ANALYSIS OF THE GENERIC SINGULARITY STUDIES BY BELINSKII, KHALATNIKOV, AND LIFSCHITZ

John D. BARROW

Department of Astrophysics, Univ. of Oxford, U.K.

and

Department of Astronomy, Univ. of California, Berkeley, Cal. 94720, U.S.A.

Frank J. TIPLER

Department of Mathematics, Univ. of California, Berkeley, Cal. 94720, U.S.A.

NORTH-HOLLAND PUBLISHING COMPANY - AMSTERDAM

PHYSICS REPORTS (Review Section of Physics Letters) 56, No. 7 (1979) 371–402. North-Holland Publishing Company

ANALYSIS OF THE GENERIC SINGULARITY STUDIES BY BELINSKII, KHALATNIKOV, AND LIFSCHITZ

John D. BARROW

Department of Astrophysics, Univ. of Oxford, U.K.
and
**Department of Astronomy, Univ. of California, Berkeley, Cal. 94720, U.S.A.*

and

Frank J. TIPLER**

Department of Mathematics, Univ. of California, Berkeley, Cal. 94720, U.S.A.

Contents:

*Lindemann Fellow 1977/78.
** Present address: Dept. of Physics, Univ. of Texas at Austin, Austin, Texas 78712, U.S.A.

Single orders for this issue

PHYSICS REPORTS (Review Section of Physics Letters) 56, No. 7 (1979) 371–402.

Copies of this issue may be obtained at the price given below. All orders should be sent directly to the Publisher. Orders must be accompanied by check.

Single issue price Dfl. 13.00, postage included.

Abstract:

We apply some modern mathematical methods of global analysis to a series of studies undertaken by Belinskii, Khalatnikov and Lifschitz (BKL) to elucidate the structure of space-time near a general cosmological singularity. A brief summary of BKL's large body of work on *inhomogeneous* cosmological models is given (their work on homogeneous models is not under discussion here). Various theorems are proven and analyses of a mathematical and physical nature are made to show that the constructions of BKL cannot be general and in some cases do not give Lorentz manifolds. We conclude that although the work of BKL has led to very significant advances in our understanding of the dynamics of *homogeneous* cosmological models, the local techniques they employ do not extend to give us reliable information about the *global* structure of generic space-times. A detailed discussion of stability, generality, function counting, linearization stability, physical singularities and fictitious singularities is given together with an outline of various physical considerations which might be useful in future studies of the structure of generic space-times.

'Any problem which is non-linear in character, which involves more than one coordinate system, or where structure is initially defined in the large, is likely to require considerations of topology and group theory for its solution'.

– M. Morse

'Approximation methods are of no avail since one never knows whether or not there exists to a particular approximate solution an exact solution . . .'

– A. Einstein

1. Introduction

The aim of this paper is to compare some results of global analysis with the general cosmological singularity studies undertaken by the collaboration of Belinskii, Khalatnikov, Lifschitz (BKL) and coworkers over the period 1961–72 [1–15, 73, 74]. By employing global techniques we are able to show that their conclusions and constructions concerning the general (*inhomogeneous*) singularity structure are either incorrect or do not describe the generic case. We shall stress the power and generality of the topological methods introduced into gravitation physics by Penrose [19], and use them to draw into question certain features of the localized description adopted by BKL by displaying a number of theorems concerning the large-scale structure of space-time.

The importance of ascertaining the structural form of generic cosmological space-times is unquestioned. Knowledge of the nature of such inhomogeneous and anisotropic cosmological models would undoubtedly shed light upon: (a) the origin and relative likelihood of the observed large scale structure in the Universe; (b) the origin and entropy content of the microwave background radiation; (c) the presence or absence of horizons near a general singularity; (d) the possible effects of particle creation near $t_{\rm pl} \sim 10^{-43}$; (e) the binding energy of the universe; and (f) chaotic gravitational collapse.

The large body of work performed by BKL on *inhomogeneous* cosmologies has concentrated upon the use of local differential techniques to conclude that two of the most general spatially homogeneous metrics of Bianchi types VIII and IX contain all the essential dynamical features of a general integral of the Einstein equations in some open neighbourhood. This work began in 1961 motivated by the Landau–Komar–Raychaudhuri [14, 20, 21] discovery of an apparently inevitable singular infinity in the Einstein equations for the density of the universe some finite time ago with respect to a set of freely-falling coordinates. A series of studies argued initially that such "singularities" were in general either fictitious – related only to the breakdown of the accompanying coordinate system, or unstable, created by the special symmetries of those simple models investigated exactly [6]. The general structure of such "fictitious" singularities was investigated using local analyses together with a function counting criterion to establish their relative generality.

Elsewhere, studies of this "apparent singularity" proceeded using global techniques of differential topology to describe the conformal structure of space-time This programme was initiated by Penrose and [19] produced rigorous results of a very general but less dynamically detailed character [22]. In particular it revealed that, contrary to the BKL conclusions, the singularities arising in the cosmological problem were in fact physically real, coordinate independent and completely general features of the universal gravitational field [23, 61].

The early conflict between the results of these two distinct approaches to the cosmological problem is now well known and BKL in particular have discussed the relevance of the global theorems [12, 73]. However, following this initial debate concerning the physical reality of the singularity these two approaches do not seem to have again encountered common ground. We feel that it is essential that the subsequent results obtained using these different methodologies be confronted again in order to reveal the limitations and pitfalls of basing analyses upon the local structure alone or upon some approximation scheme. The question is particularly apposite since BKL have pursued a long programme of work to establish the form of the general (inhomogeneous) cosmological behaviour near a physical singularity using the "Mixmaster" evolution as a paradigm. The results of this work have been widely quoted in the literature as representing the general solution to the Einstein equations near a cosmological singularity. In this later work similar techniques to those which had previously led to incorrect conclusions regarding the reality of physical singularities have continued to be employed. We shall show that many of the later conclusions drawn with regard to inhomogeneous cosmologies are also false and in addition we shall reformulate many of the problems investigated in more precise and well-defined fashion. We proceed as follows: In section 2 a summary is given of the concepts and mathematical machinery required to formulate later results. In particular precise definitions are given of stable and generic solutions to the field equations; the function counting technique is rigorously formulated and the correspondence between the Penrose and BKL notions of physical singularity indicated. Sections 3 and 4 give the principal theorems and between them we divide the BKL work into two periods, designated pre- and post-Penrose, reflecting the slight change of emphasis in these studies following emergence of the singularity theorems. In each of these sections we give a brief resumé of the arguments and conclusions of their original papers before commenting upon their results in the light of recent developments in global analysis. In section 5 additional considerations are described, some of which are concerned with the innate limitations of an approximation method such as the one used by BKL. Others are merely aimed to highlight certain features which must be encompassed by any analysis of the structure of a generic singularity, whether global or local. Finally a brief conclusion is drawn up: we feel that the local approach to the analysis of singularities used by BKL to study *homogeneous* space-times does not appear to give correct descriptions of generic, *inhomogeneous* space-times.

Notation. Unless otherwise specified the notation will be that of Hawking and Ellis [61]. In particular the Einstein equations are $R_{ab}-\frac{1}{2}g_{ab}R=8\pi T_{ab}$. Also h_{ab} is the induced metric on a spacelike hypersurface and χ_{ab} is the extrinsic curvature of the hypersurface.

2. Definitions and mathematical background

2.1. Space-time, stability and generality

The object of study in this paper is a *space-time*, which is defined to be a pair, (M, g), where M is a real, four-dimensional, orientable and connected C^∞ Hausdorff manifold without boundary and g is a C^∞

(C^2 would be sufficient), Lorentz metric defined on M. (M, g) is C^∞ inextendible and space and time orientable [63].

The only physically interesting properties of space-times are those which are *stable*, i.e., those properties which still occur when the initial data is perturbed slightly. To be more precise:

Definition 1 [64]. A space-time property will be said to be D-*stable* (or more simply, *stable*) about a spacelike hypersurface S with initial data $(h^0_{ab}, \chi^0_{ab}, \psi^0_{(i)})$ if the property exists in all space-times maximally developed from the initial data $(h_{ab}, \chi_{ab}, \psi_{(i)})$ in some open neighbourhood of the initial data $(h^0_{ab}, \chi^0_{ab}, \psi^0_{(i)})$ on S in the original space-time. Here $\psi_{(i)}$ denotes the non-gravitational fields and their derivatives on S. We use the original metric h^0_{ab} on S to define a distance function and hence a topology on the space of all initial data on S, i.e., we use the C^∞ open topology on S, see refs. [64] and [65] for more details.

It is of course possible for a space-time property to be stable and yet be of no physical interest: stability is a necessary but not a sufficient condition for physical relevance. For example, a property could be stable only in a neighbourhood of initial data sets corresponding to presently highly anisotropic universes. Since the actual universe is n v apparently almost isotropic such a property could not occur in our universe. A sufficient condition for a stable property to be of physical interest is the requirement that the property be *generic* [65], (or *general*). Roughly speaking, a stable property is generic if it occurs near *every* initial value set on S. More precisely:

Definition 2. A space-time property will be said to be *generic* to (maximally extended) space-times evolved from a spacelike hypersurface S if the property is open dense on the space of all initial data $(h_{ab}, \chi_{ab}, \psi_{(i)})$ on S, where h_{ab} is positive definite and the non-gravitational fields $\psi_{(i)}$ are restricted in some specified way to "physically realistic" fields. (The Einstein equations, various equations of state, etc. are assumed to hold on the evolved space-times.)

We shall occasionally speak of "generic" (or general) space-times. By this phrase we shall mean that the properties of the space-time structures under discussion are generic in the set of all initial data which allow the given structures to exist. For example, if we speak of "the properties of singularities in generic space-times", we shall be talking about those properties of singularities which exist in an open dense subset of all initial data which evolve singularities.

There are several topologies one could use on initial data space in order to define the concepts of "stable" and "generic". The topologies used in the above definitions are probably sufficiently fine so that properties which are stable and/or generic by the above definitions would be respectively stable and/or generic in any other reasonable topology. However, for certain applications one might want to use a coarser topology. For example, the above topology does not distinguish between variations of the initial data due purely to coordinate transformations, and variations of the true gravitational degrees of freedom. Thus, it might happen that a space-time property is stable under small variations of the true gravitational degrees of freedom, and yet is unstable by the above definition because the property depends in some way on the coordinate system, or rather on the choice of the initial hypersurface from which the space-time is evolved.

In this paper we shall be concerned with properties which continue to exist in $D^-(S)$ when the data on S is perturbed. ($D^-(S)$ is the past maximal Cauchy development of S; see HE [61] page 201 or below for the definition.) Clearly the question of which properties are stable in this sense is going to depend on which S is chosen. Two hypersurfaces S and S′ are said to have the same initial data in the space of true gravitational degrees of freedom if D(S) is isometric to D(S′), even though $D^-(S)$ is not isometric to

$D^-(S')$, where $D(S) \equiv D^+(S) \cup D^-(S)$ and $D^+(S)$ is the future Cauchy development of S. Thus we shall want a finer topology than the topology on the space of the true gravitational degrees of freedom; we shall sometimes wish to say that a property is unstable if it is unstable in the quotient topology defined by the topology used in the above definition of stability modulo the following identification: two initial data sets on S, $(h_{ab}, \chi_{ab}, \psi_{(i)})$, $(\hat{h}_{ab}, \hat{\chi}_{ab}, \hat{\psi}_{(i)})$ are said to be the same if $D^-(S)$ is isometric to $\hat{D}^-(S)$, and we shall not distinguish between diffeomorphisms of the initial data which preserve the isometry $D^-(S) \triangleq \hat{D}^-(S)$. We shall call the notion of stability with this topology *past* D-stable. In theorem 5 below (section 4), the existence of synchronous coordinate system (based on S) which reaches the singularity will be shown to be both D-unstable and past D-unstable. We shall also show that the BKL "general" solution with a physical singularity is unstable to infinitesimal perturbations of the true gravitational degrees of freedom.

