going from a condition on T_{ab} to a condition on R_{ab}. There is no real justification – theoretical or experimental – for such a step. Another possibility is (as was discussed earlier) that there might be violations of the energy conditions large enough to allow assumption (ii) to be violated. Some discussions of wormhole creation are, for example, based precisely on large violations of the energy condition (see ref. [17] and other references cited therein). The other assumption is fairly benign. It might be interesting to try and weaken it even more, but this is unlikely to allow topology change. One possible weakening would replace "every full null geodesic" in the statement by "almost every full null geodesic". It has been argued by Sorkin [54] that such a statement would more truly be a 'generic condition'. Singularity theorems have been proved with this kind of weaker generic condition [55] and it would be worth trying to do the same here.

C] Degenerate metrics

Finally, we return to the interesting idea – discussed by Sorkin [13], Ashtekar [24], Horowitz [20], and others, and mentioned briefly in the Introduction – that it might be possible to allow the metric to vanish or to become degenerate at isolated points and to use these kinds of singularities in order to get topology change. Now, one problem with allowing the metric to vanish is that we cannot compute its inverse and so cannot calculate the curvature at the points where

it vanishes. It appears, however, that the new spinorial variables introduced by Ashtekar [56] to describe relativity might allow this to be done. In this approach we can formulate all the basic equations without having to 'raise and lower indices', i.e., without having to use the inverse of the metric. Also, even in standard general relativity (couched in first-order language) Horowitz [20] has shown that it is possible to construct reasonable topology-changing spacetimes if degenerate metrics are allowed. So, this might well prove to be the correct approach to describing topology change.

It is worth pointing out here that there is a further problem with degenerate (or vanishing) metrics: the causal structure that is normally associated with a Lorentz metric will not necessarily be well-defined. But there is a way around this that allows in some cases a definition of causal relationships, even at points where the metric behaves badly. This will be discussed in detail elsewhere [57].

Acknowledgements

It is a pleasure to thank Rafael Sorkin for sparking my interest in topology change and for several extremely helpful comments on the manuscript. I also thank him, Luca Bombelli and Tom Roman for many stimulating discussions on the topics discussed here, and Matt Visser for some helpful comments on an early version of this paper. Financial support was provided by NSF grants PHY8318350 and PHY8310041 to the Relativity Group of Syracuse University in the initial stages of this work and by a grant from Long Island University during the final stages. The final version of the paper was written when I was a guest, first of the High Energy Theory Group of Brookhaven National Laboratory, and then of the Institute of Cosmology at Tufts University; I thank both institutions for their hospitality.

References

1. J. Wheeler, in *Relativity Groups and Topology*, edited by B.S. Dewitt and C.M. DeWitt, Gordon and Breach, New York (1963).
2. J. Wheeler, *Geometrodynamics*, Academic Press, New York (1962).
3. S.W. Hawking, *Nuclear Physics*, **B144**, 349 (1978); in *General Relativity: An Einstein Centenary Survey*, edited by S.W. Hawking and W. Israel, Cambridge University Press, Cambridge (1979).
4. C.W. Misner and J. Wheeler, *Ann. of Physics (NY)*, **2**, 525 (1957).
5. R.P. Geroch, *J. of Math. Phys.*, **8**, 782 (1967).