The initial data will be imposed on a surface S called a *partial Cauchy surface*, which is defined to be a C^2 space-like boundaryless hypersurface which no timelike or null curve intersects more than once. The *past Cauchy development* $D^-(S)$ of S is the set of all points p such that all future-directed timelike or null curves intersect S. The Cauchy development $D^+(S)$ is defined analogously. If the Cauchy development $D(S) \equiv D^-(S) \cup D^+(S)$ is the entire space-time, then $(\mathcal{M}, g)$ is said to be *globally hyperbolic* and S is called a *Cauchy surface*.

2.2. Singularities

The following definitions make precise the notions of physical and fictitious singularities in our usage.

Definition 3. A timelike geodesic generator γ of a synchronous coordinate system will be said to terminate in a fictitious singularity at the point p on γ if p is a point conjugate (see HE [61] p. 100) to S along γ, where S is a C^2 spacelike hypersurface which is orthogonal to all timelike geodesic generators of the synchronous coordinate system [109]. (S is a t = constant slice of the coordinate system.)

Intuitively speaking, there is a fictitious singularity at p if the timelike geodesics which are orthogonal to S and close to γ, intersect γ at p forming a caustic of the geodesic congruence.

This definition corresponds to that used by BKL.

Definition 4. A non-spacelike geodesic γ will be said to terminate in a *physical singularity* if γ is incomplete.

It is interesting to note that by the above definition 4, a physical singularity need not correspond to a divergence of curvature invariants (such as $R_{ab}R^{ab}$, R, etc.), as the finite limit of proper time is approached along γ. That is, γ need not terminate in a *scalar polynomial curvature singularity* (s-p curvature singularity). However, Clarke [67] has shown that if the space-time is inextendible (to a certain order of diffentiability) and if the space-time curvature tensor does not approach type D in the limit, then a physical singularity corresponds to a *p-p curvature singularity*. That is, some component of the curvature tensor in a frame parallel propagated, (p-p), along γ must diverge as the physical singularity is approached along γ. Now Siklos [62] has shown that p-p curvature singularities which are not s-p curvature singularities are a set of measure zero in the space of all homogeneous initial data. This strongly suggests that in physically realistic space-times s-p and p-p curvature singularities are equivalent.

The term "physical singularity" is used by BKL [8] to mean s-p curvature singularity. Thus the work of Siklos and Clarke shows that, in any physically realistic space-time, the BKL definition of physical singularity is probably equivalent to that of definition 4.

With the conceptual machinery introduced in the preceding paragraphs of this section it is easy to prove a simple singularity theorem.

Theorem **1. Let S be a Cauchy surface with $\chi^a_a \geq C > 0$, where C is a constant, and suppose $R_{ab}V^aV^b \geq 0$ for all timelike vectors V^a. Then all inextendible timelike geodesics terminate in the past in a physical singularity.**

Proof. **Landau and Lifschitz [75] have shown (using the Landau–Komar–Raychaudhuri equation) that $R_{ab}V^aV^b \geq 0$ and $\chi^a_a \geq C > 0$ implies every past-directed timelike geodesic γ orthogonal to S must encounter a fictitious singularity (i.e. there must be a point conjugate to S along γ) within a distance $3/C$ of S, providing γ does not terminate in a physical singularity first. Since S is a Cauchy surface, (M, g) is globally hyperbolic. Using global analysis it can be shown (HE [61] p. 217) that to each point q in $D^-(S)$ there is a timelike geodesic orthogonal to S which is of maximal proper time length from S to q, and which contains no fictitious singularity between S and q (that is, the synchronous coordinate system can be contained from S to q). Since every timelike past-directed geodesic orthogonal to S has either terminated in a physical singularity or encountered a fictitious singularity within a distance $3/C$, this means that any timelike geodesic which has a length to the past of S greater than $3/C$ must leave $D^-(S)$. But, S is a Cauchy surface and, by definition of Cauchy surface, *all* inextendible timelike curves hit S and pass into $D^-(S)$. Furthermore, since (M, g) is globally hyperbolic, $D(S)$ ($\equiv D^+(S) \cup D^-(S)$) is the entire space-time, and so a timelike curve entering $D^-(S)$ never leaves it. Thus, all inextendible timelike geodesics terminate in the past in a physical singularity. Q.E.D.**

2.3. *Function counting*

It appears that the first attempts to make function counting a rigorous criteria for characterizing solutions to sets of differential equations were those of Einstein [120, 121]. In the 1955 edition of his book "*The Meaning of Relativity*" he introduced the notion of the *strength* of a system of differential equations as a means of analysing both the generality and compatibility of entire sets of field equations. His intuitive idea was that 'the smaller the number of free data consistent with the system of field equation, the stronger is the system'. To quantify this idea he first expanded all the dependent field variables in Taylor series and then determined the number of relations among the various nth order coefficients that are imposed by the differential equations themselves. In general the number of coefficients in the nth order expansion of an analytic function on an N-dimensional manifold is

$$\begin{bmatrix} N \\ n \end{bmatrix} = \frac{(n+N+1)!}{n!(N-1)!}.$$

Taking into account remaining coordinate transformations, the number of constraints which limit the way in which the initial data may be chosen can also be evaluated combinatorically and subtracted from $[^N_n]$. This gives Z_n, the number of free coefficients of order n, where Z_n must be non-negative for the equations in the system to be compatible. Einstein found that for the systems he evaluated, including the vacuum Einstein equations, the limit as $n \to \infty$ of $Z_n/[^N_n]$ is $O(n^{-1})$ and he defined the *strength* of the system to be the coefficient of n^{-1} in this limit. The larger the value of this coefficient the more weakly constrained is the system of equations.

The BKL studies use the function counting criterion to establish the generality of their metrics and approximations. We now give a precise formulation of this technique for vacuum space-times which will be sufficient for our present investigations.

If S is a C^2 spacelike partial Cauchy surface, then it will have a unique maximal Cauchy development D(S), which will coincide with the entire space-time (M, g) if (M, g) is globally hyperbolic. If the matter equations are well-posed and if the Einstein equations hold, then D(S) is uniquely and completely determined by the initial data on S. It has been rigorously shown by York [70] and by Fischer and Marsden [76] that for matter-free closed universes away from initial data with symmetries, the initial data for the Einstein equations can be completely defined by four functions of three variables. That is, there is locally a diffeomorphism from the space of pairs of symmetric three-tensors (h_{ab}, χ_{ab}) on a compact spacelike hypersurface S (where h_{ab} and χ_{ab} satisfy the empty space Einstein equations and tensor fields which differ only by a coordinate transformation are identified) onto the Hilbert space consisting of four arbitrary functions of three variables. (As will be shown in theorem 7 below, there is no such diffeomorphism near spacetimes with symmetries.) Thus, these four functions define D(S) uniquely and completely.

However, (M, g) may not be globally hyperbolic – indeed, global hyperbolicity is not a stable property for all initial data sets, since in any neighbourhood of Schwarzschild initial data there is Reissner–Nordstrom initial data and Reissner–Nordstrom space is not globally hyperbolic. Thus, the four functions specified on S may not be sufficient to determine completely the entire structure of (M, g). For example, if a timelike singularity occurs to the future of a partial Cauchy surface S, then four functions given on S will determine D(S), but additional functions must be given to specify what emerges from the timelike singularity (roughly speaking, a casual curve γ is said [99, 107] to terminate in a *timelike singularity* if there is an observer for which γ is initially in his future but later in his past). Thus, if a space-time contains a timelike singularity then *more* than four independent functions must in general be given in order to define the space-time. In particular, if a space-time contains a timelike singularity, then a demonstration that g is defined by four independent, arbitrary, functions of three variables does not imply that the properties of g are "generic" in any meaningful sense.

What happens in the timelike singularity case is that although four independent functions are sufficient to determine a stable non-globally hyperbolic solution *locally*, they are insufficient to determine it *globally*. Some additional information must be supplied to determine the global structure. This information will usually be of a global nature, and so local analyses such as the BKL approach will give no indication that information beyond four functions is required. What a local analysis does is use four functions to determine a solution in some small coordinate neighbourhood, and then piece together neighbourhoods which cover the entire space-time. Matching conditions are imposed in an attempt to prevent additional information from entering. However, it could be that these coordinate neighbourhoods could be matched in inequivalent ways, or that a given coordinate neighbourhood could be matched at the same point to any one of several inequivalent other neighbourhoods. As an example at the first possibility, recall that opposite ends of a square can be identified in one way to form a cylinder but in another way to form a Möbius strip. A second example is Taub space, which has two different analytic extensions into NUT space, depending on which family of null geodesics is assumed to go through the boundary between Taub space and NUT space ([61] p. 170). The exterior Schwarzschild solution provides an example of the second possibility. There are many C^2 extensions through the event horizon besides the analytic Kruskal extension [116–119]. The requirement that a region be in the Cauchy development of another is much stronger than merely requiring that the region be an extension of another. The additional information required in the timelike singularity case would probably be due to the second possibility.

Since non-globally hyperbolic space-times require in general more than four independent functions for their complete specification, one might be tempted to regard such space-times as more "general" than globally hyperbolic space-times. However, this need not be the case. The relative degree of "generality" of a space-time property is determined primarily by the relative "size" of the region in initial data space for which the property holds in the space-times maximally developed and extended from initial data "points" in the region. To define "size", we must of course put a measure on initial data space, and the problem of which is the best measure to use is still unsolved. In particular, the relative sizes of the two initial data space regions which respectively evolve globally hyperbolic and non-globally hyperbolic space-times is unknown. It is even possible that the non-globally hyperbolic region is vastly greater in size for any reasonable measure, and yet all the "physically realistic" initial data sets lie in the globally hyperbolic region. Penrose has conjectured that *all* physically realistic initial data sets lie in the globally hyperbolic region. He calls this conjecture the postulate of *strong cosmic censorship* [99].

If it is in fact the case that physically realistic space-times are not globally hyperbolic, then the degree of generality of a property will in reality be determined by both the generality of the property in the initial data region, *and* by the generality of the property in what we might call "extension space". That is, by the relative "number" of extensions from a given D(S) which have the property compared to the "number" of extensions which do not.

The foregoing discussion of generality in non-globally hyperbolic space-times is of course very vague, but it will suffice for our purposes, since we are concerned in this paper primarily with properties defined in D(S). To discuss stability and generality properly in non-globally hyperbolic space-times one really should use the notion of G-stability [64].

It was remarked above that in general a stable solution of the empty space Einstein equations is defined in D(S) by four independent functions given on S. However, the topology of S will sometimes impose global constraints on these functions. For example, suppose S has the topology of the three-torus and suppose we choose one of the independent functions to be the trace of the extrinsic curvature of S. If we were to set $\chi^a_a = 0$, then one might expect that we would be free to set 3 more functions arbitrarily on S. (In fact, in York's initial value scheme [70], the condition $\chi^a_a = 0$ is regarded as a mere coordinate condition on the space-time; one still has four functions free.)

This is not true in general. Schoen and Yau have shown [71] that if the strong energy condition holds, and $\chi^a_a = 0$ on an S with topology T^3, then the entire space-time is flat. That is, if the topology is T^3, then the entire space-time geometry is determined by the single condition $\chi^a_a = 0$.

This illustrates a key problem: in order to show that the general initial value problem has been well-posed, one must show that the four functions of three variables (or more generally, the arbitrary number of functions of an arbitrary number of variables) are in fact independent and that they completely parameterize the initial value space $\mathscr{C}$ (see ref. [76] for a precise definition of $\mathscr{C}$). Now all possible four functions of three variables on S form a W^s Hilbert space $\mathscr{H}$ (that is, $\mathscr{H}$ is a Sobolev space W^s – see ref. [61] p. 234 for a definition of W^s. We will choose $s \geq 3$ so that the maximal development will be unique [100]). Thus $\mathscr{C}$ can be parameterized by four independent functions of $\mathscr{H}$ if and only if there is a local diffeomorphism f: $\mathscr{C} \to \mathscr{H}$. The map f will be a local diffeomorphism if and only if the Frechet derivative Df of f is locally 1–1 and onto $\mathscr{H}$. (In a finite number of dimensions this is equivalent to requiring that the Jacobian of the transformation be non-zero. See ref. [111] for a discussion of this point and the definition of the Frechet derivative.) The usual way of demonstrating that Df is locally 1–1 and onto is to first show that Df is an elliptic operator with trivial kernal (this shows that Df is 1–1), and to then prove that the adjoint operator D*f also has trivial kernal. By the Fredholm Alternative Theorem [105], this implies that Df is locally 1–1 and onto.

Fischer, Marsden and Moncrief have shown [77] that Df will be 1–1 and onto a neighbourhood U of $\mathscr{H}$ if and only if the space-time which is evolved from a point in $\mathscr{C}$ is linearization stable.

2.4. Linearization stability

BKL express their "general" metric in terms of a series expansion (see sections 3 and 4 for details and references), and they focus attention on the first order terms. This procedure raises the question of whether the series converges. In particular, is the first order term a good approximation to the actual metric? One way of attacking these questions is to ask "Is the space-time linearization stable?"

If a space-time is linearization stable, then the first two terms of the expansion

$$g_{ab} = g^0_{ab} + \epsilon g^1_{ab} + \epsilon^2 g^2_{ab} + \cdots$$

where g_{ab} and g^0_{ab} are solutions to the full Einstein equations, and g^1_{ab} is a solution to the linearized equations, can be completed to a convergent expansion. That is, there exists a neighbourhood (in solution space) of the original metric g^0_{ab} such that every metric g_{ab} in the neighbourhood can be expressed in terms of a convergent power series, with $g^0_{ab} + \epsilon g^1_{ab}$ being a good first order approximation to g_{ab}.

Fischer and Marsden [68], together with Moncrief [69] have shown that g^0_{ab} is linearization stable if and only if g^0_{ab} has no Killing fields. (Their theorem assumes that (M, g^0) contains a compact Cauchy surface, and that $T_{ab} \equiv 0$. Nothing is known about the linearization stability properties of non-asymptotically flat, open universes.) Thus, if g^0_{ab} contains Killing fields, then there exist solutions g^1_{ab} to the linearized equations such that $g^0_{ab} + \epsilon g^1_{ab}$ *cannot* be completed to a convergent series; the expression $g^0_{ab} + \epsilon g^1_{ab}$ is *not* a good first order approximation to a real solution g_{ab} to the full Einstein equations. Such solutions g^1_{ab} to the linearized equations are called *spurious perturbations*. In particular, if one is perturbing a Bianchi model, the linearized equations will have solutions which are spurious perturbations

It is possible that there exist solutions g^1_{ab} to the linearized equations such that $g^0_{ab} + \epsilon g^1_{ab}$ can be completed to a convergent series even though g^0_{ab} is not linearization stable, Fischer, Marsden and Moncrief [77] have shown that a given g^1_{ab} is not a spurious perturbation if and only if g^1_{ab} satisfies a certain 2nd order constraint (integrals of the Taub conserved quantity vanish; i.e., g^1_{ab} satisfies the equation in Theorem 2 of ref. [79]). Thus, if g^0_{ab} contains Killing fields, then each solution g^1_{ab} to the linearized equations must be checked to see if it is a real and not a spurious perturbation by checking if it satisfies a certain complicated equation.

Fischer, Marsden and Moncrief [77] have shown that a sufficient condition for the existence of non-spurious solutions to the linearized equations is the existence of a maximal spacelike hypersurface (or more generally, a χ^a_a = constant hypersurface) with a Killing field tangent to the hypersurface. Thus, the Bianchi models have both spurious and non-spurious perturbations.

The above discussion on linearization stability applied to the matter-free case, which is sufficient for our purposes since BKL also restricted their arguments in most cases to the matter-free case. Nevertheless, if should be mentioned that the presence of matter can affect linearization stability. For example, D'Eath has shown [112] if the matter is in the form of a perfect fluid, and the equations for the fluid and the space-time geometry are formulated in such a way as to form an unconstrained Hamiltonian system, then the space-time is linearization stable even if Killing vectors are present. However, if these equations are formulated in terms of a constrained Hamiltonian system and the constraints generate gauge transformations (as would be the case if the fluids were represented by Schutz potentials [113]), then as in the empty-space case the space-time is linearization *unstable* if it contains Killing vectors [114]. Furthermore, a closed universe with Killing vectors is linearization unstable if the matter is in the form of gauge fields or scalar fields [78]. Thus the question of linearization stability in the presence of matter depends on the way the matter tensor is represented.

3. Pre-Penrose work of BKL

This period of work by BKL before the proof of the singularity theorems we shall take to encompass the following pieces of published work by the Soviet school [1–9, 74]. A summary of the principal conclusions of [1–4, 74] is contained in the well-known review [6]. Although chronologically the paper [8] belongs in the post-Penrose period, we include it here using acknowledgement of the existence of physical singularities as the criterion for admission. The paper [4] will also be discussed in section 5 since it is useful as an example of the ambiguity of local coordinate dependent techniques. Various parts of the work in [1–9, 74] are quoted and regarded as canonical in the most recent edition of the text-book [14].

For ease of subsequent discussion we categorize and briefly summarize the principal ideas and conclusions of this body of work before making some comments in the light of global methods. The central ideas are: (1) The mathematical methodology; (2) physical singularities; (3) fictitious singularities. We now treat these in turn.

3.1. Mathematical methodology

Local series techniques and function counting.

Resumé. The goal of the programme was to elucidate the nature of the general cosmological solution to the Einstein equations. They explain [80], "By general solution we mean a solution in which the physical arbitrariness is determined by four arbitrary functions of three spatial coordinates in vacuum and by eight such functions in a space with matter". The manner in which such a solution is sought is described by Lifschitz [81], "Assuming the existence of a singularity we must find in its neighbourhood the widest class of solutions and then judge by the number of arbitrary functions whether it is the general solution". Local series methods were employed to expand the field equations in power series within a small neighbourhood of the Friedmann and Kasner metrics; the cardinality of these solutions being then ascertained by calculating the number of arbitrary functions remaining after the coordinate system was fixed. Thus it was claimed that such a neighbourhood of the Kasner metric near a regular point was generic [6, 14]. These local series techniques were used throughout the analyses of [1–9, 74] in the treatment of the general structure problem and in fact also in the later work; see section 4.

Comments. We begin by pointing out that, in the light of definitions 1 and 2 of section 2 the stated aim of the BKL programme is actually to find a *stable* rather than a generic singularity structure. In the work cit. d above, series techniques were employed to construct local neighbourhoods around particular solutions on initial hypersurfaces where four 'arbitrary' functions of the three spatial variables would specify the metric. Referring back to section 2 we point out that for such a methodology to succeed the space-time must of course be globally hyperbolic and in order for four functions to completely and uniquely specify the development of the entire space-time they must be given on a global Cauchy surface. In addition, if part of the initial singularity were of timelike character then four arbitrary functions of three variables given on any spacelike hypersurface would be inadequate to develop the entire space-time; see section 2. BKL also use the four function criterion *locally* to claim small neighbourhoods can be representative of the generic global case if they may be specified by four such independent functions of three variables. We point out that their claim is unjustified, as the following illustrates:

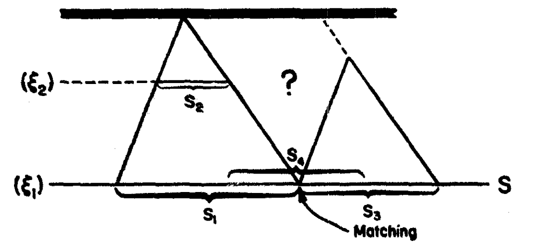

Fig. 1. Initial data specified on non-overlapping regions S_1 and S_2 determines evolution only in the triangular regions. The "?" region is contingent upon the matching conditions between these two regions. The symbols ξ_1, ξ_2 are defined in subsection 4.3.

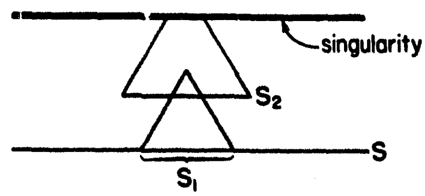

Fig. 2. Data given on S_2 cannot be 'generic' in any meaningful sense if the development of S_2 is to reach the singularity. A sufficiently large change of the initial data on S_2 will change it into that data appropriate to the neighbourhood S_1, whose development does not reach the singularity.

Fig. 1 is a generic Penrose diagram for any space-time with event horizons (or particle horizons for the time reverse of fig. 1).

If data is specified on the two adjacent neighbourhoods S_1 and S_3 of the hypersurface S then we can determine only what lies in the triangular developments of S_1 and S_3. Now as can be seen from fig. 1, most of the information about the singularity lies in region "?" whose properties can be determined only when we have information about the matching region. Thus, a study of S_1 will not reveal anything about the large scale structure of the singularity; this global structure is contingent upon the matching conditions between S_1 and S_3 on S. We have stressed that BKL seek a *stable* rather than a *generic* data set of S. Now suppose we can pick a neighbourhood overlapping S_1 and S_3, then by taking a suitable number of such overlaps we can sweep out the structure of the entire singularity. However, this can be done only if the solution we have constructed in S_1 is not merely stable but has the same properties on a neighbourhood (in solution space) of large *finite* size rather than merely an infinitesimal neighbourhood, and this is evidently not the case for BKL's constructions as can be seen from fig. 2.

We note that even if four functions of three variables are given on a global Cauchy surface they must be shown to be *independent* by proving that the map from true-degrees-of-freedom space into the Hilbert space of four functions of three variables is locally a diffeomorphism. This does not appear to have been done by BKL, and indeed we shall demonstrate below (theorem 6) that no such diffeomorphism exists near homogeneous closed universes. Finally, we remark that it is desirable that any series technique display an accompanying convergence condition which may be explicitly checked. BKL have not done this and we shall see in what follows that their series representations evidently diverge from any real solutions to the Einstein equations since they lead globally to the construction of artefacts which cannot in fact even be space-times.

3.2. Physical singularities

Resumé. Motivated by the Landau–Komar–Raychaudhuri [14, 20, 21] discovery that the determinant of the spacelike three-metric of a synchronous coordinate system must vanish at some finite time in the past, it was argued [6] that this result displayed *only* the tendency for the coordinate volume described in the synchronous reference system of geodesics to disappear. Thus it was felt that no conclusions drawn from it regarding infinities in physical quantities could be invariant in general, as [82], "this singularity which is inevitable in the synchronous system is actually fictitious in the general case". In their work [83], "a physical singularity is taken to be a singularity which appears in scalar quantities, with a direct physical meaning, such as the energy density of the matter, or in second-order invariants of the curvature tensor...". Regarding exact solutions, like Friedmann's which manifestly contain such physical singularities it was judged that these [84], "...have a generality which is insufficient to take into account arbitrary initial conditions". They claim that such physical singularities may in general be removed from these special space-times since [84], "there exist small perturbations of a type such that their superposition leads to the vanishing of the singularity, this means that it should go over into a fictitious singularity as a result of the perturbation". It was thought that when the geodesic congruence forming the synchronous reference system became singular a transformation could always be made to a new synchronous system and this used to trace the evolution of the space-time until that in turn also broke down, and so on... ad infinitum. No consideration of the global behaviour in the limit for this procedure was given and they draw the [85] "...conclusion that the general solution has no true singularity" using these local arguments.

Comments. These early conclusions of BKL concerning physical singularities have received much comment in earlier work [25] and thus we shall not labour our discussion of this point. The most instructive discussion of these results is that of MacCallum [26] who stresses how the vital question of what occurs in the limit of these successive synchronous coordinate system exchanges was not considered. The result, now well known, is that a real physical singularity inevitably occurs in general and this is precisely stated by the Hawking–Penrose theorem [23, 61]. We have given in section 2 an argument for the equivalence, in practice, of the Penrose and BKL definitions (see section 2), of physical singularity and this clarifies some ambiguities that still seem to exist over this question [14].