6. R.P. Geroch, *Singularities in the spacetime of General Relativity*, Ph.D. Dissertation, Princeton University (1968).
7. F.J. Tipler, *Causality Violations in General Relativity*, Ph.D. Dissertation, Univ. of Maryland (1976).
8. F.J. Tipler, *Ann. of Physics (NY)*, **108**, 1 (1977).
9. F.J. Tipler, *Phys. Lett.*, **165B**, 67 (1985).
10. J.L. Friedman and R. Sorkin, *Phys. Rev. Lett.*, **44**, 1100, (1980).
11. A. Strominger, *Phys. Rev. Lett.*, **52**, 1733 (1984).
12. R. Sorkin, in *Topological Properties and Global Structures of spacetime*, edited by P.G. Bergmann and V. de Sabhata, Plenum, New York (1986).
13. R. Sorkin, *Phys. Rev.*, **D33**, 978 (1986).
14. A. Anderson and B. DeWitt, *Found. of Physics*, **16**, 91 (1986).
15. F.A. Bais, C. Gomez and V.A. Rubakov, *Nucl. Phys.*, **B282**, 531 (1987).
16. C.A. Manogue, E. Copeland and T. Dray, *Pramana*, **30**, 279 (1988).
17. M.S. Morris, K.S. Thorne and U. Yurtsever, *Phys. Rev. Lett.*, **61**, 1446 (1988).
18. M. Visser, *Phys. Rev.*, **D41**, 1116 (1990).
19. J. Friedman, p. 539 in *Conceptual Problems of Quantum Gravity*, edited by A. Ashtekar and J. Stachel, Birkhauser, Boston (1991).
20. G. Horowitz, *Class. Quant. Grav.*, **8**, 587 (1991); p. 1167 in the *Proceedings of the Sixth Marcel Grossmann Meeting (Kyoto, Japan)*, World Scientific, Singapore (1992).
21. G. Gibbons and S.W. Hawking, *Comm. in Math Phys.*, **148**, 345 (1992); *Phys. Rev. Lett.*, **69**, 1719 (1992).
22. J. Friedman, K. Schleich and D. Witt, *Phys. Rev. Lett.*, **71**, 1486 (1993).
23. A. Vilenkin, Tufts Institute of Cosmology preprint (1993).
24. A. Ashtekar, *Lectures on Non-perturbative Canonical Gravity*, World Scientific, Singapore (1991).
25. S.W. Hawking and G.F.R Ellis, *The Large Scale Structure of spacetime*, Cambridge Univ. Press, Cambridge (1973).
26. R.P. Geroch and G.T. Horowitz, in *General Relativity: an Einstein Centenary Survey*, edited by S.W. Hawking and G.F.R. Ellis, Cambridge University Press, Cambridge (1979).
27. P. Yodzis, *Comm. in Math. Phys.*, **26**, 39 (1972); *Gen. Rel. and Grav.*, **4**, 299 (1973).
28. J.W. Milnor and J.D. Stasheff, *Characteristic Classes*, Princeton University Press, Princeton (1974).

29. J.W. Milnor, *Topology from the Differentiable Viewpoint*, Univ. Press of Virginia, Charlottesville (1965).
30. B.L. Reinhart, *Topology*, **2**, 173 (1963).
31. E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw Hill, New York (1955).
32. S.W. Hawking and R. Penrose, *Proc. Roy. Soc. Lond.*, **A314**, 529 (1970).
33. J.K. Beem and S.G. Harris, *Gen. Rel. and Grav.*, **25**, 939 (1993); *Gen. Rel. and Grav.*, **25**, 963 (1993)
34. A. Borde, *Class. and Quant. Grav.*, **2**, 589 (1985).
35. A. Borde, *Class. and Quant. Grav.*, **4**, 343 (1987).
36. A. Borde, *Phys. Lett.*, **102A**, 224 (1984).
37. R. Sorkin, *Phys. Rev. Lett.*, **51**, 87 (1983); **54**, 86(E) (1985).
38. F.J. Tipler, *Phys. Rev.*, **D17**, 2521 (1978); *J. of Diff. Eqns*, **30**, 165 (1978).
39. T.A. Roman, *Phys. Rev.*, **D33**, 3526 (1986); *Phys. Rev.*, **D37**, 546 (1988).
40. C. Chicone and P. Ehrlich, *Manuscripta Math.*, **31**, 297 (1980).
41. U. Yurtsever, *Class. Quant. Grav.*, **7**, L251 (1990).
42. G. Klinkhammer, *Phys. Rev.*, **D43**, 2542 (1991).
43. R. Wald and U. Yurtsever, *Phys. Rev.*, **D44**, 403 (1991).
44. L. Ford and T. Roman, Tufts University preprint (1994).
45. S.W. Hawking, *Comm. in Math. Phys.*, **43**, 199 (1975).
46. J. Bardeen, *Phys. Rev. Lett.*, **46**, 382 (1981).
47. R. Penrose, unpublished remark.
48. L. Ford and T. Roman, Tufts University preprint (1994).
49. R. Sorkin, *Int. J. of Theor. Phys.*, **25**, 877 (1986).
50. J. Munkres, *Elementary Differential Topology*, Princeton University Press, Princeton (1966).
51. K. Uhlenbeck and D. Freed, *Instantons and Four-Manifolds*, Springer Verlag, New York (1984).
52. E. Witten, *Comm. in Math. Phys.*, **100**, 197 (1985).
53. J. Milnor, in *Lectures in Modern Mathematics, II*, edited by T.L. Saaty, Wiley, New York (1964).
54. R. Sorkin, unpublished remark.
55. A. Borde, *J. of Math. Phys.*, **28**, 2683 (1987).
56. A. Ashtekar, *Phys. Rev. Lett.*, **57**, 2244 (1986); *Phys. Rev.*, **D36**, 1587 (1987).
57. A. Borde and R. Sorkin, in preparation.