3.3. Fictitious singularities

Resumé. It was concluded from the arguments described above that the general solution could contain only a fictitious singularity generated by the geometric specification of a particular coordinate system, and which would disappear upon transfer to another synchronous reference system. BKL then proceeded to determine the nature of the general solution to the vacuum gravitational field equations containing a fictitious singularity. Local series expansions of the Einstein equations were used to analytically construct such a solution and it was claimed that the fictitious singularity was, in the general (four function) case, *simultaneous.* Thus, it was claimed [86], "the general solution to the equations of gravitation can also be represented (by suitable choice of the synchronous frame) in a form in which the singularity is simultaneous for all of space". Belief in the validity of these results was maintained in later analyses since it was necessary to have such a synchronous coordinate grid which could be followed all the way into the singularity, breaking down simultaneously everywhere, in order to meaningfully apply local analyses to the singularity problem. The result is also described as canonical in the latest edition of the text [14] and by Belinskii [15]. The literature thus gives the impression that BKL still believe this construction to be valid even though they no longer believe their original arguments against the existence of physical singularities.

Comments. It is with regard to the above cited assertions that the application of global techniques to BKL's results become particularly interesting since this work is still claimed to be general [15]. We now prove two theorems which demonstrate, that: (i) In a universe with non-compact space sections the simultaneous fictitious singularity BKL claim to have constructed can only occur on a hypersurface which is not a global Cauchy surface; thus *four arbitrary functions of three variables are insufficient to establish generality*; (ii) In the compact case we prove an even stronger result; that it is simply *impossible for every geodesic of a synchronous coordinate system to terminate in a fictitious singularity.*

Theorem 2. Let S be a spacelike boundaryless C^2 hypersurface of a space-time (M, g). Suppose that the synchronous coordinate system generated by the timelike geodesics orthogonal to S terminates in a simultaneous fictitious singularity. Then S is non-compact and furthermore S is *not* a Cauchy surface for (M, g).

Proof. We first note that S can be assumed to be a partial Cauchy surface, by the remarks of HE [61] p. 273. To fix ideas we assume that the simultaneous fictitious singularity occurs to the future of S at the time t_0 as measured in the synchronous coordinate system. Since a given geodesic generator γ of the synchronous coordinate system encounters a fictitious and not a physical singularity at the point p at proper time t_0, and since (M, g) has no boundary, it follows that the geodesic segment γ can be extended beyond the point p. However, by the corollary on page 217 of HE [61], to each point $q \in D^+(S)$ there is a future-directed timelike geodesic curve orthogonal to S from S to q which does not contain any point conjugate to S between S and q. Since there is a point conjugate to S along *every* timelike geodesic orthogonal to S at proper time t_0, it follows that no point in the extension of γ beyond p can lie in $D^+(S)$. If S were a Cauchy surface, then *all* points to the future of S would have to lie in $D^+(S)$. Thus S is not a Cauchy surface.

To show that S is non-compact, we proceed as in the proof of theorem 4 of HE [61] (p. 273). The region of space-time $S \otimes [0, t_0]$, which is swept out by the synchronous coordinate system would be compact if S were compact. By the argument above, $S \otimes [0, t_0]$ contains $\bar{D}^+(S)$, (and hence $D^+(S)$ and $H^+(S)$). Thus $H^+(S)$ and $\bar{D}^+(S)$ would be compact. Consider a point $q \in H^+(S)$. Since every past-directed non-spacelike curve from q to S would consist of a (possibility zero) null geodesic segment in $H^+(S)$ and then a non-spacelike curve in $D^+(S)$, it follows that $d(S, q)$ would be less than or equal to t_0. Thus, as d is lower semi-continuous, one could find an infinite sequence of points $r_n \in D^+(S)$ converging to q such that $d(S, r_n)$ converged to $d(S, q)$. To each r_n there will be a segment of a timelike geodesic generator of the synchronous coordinate system. Since $S \otimes [0, t_0]$ is compact there is a segment of a timelike geodesic generator of the synchronous coordinate system which is a limit curve of the segments defined by the r_n sequence. By continuity this geodesic has length $d(S, q)$ and future endpoint on q. Thus to every point $q \in H^+(S)$ there would be a segment of a timelike geodesic generator γ_q of the synchronous coordinate system from S to q, with γ_q having length $d(S, q)$. Now let $q_1 \in H^+(S)$ lie to the past of q on the same null geodesic generator λ of $H^+(S)$. Joining the geodesic of length $d(S, q_1)$ from S to q_1 to the segment of λ between q_1 and q, one would obtain a non-spacelike curve of length $d(s, q_1)$ from S to q which could be varied to give a longer curve between these endpoints (proposition 4.5.10, HE [61]). Thus $d(S, q)$, $q \in H^+(S)$, would strictly decrease along every past-directed generator of $H^+(S)$. By proposition 6.5.2 [61], such generators could have no past endpoints, and so $d(S, q)$ cannot have a minimum. However, as $d(S, q)$ is lower semi-continuous in q, it must have a minimum on the compact set $H^+(S)$. Thus if S is compact, it is not possible for *all* synchronous coordinate system generators to terminate in a fictitious singularity. Q.E.D.

If a space-time (M, g) contains a non-compact Cauchy surface then it may be possible to find a *partial* Cauchy surface such that the synchronous coordinate system orthogonal to this hypersurface terminates in a simultaneous fictitious singularity. For example, consider the partial Cauchy surface S defined by: $x_4 < 0$; $x_1^2 + x_2^2 + x_3^2 - x_4^2 = -1$ in Minkowski space [61]; for this hyperboloid $\chi_a^a = -3$ at all points, see fig. 3.

This partial Cauchy surface is a Cauchy surface for the past light cone of the origin of coordinates in Minkowski space. All timelike geodesics normal to this partial Cauchy surface terminate in a simultaneous fictitious singularity at the origin. Although S is a Cauchy surface for the past light cone, it is not a Cauchy surface for the entire Minkowski space.

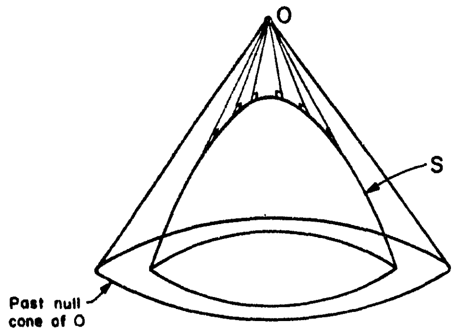

Fig. 3. S is a spacelike hypersurface in Minkowski space. S is a Cauchy surface for $I^-(0)$, but only a partial Cauchy surface for the entire space-time. The origin O is a simultaneous fictitious singularity in the synchronous coordinate system generated by the timelike geodesics orthogonal to S.

Theorem 3. Let S be a compact spacelike boundaryless C^2 hypersurface of a space-time (M, g). Then it is *not* possible for *every* timelike geodesic generator of the synchronous coordinate system formed by the timelike geodesics orthogonal to S to terminate in a fictitious singularity to the future of S. (A similar theorem can be proven with 'future' replaced by 'past'.)

Proof. Suppose on the contrary, that all timelike geodesic generators of the synchronous coordinate system do terminate in a fictitious singularity to the future of S. By definition of fictitious singularity, this means that every timelike geodesic γ orthogonal to S can be extended into the future until it reaches a point $p \in \gamma$ which is conjugate to S. By proposition 7.24 of Penrose [72], the location of the first conjugate point p to S varies continuously with γ, and thus the proper time distance t to the first conjugate point to S along the timelike geodesics orthogonal to S is a continuous function of S. Since S is compact and $t(S)$ is continuous, we can renormalize proper time along the timelike geodesics normal to S so that $t(S) = t_0$. Let $\beta: S \otimes [0, t_0] \to M$ be the differentiable map which takes a point $p \in S$ a distance $s \in [0, t_0]$ up the future-directed geodesic through p orthogonal to S. Then β $(S \otimes [0, t_0])$ would be compact and would contain $\bar{D}^+(S)$, for reasons given in theorem 2; the remainder of the proof is the same as the latter part of the proof of theorem 2 above. Q.E.D.

A slight modification of the proofs of theorems 1 and 2 will show that if the geodesics normal to a spacelike Cauchy surface S terminate in a simultaneous singularity, then for *all* the geodesics this singularity must be physical, not fictitious. If the hypersurface S is just a partial Cauchy surface, or if S is a hypersurface with boundary, then it can be shown that if some of the geodesics terminating in a future simultaneous singularity in fact terminate in fictitious singularities, then the fictitious singularities are in $J^+(H^+(S))$, and hence the structure (and even the existence) of these singularities is not determined by the four functions given on S.

Finally we notice, as an aside, that in later work the desire to construct a simultaneous *physical* singularity possibly springs from these claims regarding the existence of a general simultaneous *fictitious* singularity. It seems as though whenever in their non-generic solutions BKL obtained a physical singularity at the point where the synchronous coordinate system broke down, it was claimed in the generic case this physical singularity would be converted into a fictitious singularity. After the singularity theorems were proven they abandoned this claim. However, the methodology of using a synchronous reference system which could be continued all the way into a physical singularity was retained.

4. Post-Penrose work of BKL

Following acceptance of the singularity theorems [73] BKL concentrated upon constructing the general solution to Einstein's equations near a physical spacelike singularity. This work appears in the four papers [10–13] and is briefly described in the textbook [14]. (Note we are not here discussing any of the work on *homogeneous models* [12, 14], which is of a different nature.) Again we divide our discussion as follows: (1) *physical singularities,* (2) *the 'general solution'*, (3) *methodology*, (4) *future singularities.*

4.1. Physical singularities

The Kompanyeets metric.

Resumé. The aim of this work was to [87] "investigate a general solution containing a *simultaneous physical singularity in time*". It was argued that a metric which was first studied by Kompanyeets [66] is sufficiently arbitrary to represent an interval of the general solution corresponding to a period of Kasner-like evolution. It is shown how such a metric admits the homogeneous Bianchi VIII and IX metrics as particular specializations and thus it is claimed that the oscillatory structure of these homogeneous models provides a behavioural paradigm for the general case. An approximation technique relying on local Fourier series expansions is employed. Continuing this analysis [12], it was argued in [13] that the interchange of Kasner periods takes place in a quantitatively similar fashion to the Taub–Bianchi II [26] bounce law which describes a single oscillation in the 'Mixmaster-like' cases of homogeneous Bianchi types VIII and IX. The single structure constant governing the bounce-inducing, curvature anisotropy is assumed to take on a spatial variation in general, as a manifestation of the inhomogeneity. It is also felt that [88] "The generality of the oscillatory solution gives one a basis for assuming that the singularity reached by a finite body in its collapse below the event horizon in the comoving reference frame has this same character".

Comments. As quoted above, BKL claim to have constructed a simultaneous physical singularity in space. The following theorem proves that, for black hole singularities in asymptotically flat space, such a construction can only occur on a hypersurface which is not a global Cauchy surface, and thus *more* than the *four* arbitrary functions BKL display would be necessary to characterize the entire space-time.

Theorem 4. Let (M, g) be a weakly asymptotically simple and empty space-time, and let S be a spacelike C^2 hypersurface. If every timelike geodesic orthogonal to S is incomplete to the future of S, with every such geodesic having the *same* proper time length in $J^+(S)$, then S is *not* a Cauchy surface for (M, g).

Proof. If S were a Cauchy surface for (M, g), then $\mathscr{I}^+ \subset \overline{D^+(S, \bar{M})}$, where closure is taken in the space $\bar{M} = M \cup \partial M$ (see HE [61] p. 222 for notation). If t_0 is the limit to the proper time length of the timelike geodesic orthogonal to S, we can find a timelike curve greater than t_0 by choosing the future endpoint of γ sufficiently close to $\mathscr{I}^+$. But, this contradicts the corollary on p. 217 of HE [61] and the assumption that S is a Cauchy surface which implies $\gamma \subset D^+(S)$. Q.E.D.

We can extend this theorem with the following result which applies to open universes containing a black hole.

Theorem 5. Let (M, g) be a space-time which contains a future incomplete timelike geodesic and a future complete timelike geodesic. Then there does *not* exist a C^2 spacelike hypersurface S such that

(i) Every timelike geodesic orthogonal to S has the *same* proper time length in $J^+(S)$;

(ii) S is a Cauchy surface.

Proof. This is an obvious modification of the proof of theorem 3.

For the case of a past cosmological singularity, it is also true that even if the space-time is globally hyperbolic, it is possible that a hypersurface with a synchronous coordinate system for which the singularity is simultaneous might not be a Cauchy surface for the whole space-time. For example, in the open Friedmann universe we can pick a single point on the singularity and consider a congruence of timelike geodesics emerging from this point such that this congruence fills the future light cone of the point, as in fig. 4.

The t = constant surfaces of this congruence are partial Cauchy surfaces and in this synchronous coordinate system the singularity occurs simultaneously at $t = 0$. However, the t = constant surfaces are not Cauchy surfaces for the entire space-time and the coordinate system allows us to analyse *only* the structure of a single point on the singularity boundary (C-boundary; see ref. [107] for more details).

The above example shows that even when a Cauchy surface is present the BKL construction may allow one to see only a minute part of the singularity structure. The following example [58] shows that if *no* Cauchy surface exists then the BKL construction could omit the most interesting portions of the

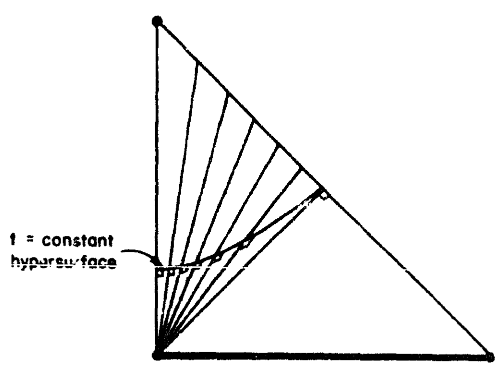

Fig. 4. A synchronous coordinate system with a simultaneous past physical singularity in an open Friedman universe. The t = constant surfaces provide a coordinate system allowing us to analyse the structure of only a single point on the singularity boundary (C-boundary).

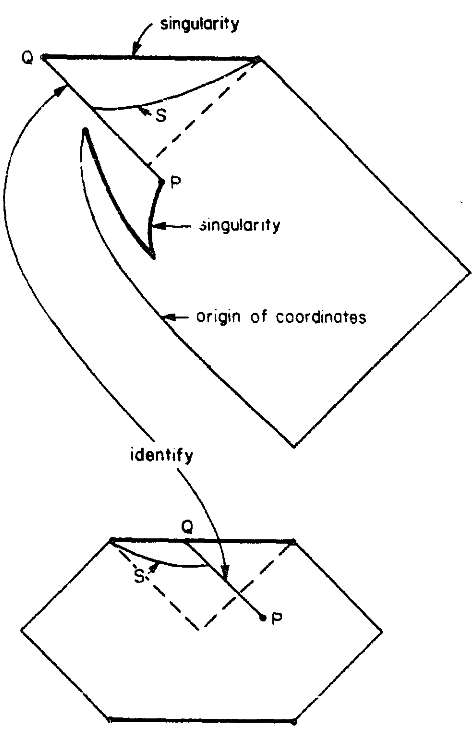

Fig. 5. Penrose diagram for a spherically symmetric gravitational collapse configuration which violates cosmic censorship. This diagram is adapted from Yodzis et al. [58]. The surface QP in both diagrams is identified. The synchronous coordinate system generated from S has a simultaneous physical singularity to the future of S. The entire Cauchy development of S does not include the interesting portion of the singularity.

singularity, i.e. the timelike portions. In fig. 5, the partial Cauchy surface S has a simultaneous singularity in $D^+(S)$ and the BKL procedure tells us that this singularity has the same structure as the Schwarzschild singularity. However, as is obvious from the figure the singularity structure of the entire space-time is radically different from Schwarzschild. Note that D(S) intersects *no* portion of the singularity; initial data on S cannot even *hint* at the existence of the timelike singularity structure.

The above examples apply to open cosmological and black hole type singularities respectively. More generally, we can show that for *all* physical singularities the simultaneous physical singularity is in fact not a stable property of any hypersurface in any space-time with particle horizons.

Theorem 6. Let S be a C^2 spacelike, boundaryless hypersurface in a space-time (M, g). Suppose S is a partial Cauchy surface with the property that every timelike geodesic orthogonal to S terminates in a physical singularity in $J^-(S)$ at the same proper time t. Then, this property is not a stable property (in either the D-stable or past D-stable topology) of any initial data set, $(h_{ab}, \chi_{ab}, \psi_{(i)})$, such that for the set:

(1) There exists an open neighbourhood U in S with $\bar{U}$ compact such that at least one inextendible past-directed timelike geodesic α orthogonal to S remains in int $D^-(U)$.

(2) There exists at least one inextendible past-directed timelike geodesic γ orthogonal to S satisfying $\gamma \cap J^-(\bar{U}) = \emptyset$.

Proof. We can change the initial data infinitesimally (in either the D-stable or past D-stable topology) so that the new initial data on S is the same as would be obtained by deforming $S \cap U$ into the past, keeping α orthogonal to the deformation and not changing S outside of U. This will change the time needed for α to reach the singularity, but by (2) it will not change the length of γ. Thus in the new initial data, the singularity no longer occurs at the same proper time t (see fig. 6). Q.E.D.

It should be emphasized that the above proof of instability of the simultaneity of the singularity depends crucially on the topology chosen in initial data space. The topology used in the proof of theorem 6 is not the usual topology of true-degrees-of-freedom space [76]. In particular the theorem explicitly assumes that a particular hypersurface is chosen upon which to impose initial data; we have *not* been able to prove that when the true degrees of freedom are varied, a synchronous coordinate system with a simultaneous physical singularity ceases to exist. What *could* happen is that when the true degrees of freedom are varied, such a coordinate system continues to exist but its $t' =$ constant slices move away *globally* from the $t =$ constant slices in the original coordinate system; see fig. 7.

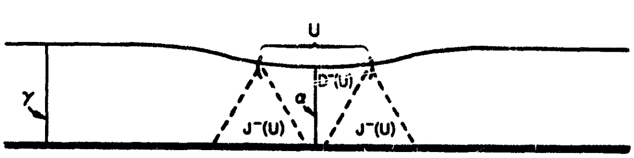

Fig. 6. Figure for the proof of theorem 6.

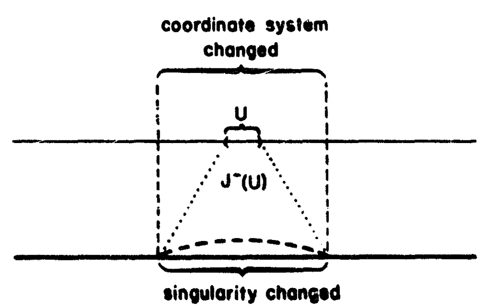

Fig. 7. The BKL "general" metric cannot be stable. If the initial data is changed in the small region U, tnis changes the time to the singularity in a much larger region. Since the singularity is required to be simultaneous, this necessarily changes the metric in an initial region larger than U. However, this necessary change does not occur when the BKL metric is perturbed.

This global movement *must* occur if a synchronous coordinate system with a simultaneous singularity continues to exist under variation of the true degrees of freedom, since we expect that a small, localized (in U) variation would cause a change in the time to the singularity (as measured in the original coordinate system in the region outside $J^-(U)$). If the singularity is to remain simultaneous, the coordinate system must therefore change *globally*; in particular, the relationship between the induced metric in the t = constant slices and the extrinsic curvature must change in a large region when the functions defining the true degrees of freedom are changed locally in U.

This does not occur in the BKL "general" metric, which has the form [73]

$$ds^2 = dt^2 - (a^2 l_\alpha l_\beta + b^2 m_\alpha m_\beta + c^2 n_\alpha n_\beta)\, dx^\alpha\, dx^\beta \tag{1}$$

$$a = t^{p_1},\ b = t^{p_2},\ c = t^{p_3}$$

with

$$p_1 + p_2 + p_3 = p_1^2 + p_2^2 + p_3^2 = 1. \tag{2}$$

Vectors l_α, m_α, n_α together with the scalars p_1, p_2, p_3 are functions of the space coordinates *only*. Clearly a small, localized change in these vectors and scalars (subject to the constraints (2) and the field equations) will not cause a *global* change in the metric (1). A similar argument applies even if we only wish to claim validity for (1) inside a small neighbourhood S (see fig. 1), whose past Cauchy development includes part of the past singularity. We can then choose the small, localized region U to be a small region of S_1. Thus the BKL metric does *not* possess the necessary properties of a *stable* metric expressed in a synchronous coordinate system with a simultaneous physical singularity.

We have not been able to prove or disprove the existence in general of a synchronous coordinate system with a simultaneous past singularity. However, we have been able to prove that *if* such a coordinate system exists in a closed universe, then it is unique.

Proposition 1. Suppose (M, g) is globally hyperbolic, and suppose there exists a compact C^2 boundaryless spacelike hypersurface S such that the past-directed timelike geodesics orthogonal to S form a synchronous coordinate system every timelike geodesic generator of which terminates in a physical singularity at the *same* proper time t_0. Then, this synchronous coordinate system is unique; there is no synchronous coordinate system in (M, g) covering all or part of (M, g) which has compact spacelike sections and which terminates in a simultaneous past physical singularity.

Proof. Suppose on the contrary that another such synchronous coordinate system existed. Let S′ be a t' = constant spacelike section of this other coordinate system with $S' \subset D^-(S)$. (Primes will denote quantities associated with the other synchronous coordinate system. Unprimed quantities are quantities associated with the synchronous coordinate system based on S.) Since the two coordinate systems do not coincide, there exists a t = constant section S_1 of the unprimed system such that $I^+(S_1) \cap S'$ and $I^-(S_1) \cap S'$ are non-empty; see fig. 8. By the corollary on page 217 of HE [61], (and by the compactness of S′ and S_1 which implies S′ and S_1 and Cauchy surfaces; see [110]) there exists a non-zero length timelike geodesic segment τ_- of maximal length from S_1 to $D^-(S_1) \cap S'$, and a non-zero length timelike geodesic segment τ_+ of maximal length from S_1 to $D^+(S_1) \cap S'$. Since τ_- and τ_+ are of maximal length, they must be orthogonal to both S_1 and S′ [104]. That is, both τ_1 and τ_+ are segments of timelike geodesic generators of *both* synchronous coordinate, systems. But the time to the singularity from S′ is t + (length of τ_+) as measured along the primed generator of which τ_+ is a part, and t − (length of τ_1) as measured along the primed generator of which τ_- is a part. This contradicts the fact that the time is t' along *both* generators. Q.E.D.

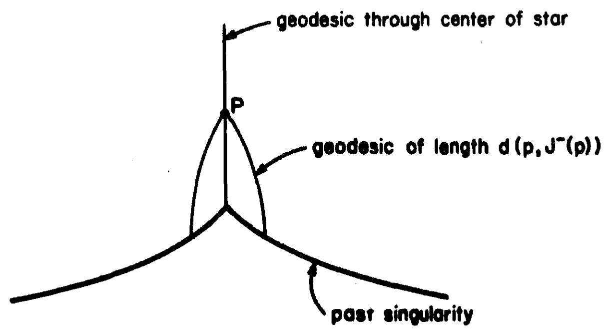

Fig. 9. Penrose diagram for a spherically symmetric space-time containing a "cusp" in the past singularity. In this space-time there is no synchronous coordinate system with simultaneous past singularity in any neighbourhood of the entire singularity. The Eardley foliation is only C^{1-}.

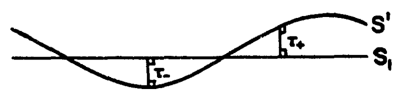

Fig. 8. Figure for the proof of proposition 1.

Since such a coordinate system is unique (if it exists) we might expect it to coincide with other unique, geometrically defined coordinate systems. For example, in the homogeneous 'Mixmaster' universe, the synchronous coordinate system with a simultaneous physical singularity *does* exist and the t = constant hypersurfaces coincide with the χ^a_a = constant hypersurfaces. Now Goddard has shown [105] that if a χ^a_a = constant foliation of space-time exists, then this foliation is unique. Marsden and Tipler have shown [100] that generic Wheeler universes – globally hyperbolic closed universes with all-encompassing past and future singularity; see ref. [104] for a precise definition – whose singularities are of the strong curvature type [101], probably possess a foliation of χ^a_a = constant hypersurfaces. Since all physically realistic globally hyperbolic closed universes are expected to be of the Wheeler type, and since generic singularities are expected to be of the strong curvature type (the Schwarzschild, Friedmann, and even the BKL "generic" singularity is of this type), it follows that all physically realistic globally hyperbolic closed universes can probably be foliated by χ^a_a = constant Cauchy surfaces.

However, the hypersurfaces of this foliation will not in general coincide with the t = constant hypersurfaces of a synchronous coordinate system with a simultaneous past singularity (assuming such a system exists). From the Landau–Komar–Raychaudhuri equation, it follows that a necessary condition for these two sets of hypersurfaces to coincide is $R_{ab}N^aN^b + K^{ab}K_{ab}$ = constant over each hypersurface, where N^a is the unit normal to the hypersurface. The condition holds in homogeneous universes, but it clearly would not hold in general.

Although we have not been able to prove it rigorously, we believe that a synchronous coordinate system with a simultaneous past singularity will *not* exist in stable, inhomogeneous space-times with an all-encompassing past singularity no matter how close one goes to the singularity. Eardley [60] has attempted to construct such a system by taking level surfaces of the function $d(\mathrm{p}, \mathrm{J}^-(\mathrm{p}))$, for all points $\mathrm{p} \in M$. (For a precise definition of the function d see HE [61] p. 215.) If $d(\mathrm{p}, \mathrm{J}^-(\mathrm{p}))$, is finite for all $\mathrm{p} \in M$ (i.e., if the past singularity is all-encompassing) then this construction will exist, will be unique, and will probably coincide with a synchronous coordinate system having a simultaneous past singularity, if such a system exists. However, the Eardley construction will in general give only C^{1-} hypersurfaces, not the C^2 hypersurfaces which are needed to define a synchronous coordinate system.

To see why in general the Eardley construction will not generate a synchronous coordinate system, see fig. 9. This figure is a Penrose diagram for a spherically symmetric but inhomogeneous space-time with an all-encompassing past singularity – one can regard it as a cosmology which has a homogeneous matter distribution except for a single, spherically symmetric star. The star causes a "cusp" in the past singularity, the presence of the cusp being detected by the fact that the timelike geodesics of length $d(\mathrm{p}, \mathrm{J}^-(\mathrm{p}))$ for all events p in the centre of the star, lie not through the centre of the star, but a little to the side. Since the space-time is spherically symmetric, there will be infinitely many timelike geodesics of length $d(\mathrm{p}, \mathrm{J}^-(\mathrm{p}))$ from the singularity to the point p, one geodesic through each point of a sphere surrounding p. However, if the geodesics of length $d(\mathrm{p}, \mathrm{J}^-(\mathrm{p}))$, for all p sufficiently close to the singularity, are to define a synchronous coordinate system, there must be a *unique* geodesic from the singularity to the point p. Since approach to the singularity would in general *increase* inhomogeneities rather than smooth them out, we would expect such cusps to occur in generic singularities, causing the geodesics of length $d(\mathrm{p}, \mathrm{J}^-(\mathrm{p}))$ to be non-unique. (The absence of spherical symmetry would not necessarily invalidate this argument.)

Misner [25] has given essentially the same argument for doubting "... that synchronous coordinate systems can be chosen in such a way that they reach the singular points of the solution." Misner pointed out that a synchronous coordinate system would survive for only about the free-fall time through an inhomogeneity of space-time; for a synchronous coordinate system covering the sun this is about an hour, not the 10^{10} years required if the coordinate system were capable of extension to the initial singularity. He went on to emphasize: "It does not help to choose the initial hypersurface of a synchronous coordinate system closer to the singularity, since the generic solution which resembles our Universe now will, if extrapolated back toward the singularity, be even more irregular near the singularity, so synchronous coordinates will then extend an even smaller fraction of the required distance to the singularity than now". Misner directed the above argument against the BKL pre-Penrose work, but it applies even more forcibly against the BKL post-Penrose work.

It should be noted that the above arguments do not rule out the *local* existence of a synchronous coordinate system with a simultaneous singularity – any geodesic γ of length $d(\mathrm{p}, \mathrm{J}^-(\mathrm{p}))$ from p into the singularity has a geodesic congruence in an *infinitesimal* neighborhood of γ such that the congruence defines a synchronous coordinate system with a simultaneous physical singularity. In fact, it is the existence of such a *local* congruence that allowed Tipler [101] to obtain limits to the rate of growth of curvature near *generic* singularities. The above arguments do, however, strongly indicate that such a synchronous coordinate system is unlikely to exist over a region of sufficient size to give any *global* information about the singularity.

4.2. The general solution

Resumé. The results described in subsection 4.1 above [89] led to a claim, "that the general solution describing one era is close to the solutions for the homogeneous models of types VIII and IX depending on whether the function $\rho(x, y)$ vanishes or not ... The type of model approximating the structure of space can change from point to point". They claim their analysis reveals no essential structural differences between compact and non-compact universes and [10], "so far it can be stated only that there is no direct connection with the finite or infinite nature of the space ...".

Comments. BKL claim that roughly the general behaviour appears 'Bianchi type VIII or IX-like' locally according as a certain arbitrary two-dimensional function $\rho(x, y)$ is zero or non-zero respectively. However, since the type IX-like case has compact topology the arbitrary functions specifying it should be globally periodic. No such restriction on the behaviour of the arbitrary functions in these two distinct cases arises in their calculations [12]. It also appears ad hoc that the homogeneous Bianchi VIII and IX models are chosen as paradigms for the general case. Although they are indeed specified by four arbitrary *constants* on a $t=$ constant hypersurface so are the Bianchi VI_h and VII_h models [92]. BKL feel [12, 14] that only the VIII and IX models exhibit an oscillatory phase yet recent work using modern techniques in the qualitative theory of autonomous differential systems by Peresetskii [91] claims that in general VI_h and VII_h metrics also exhibit an oscillatory phase in the presence of matter motion relative to the synchronous reference system. Thus there appear no grounds for picking either VIII or XI models as generic amongst the homogeneous models.

Furthermore, the following theorem shows that a stable solution which is parameterized by four arbitrary functions of three variables cannot exist close to *any* closed spatially homogeneous universe. In particular, contrary to the above-mentioned claims of BKL, such a stable solution cannot admit a spatially homogeneous closed universe as a special case.

Theorem 7. Suppose (h_{ab}, χ_{ab}) is a point in the true-degrees-of-freedom space $\mathscr{C}$ of the empty-space Einstein equations. If (h_{ab}, χ_{ab}) is given on a *compact* spacelike hypersurface S and has Killing vectors, then it is *not* possible to parameterize *any* neighborhood of (h_{ab}, χ_{ab}) in $\mathscr{C}$ with four functions of three variables. That is, there does not exist a diffeomorphism from a neighborhood of (h_{ab}, χ_{ab}) in $\mathscr{C}$ onto a neighborhood of the W^s $(s \geq 3)$ Hilbert space $\mathscr{H}$ of four functions of three variables. Furthermore, this result holds *locally* on S; that is, it also holds for a solution (h_{ab}, χ_{ab}) given just on a neighborhood U of S, provided (h_{ab}, χ_{ab}) given on U can be extended to a solution on the whole of S.

Proof. (The above theorem, its proof, and its significance were worked out jointly by the authors and Professor Jerrold E. Marsden. The authors would like to thank Professor Marsden for his help.) Fischer, Marsden, and Moncrief have shown [77] that if a solution (h_{ab}, χ_{ab}) on a compact S to the empty-space Einstein equations has Killing vectors, then the solution space $\mathscr{C}$ is not differentiable (in the Frechet sense) at the point (h_{ab}, χ_{ab}). This rules out the possibility of a diffeomorphism from $\mathscr{C}$ to $\mathscr{H}$ in a neighborhood of (h_{ab}, χ_{ab}). Furthermore, the Fischer–Marsden–Moncrief result holds *locally* on S. Q.E.D.

As discussed in subsection 2.4 on linearization stability, what happens near solutions with Killing vectors is that the four functions parameterize non-solutions as well as solutions. At (h_{ab}, χ_{ab}), the non-solutions are tangent to the spurious perturbations of the linearized equations. Since the non-spurious perturbations must satisfy a constraint which is not imposed on the spurious perturbations, it follows that these spurious perturbations are open dense in the space of solutions to the linearized equations. Thus the set of tangent directions of the non-solutions parameterized by the four functions is open dense at (h_{ab}, χ_{ab}). Thus in this sense the true solutions near (h_{ab}, χ_{ab}) are a set of "measure zero" in the set of all "solutions" parameterized by the four functions. More precisely, in the language of subsection 2.1 it is *generic* for a metric such as the BKL metric (1) *not* to satisfy the Einstein equations near (h_{ab}, χ_{ab}). This shows that the BKL power series *almost everywhere* does *not* converge to a true solution near (h_{ab}, χ_{ab}).

As stated in theorem 6, these results hold even if (h_{ab}, χ_{ab}) is given only on a neighborhood U of S, provided the solution on a neighborhood can be extended to a solution of the whole of S. This assertion actually does not have too much content since the initial value equations are elliptic and hence would probably give a unique extension from U to the whole of S. However, it does point out certain defects in a method which attempts to piece together a global solution on S from solutions in small neighborhoods: if the method patches together approximately homogeneous neighborhoods, then the conclusions of theorem 6 are avoided only if the errors made in the patching procedure strongly influence the evolution of the initial data (see also subsections 4.3 and 5.3 on this point).

The requirement that the map from $\mathscr{C}$ to $\mathscr{H}$ be a diffeomorphism and that $\mathscr{H}$ be a W^s space with $s \geq 3$ is a necessary requirement if one assumes that the metric tensor varies differentiably as one varies the four functions, and this differentiability is necessary if one wants to do analysis on the space of solutions. It can be shown, however, that a local *homeomorphism* from $\mathscr{C}$ to $\mathscr{H}$ always exists (M.W. Hirsch, private communication).

The question of whether theorem 6 holds for non-compact S is unsolved in general. It is known [77] that if S is asymptotically flat, then the theorem is false; there does exist a local diffeomorphism from $\mathscr{C}$ to $\mathscr{H}$ near a solution with Killing vectors. Whether or not the structure of $\mathscr{C}$ near a non-compact homogeneous solution resembles the compact case is not known.

4.3. Mathematical methodology

Resumé. BKL claim to be using a local approximation and construction technique so that [89], "the general solution obtained is valid only in a certain limited interval of ξ (time). We can therefore always separate in space a region S_1 bounded in z at the initial instant ξ_1, and a region S_2 at the instant ξ_2, such that the initial perturbations from the regions outside S_1 do not have time to influence the character of the solution in S_2 at the instant ξ_2 (see fig. 1). Thus, if we use a Fourier expansion in the region S_1, then it can be assumed that the solution obtained in this manner will actually be general for all instants of time in the interval $\xi_1 - \xi_2$ and in a certain bounded region of z, the dimensions of which are determined by those of S_2 at the final instant ξ_2 of the era". By further considerations of their local approximations in different asymptotic regions of a Kasner period they go further and [93] "conclude that the subsequent behaviour of the solution in the general case (after going outside the region of applicability of the approximation constructed by us) will be qualitatively the same as in the particular case of the metric (of type IX)". Concerning the validity of the series expansions, Belinskii claims [15], "that in physically significant situations convergence or the asymptotic property of the series is always guaranteed to that extent which is required. I think that the convergence requirement could at worst only impose some weak restrictions, such as inequalities, on the arbitrary parameters which appear in our solution".

Comments. On some occasions BKL claim to be deriving a local description of the singularity structure. In subsection 3.1 above we displayed how the knowledge of four function data on small neighborhoods is insufficient to derive the structure of the entire singularity unless the solutions are either essentially generic, or stable in a large, finite neighborhood of solution space. Since BKL have obtained only stable solutions – i.e. solutions whose basic structure does not change when the data is perturbed infinitesimally – their techniques cannot give a complete description of the singularity structure; see subsection 3.1 and fig. 1.

In the construction of the oscillatory behaviour over long time intervals claimed by BKL, the number of required coordinate transformations becomes infinite as the singularity is approached. (Increasingly more frequent transformations to smaller regions being required to maintain the approximation.) In essence this procedure is the same as approximating the sphere by a sequence of planes and concluding that the earth is flat. What is happening in both cases is that each time a coordinate transformation is made, an approximation is also made and this approximation introduces de facto new initial data into the solution. These new initial data change the time evolution of the solution and invalidate the function counting argument.

4.4. Future singularities

Resumé. Regarding the existence of singularities in the future for the general case it is claimed that [88], "a singularity in the future can have physical meaning only if it is reachable for arbitrary initial conditions assigned at some preceding moment of time. Clearly there is no reason why the distribution of matter and field that was reached at some moment in the process of evolution of the universe should have corresponded to the precise conditions required for the appearance of some particular solution of the Einstein equations", it is claimed that [94] "the existence of a general solution possessing a singularity does not therefore preclude the existence of other general solutions that do not have a singularity. For example, there is no reason to doubt the existence of a general solution, without singularities, that describes a stable isolated body with not too large a mass". These views were also reiterated by Belinskii recently [15].

Comment. The following theorem of Hawking shows that for certain sets of initial data the existence of a future cosmological singularity in a closed universe is a stable property of the initial data. The work of subsection 2.2 indicates (but does not rigorously prove) that there is an equivalence between the BKL and Penrose notions of singularity.

Theorem (HE [61] p. 272). Space-time is not future timelike geodesically complete if
(i) $R_{ab}V^a V^b \geq 0$ for every non-spacelike vector V^a.
(ii) There exists a compact spacelike three-surface S (without edge).
(iii) The unit normals to S are everywhere converging on S.
This theorem established that singularities occur in the future for a set of Cauchy data which is not of measure zero, for condition (iii) is an open condition. One can vary the normals slightly and they would still be everywhere expanding.

5. Further considerations

In the previous sections 3 and 4 we have displayed a number of comments concerning limitations inherent to the BKL methodology as applied to the generic behaviour of solutions in cosmology and gravitational collapse near their respective singularities. We shall now turn to consider some additional, physical questions that must be faced in any description of the generic cosmological solution to the Einstein equations.

5.1. The definition of inhomogeneity

The property of spatial (in)homogeneity is not an unambiguous notion. Early work [27] and virtually all recent work on physical or 'applied' cosmology has concentrated upon the Newtonian concept of 'differential homogeneity' [27, 31], defined by the absence of spatial gradients in all scalar, vector and tensor components of physical quantities. MacCallum [28] has argued very convincingly that such a definition is both too restrictive and intuitively unreasonable when compared with one obtained by using invariance properties under a 3-parameter group of motions. Such an explicitly coordinate independent criterion generates the Bianchi classification of homogeneous spaces [26, 29, 30]. The evaluation of inhomogeneities solely by examining the spatial variation in metric and other physical quantities is clearly not invariant. For, in this approach the 'homogeneity' is determined by an observer inferring equal values of a physical variable at different places *simultaneously*. However, special relativity shows such a simultaneity to be dependent upon the observer's relative motion. Thus the perceived presence and nature of the spatial gradient in a physically varying quantity depends upon the velocity four vector of the flow lines of the matter. In the cosmological context it is found that many Bianchi (group) homogeneous universes contain spatial gradients in physically observable quantities [31, 96]. Often BKL seem to use these two notions of homogeneity interchangeably and without distinction making it difficult to detect purely coordinate induced irregularities.

5.2. *Types of inhomogeneity*

The influence of inhomogeneous structure and motion in general relativity is expected to be strong in the light of the remarkable self-interacting non-linearity and hyperbolic evolution structure of the gravitational field equations. However, BKL restrict their discussion to that which we label 'passive inhomogeneity', wherein the irregularity allowed is of an entirely algebraic nature. By this we mean that the spatial variation appears decoupled from the temporal evolution and is manifested principally via integration constants. Having appeared in the constraint equations to determine an initial configuration it plays no dynamic rôle in the evolution equations. Thereby, the irregularities are 'homogeneously propagated'. Such a notion can be made precise using a Hamiltonian representation [e.g. 32]. Roughly, this means that the spatial derivatives do not occur in the evolution equations for the metric(1); see subsection 4.1. We feel that these restrictions are unnatural since the space and time dimensions appear on an equal footing within the hyperbolic equations of the theory. Consideration only of such passive inhomogeneities, although necessary for tractability, also sufficiently dilutes the notion of inhomogeneity so that description of many interesting and dynamically important features could be eliminated by hypothesis. The absence of such features as black hole formation, non-cosmological trapped surfaces [106], strong gravitational wave interaction together with their related phenomena, seems intuitively unappealing and can probably be associated with the non-general nature of their 'general' solutions revealed in sections 3 and 4. We have seen above how local techniques may inevitably reveal a homogeneous structure and we shall also see below how this passive inhomogeneity might lead to the erroneous physical assumption of an approximate superposition principle.

Our experiences with other non-linear field theories have taught us to expect new features peculiar to inhomogeneous situations [34, 35]. This leads us to pose the following problem: Do all stable inhomogeneous cosmological space-times have homogeneous specializations, and are all spatially homogeneous universes linearization stable? An exposition of this exceedingly subtle mathematical notion is given in subsection 2.4 where we explain why the homogeneous (Bianchi) models contain 'spurious perturbations' in their perturbation spectra and thus why further checks are necessary to ensure that perturbations remain in the vicinity of a space-time.

5.3. *Curvature dominated singularities and coordinate coverings*

In this subsection we explain with examples what type of approximation procedures are justified physically and also give some examples of the vacuous nature of conclusions drawn from some local coordinate dependent methods. Firstly, we confine attention to the final paper of the 'general solution' programme [13], in which it is argued, for example, that the generic cosmology displays interchange of Kasner regimes in essentially identical fashion to the Taub–Bianchi II model. The local technique of analysis confines attention to some small coordinate patch from which it is deduced that there exist a succession of periods, $\Delta_i t$, during which the spatial curvature anisotropy is dynamically negligible, interspersed by briefer periods during which these curvature anisotropy effects change the expansion rates under a Kasner transformation. Thus, it is claimed that the mere spatial variation of the group structure constant determining the curvature anisotropy in the local region is sufficient to describe all the global consequences of this behaviour for the dynamics. This seems dubious and is an artefact of considering only the 'passive inhomogeneity' described above in subsection 5.2. If the space-time is generic one expects the spatial curvature anisotropy to vary strongly from region to region. In picking one small coordinate covered region to establish the existence of a series of intervals in which the curvature is unimportant BKL neglect to consider that the interval appropriate for one such region will in general differ from that during which others are free of curvature anisotropy effects. We expect the collective gravitational Coulomb effect of other regions to severely perturb the original neighborhood being examined because of this lack of synchronization. There is no evidence that a scenario in which different regions behave as independent Mixmaster-like models is stable. Indeed it has been shown by Barrow and Carr [36] that unless inhomogeneities are bound there are complex effects arising through the non-linear coupling of the density fluctuations with the background anisotropy which dominate the local dynamics. A more familiar example of the problem arises in the use of so called 'self-modelling' solutions in describing the formation of spherically symmetric protogalaxies embedded in a background Friedmann universe [37–40]. These analyses are in the spirit of the BKL method. By using them we gain information only of the evolution of one protogalactic clump relative to the background into the non-linear regime; but not of the mutual interaction between different regions. However, N-body simulations [41, 42] reveal that these latter effects are completely *dominant* in the non-linear regime. Another familiar example from non-linear physics is the enhanced efficiency of shockwave formation obtained via the mutual interaction of members of a wave ensemble relative to shock formation solely via the self-interaction of one wave.

We have shown in sections 3 and 4 how a local coordinate dependent technique fails to give a complete description of the global and local structure of a space-time in general. Here we point out some further ambiguities of the technique which make one suspect the method to be very hazardous. BKL's local approach is equivalent to producing a coordinate covering of the generic metric structure of space-time. They have claimed for example [12] that small deviations from homogeneous Bianchi type VIII and IX models constitute such a covering (see section 4). Although we have shown in sections 3 and 4 that the arguments leading to this conclusion are incorrect we also point out that even if such a statement could be established it would give us essentially no real information about the space-time structure at all. One can cover a space-time with a coordinate patching derived from *any* other space-time so long as these coordinate patches are sufficiently small. In particular there always exists a covering in which the 'general' solution appears approximately homogeneous on small enough coordinate neighborhoods; and indeed such a local approach applied to the surface of the earth once convinced us that the earth was flat. The following examples specifically illustrate the ambiguities of this procedure:

(i) *Tolman–Bondi metrics* [43–44]. Locally these look like dust-filled Friedmann models and a covering of the space-time by Friedmann patches can be generated [37]. However, the global and singularity structures of the two are markedly different.

(ii) *Szekeres metrics.* A local covering may be generated by Kantowski–Sachs [46] like patches, yet again the global structures are quite distinct.

(iii) *Einstein–Rosen–Gowdy metrics* [47, 48]. Liang [49] has approximately analysed the behaviour of these space-times when matter is present. Specifically, in the radiation filled case he displays two coverings of the space-time by Friedmann-like and Kasner-like patchings. They lead to quite different representations and dynamical descriptions, both apparently consistent with the approximation technique.

(iv) *Kompanyeets metric* [66]. The Kompanyeets metric was covered by a certain local covering to claim it was general in the post-Penrose work of BKL. However in 1965 Khalatnikov [4] produced another local covering based on local series expansions to claim that this metric did *not* represent the general solution. Thus one covering gave a metric containing two arbitrary functions and another four.

5.4. The vacuum assumption

It has long been stressed by BKL that in the vicinity of the cosmological singularity one may usually neglect the influence of the stress-energy terms with respect to the geometric terms in the Einstein equations [16]. This important idea motivated the later work of Eardley, Sachs and Liang [49–50] who developed the more rigorous notion of a velocity-dominated singularity. Although BKL's assumption of Weyl curvature domination is probably quite satisfactory for the high density regime near the singularity in the $p = 0$ or $p = \rho/3$ states required in the Hagedorn statistical bootstrap [51] theory or asymptotically free, coloured SU(3) invariant gauge theory (QCD) [52]; there is an important and physically relevant case in which it is not. Should the asymptotic behaviour of matter be that of the 'stiff' Zeldovich form, $p = \rho$, describing baryons interacting via a spin-1 vector meson field [53–56], then the structure qualitatively changes. BKL have pointed out that the generic solution can now be dominated by the matter-field distribution so the stress tensor strongly influences the singularity structure [94]. In particular, the oscillatory Mixmaster behaviour evident in the homogeneous vacuum Bianchi VIII and IX models near the singularity need not persist in a stiff matter era. The influence of this matter can be stronger than that of the spatial curvature anisotropy and perturbations strong enough to permute the expansion rates can be suppressed. In this case the expansion scales can all fall off monotonically on approach to the singularity. Such a situation is quite compelling on many other grounds and has been described in detail by Barrow [56].

It is also interesting to consider the global constraints placed upon cosmological models by the vacuum assumption. Do, for instance, all generic vacuum solutions have matter-filled analogues, and vice-versa? And, in what way does the presence of matter affect the global structure of cosmological space-times? Indeed this first feature of 'evacuation stability' occurs for the most general matter-filled cosmological space-times so far known, the Szekeres metrics [47–57]. These inhomogeneous dust-filled universes have no cosmological vacuum specialization, reducing to the Schwarzschild space-time when the matter content is removed. The example given in section 2 regarding the instability of the globally hyperbolic character of the Schwarzschild metric on addition of electric charge is also a good example of the sensitive balance between the global structure and the matter content of a space-time. We also note that there are no non-flat singularity-free asymptotically flat solutions to the *vacuum* Einstein equations [108], though there are such solutions if matter is present.

5.5. Entropy and generality

Here we draw attention to the argument of Barrow and Matzner [56, 59] emphasizing that the observation of the specific radiation entropy of the universe from the microwave background radiation, in conjunction with the second law of thermodynamics, strongly indicates that the universe has never been dominated by any form of irregularity energy near the past singularity. The present entropy level places a fundamental upper limit upon the amount of dissipation that can have occurred in the past and hence on the allowed deviations from the isotropic Friedmann metric. The presence of many highly efficient dissipative mechanisms near the singularity and the rapid divergence of the anisotropy energy, $\sigma^2 \propto R^{-6}$, with scale factor R, on approach to the singularity indicate that the earlier irregularities are dissipated the more entropy is produced. This argument encompasses both classical and quantum processes and argues strongly against our present universe being the remnant of a generic set of initial conditions. In particular, 'Mixmaster-like' oscillations and the accompanying level of anisotropy energy would be highly efficient generators of entropy near the singularity whenever the expansion is not dominated by the isotropic matter content. This is one indication of the importance of non-gravitational considerations in deciding the manner in which our universe is geometrodynamically non-generic*.

We may also make a speculative connection between these ideas and those of Penrose regarding the entropy of the gravitational field. Penrose has conjectured [97] that the Weyl curvature tensor measures the entropy of the space-time geometry in some sense. This would in fact *explain* the quiescent initial singularity required by the low radiation entropy of the universe since the initial big-bang singularity would correspond to a minimum entropy state of the free gravitational field with zero Weyl component of the curvature and thus to dynamics dominated by the Ricci tensor (low gravitational entropy state). This is sufficient to guarantee that the initial state be Friedmann-like and therefore not highly dissipative. If this conjecture were true, it would of course also indicate the singularity in the *future* of a closed universe to be of an essentially different dynamical character to that in the past since it must correspond to a state of high 'gravitational entropy' with the now dominant Weyl tensor terms in the curvature ensuring highly anisotropic and inhomogeneous dynamics in the final collapse. These considerations, if correct, thus relegate considerations of the generic dynamics of the Einstein equations to the *future* evolution of the universe and will be discussed in more detail by the authors elsewhere [98].

6. Conclusions

In the foregoing pages we have rigorously defined certain, previously vague, pieces of methodology essential to an analysis of the general solution to the Einstein equations near a cosmological singularity. We have used this machinery specifically to analyse in detail the methods and conclusions of the long programme of work by Belinskii, Lifschitz and Khalatnikov concerning the structure of the general solution to the Einstein equations near a cosmological singularity. Our principal conclusions may be summarized as follows:

(i) We give a detailed comparative discussion of various local and global notions of stability and generality. We show that the BKL approach examines only *stable* rather than *generic* solutions to the cosmological problem.

(ii) A rigorous formulation of the function counting criterion to establish the generality of space-times in the metaspace of all Lorentz metrics is given. In particular, the global properties required of a space-time in order to establish the generality of vacuum solutions by four arbitrary three-dimensional functions are given. The neglect of such considerations by BKL has led them to construct four-function

* The details of the arguments connecting the radiation entropy per baryon and the anisotropy energy must be modified if very close to the Planck time, baryon non-conserving interactions occur. Such interactions, for example, are anticipated in the latest grand unified gauge theory based on the SU(5) gauge group [95, 122, 123].

'solutions' which cannot be general. A discussion of generality in space-times with timelike singularities is given; no function counting criteria presently exists to establish generality in this case.

(iii) A discussion of the linearization stability of the Einstein equations is given and it is noted, in particular, that there exist 'spurious' perturbations of all the homogeneous Bianchi type universes and additional calculations must be performed before reliable information can be extracted from perturbation studies of Bianchi type VIII and IX metrics.

(iv) The probable equivalence between the BKL notion of 'singularity' involving infinities in the curvature invariants and that of Penrose specifying geodesic incompleteness is indicated for the cosmological problem. The ideas of fictitious and physical singularities are rigorously defined.

(v) The early 'pre-singularity theorem' work of BKL regarding the construction of general fictitious singularities was examined and theorems proved showing the results claimed by BKL using local series approximations to be incorrect. Specifically, contrary to their claims, global analysis reveals that it is impossible for every geodesic of a synchronous coordinate system to terminate at a fictitious singularity in a space-time with a compact hypersurface. In a space-time without a compact hypersurface it is impossible for a $t = \text{constant}$ hypersurface of a synchronous coordinate system with a simultaneous fictitious singularity to be a global Cauchy surface. These results demonstrate how the local series techniques led to the construction of artefacts which were, in the compact case not even space-times.

(vi) A discussion was given displaying the limitations and pitfalls of using local techniques to build up a large-scale picture of the singularity structure. Examples are given to show how this approach may enable us only to examine the structure of a single point on the singularity boundary even when the space-time is globally hyperbolic.

(vii) In later work BKL claimed to have established the 'general' (in the function counting sense) behaviour of the Einstein equations to be of 'Mixmaster' character having a simultaneous physical singularity. We show that in many cases apparently encompassed by their analysis such a claim is impossible since a simultaneous physical singularity could only occur in the Cauchy development of a hypersurface which is not a Cauchy surface for the entire space-time and thus *more* than four arbitrary three-dimensional functions would be necessary to specify it in general. We also argue (but do not prove) that the simultaneity of a physical singularity in a synchronous coordinate system is not a stable property of space-time. In particular, we show that the form of their "general solution" with a simultaneous physical singularity is inconsistent with the supposed stability of the simultaneity. We also show that a stable closed universe cannot be parameterized by four functions of three variables near a homogeneous space-time.

(viii) A number of physical objections are raised to local analyses and to the conclusion drawn from them, namely, that the general cosmological behaviour appears to be of Bianchi type VIII or IX-like character from place to place. The limitations and implications of the vacuum assumption near the singularity are considered together with possible global constraints placed upon the space-time structure due to the presence of matter.

(ix) It is argued that observational evidence indicates the initial state of the universe to have been of a dynamically non-generic character and some conjectures made which might explain this.

(x) Various recent claims by BKL regarding future cosmological singularities and the possibility of singularity-free general solutions to the Einstein equations are shown to be unfounded when examined globally.

These investigations lead us to suspect that the successful programme of investigation pursued by BKL to elucidate the general structure of homogeneous cosmological models employing local, coordinate based techniques cannot be successfully and naturally extended to build up a reliable picture of generic, inhomogeneous space-times. Local analyses are confronted with many difficulties when combined with the use of approximation methods and a global approach may be mandatory if correct and unambiguous information about the generic, singular structure of space-time is to be obtained.

Acknowledgements

We would especially like to thank D. Eardley, M.W. Hirsch, R.A. Matzner, I.D. Novikov, R.K. Sachs, D.W. Sciama, E.H. Spanier, and J.A. Wheeler for invaluable discussions and communications, and particularly J.E. Marsden, M.A.H. McCallum and S.T.C. Siklos for detailed contributions and advice. Financial support by the Lindemann Trust (J.D.B.), and NSF grant no. PHY-77-15191 (F.J.T.), is gratefully acknowledged. J.D.B. would like to thank Berkeley Astronomical Department for their hospitality during the preparation of this paper.

References

[1] E.M. Lifschitz, V.V. Sudakov and I.M. Khalatnikov, Sov. Phys. JETP. 13 (1961) 1298.
[2] E.M. Lifschitz and I.M. Khalatnikov, Sov. Phys. JETP. 12 (1961) 108, 558.
[3] E.M. Lifschitz, V.V. Sudakov and I.M. Khalatnikov, Phys. Rev. Lett. 6 (1961) 311.
[4] I.M. Khalatnikov, Sov. Phys. JETP. 21 (1965) 172.
[5] V.A. Belinskii and I.M. Khalatnikov, Sov. Phys. JETP. 22 (1966) 694.
[6] E.M. Lifschitz and I.M. Khalatnikov, Sov. Phys. Usp. 6 (1963) 495.
[7] L.P. Grishchuk, Sov. Phys. JETP. (1967) 320; 26 (1968) 974.
[8] V.A. Belinskii and I.M. Khalatnikov, Sov. Phys. JETP. 29 (1969) 911.
[9] E.M. Lifschitz, Contemporary physics, Vol. 1: Trieste Symposium (1968) (Inst. Atomic Energy Agency 1969).
[10] V.A. Belinskii and I.M. Khalatnikov, Sov. Phys. JETP. 30 (1970) 1174.
[11] V.A. Belinskii and I.M. Khalatnikov, Sov. Phys. JETP. 32 (1971) 169.
[12] V.A. Belinskii, E.M. Lifschitz and I.M. Khalatnikov, Sov. Phys. Usp. 13 (1971) 745.
[13] V.A. Belinskii, I.M. Khalatnikov and E.M. Lifschitz, Sov. Phys. JETP. 35 (1972) 838.
[14] L.C. Landau and E.M. Lifschitz, The classical theory of fields, 4th revised English ed. (Pergamon, 1975).
[15] V.A. Belinskii, communication at 'GR8' Symposium, Waterloo (August, 1977), to appear.
[16] V.A. Belinskii and I.M. Khalatnikov, Sov. Phys. JETP. 42 (1976) 205; 45 (1977) 1.
[17] V.A. Belinskii, I.M. Khalatnikov and E.M. Lifschitz, Sov. Phys. JETP. 33 (1971) 1061.
[18] V.A. Belinskii, I.M. Khalatnikov and M.P. Ryan, Ann. Phys. 70 (1972) 301.
[19] R. Penrose, Structure of space-time, in: Battele-rencontres, eds. DeWitt and J.A. Wheeler (Benjamin, N-Y, 1968).
[20] A. Komar, Phys. Rev. 104 (1956) 544.
[21] A. Raychandhuri, Phys. Rev. 98 (1955) 1123.
[22] R. Penrose, Phys. Rev. Lett. 14 (1965) 57.
[23] S.W. Hawking and R. Penrose, Proc. Roy. Soc. Lond. A 314 (1970) 529.
[24] C.W. Misner, Phys. Rev. Lett. 22 (1969) 1071.
[25] C.W. Misner, Phys. Rev. 186 (1969) 1328.
[26] M.A.H. MacCallum, in: Cargese lectures in physics, Vol. 6, ed. E. Schatzman (Gordon and Breach, N-Y, 1973).
[27] L.P. Grishchuk, Sov. Astron. 11 (1968) 881.
[28] M.A.H. MacCallum, Phys. Lett. 40A (1972) 325.
[29] L. Bianchi, Mem. Soc. It. Della. Sc. (Dei XL) (3) 11 (1897) 267.
[30] M.P. Ryan and L.C. Shepley, Homogeneous relativistic cosmologies (Princeton, 1975).
[31] M. Demianski and L.P. Grishchuk, Commun. Math. Phys. 25 (1972) 233.
[32] F. Lund, Phys. Rev. D 8 (1973) 3253, 4229.
[34] A.C. Scott, F.Y.F. Chu and D.W. McLaughlin, Proc. IEEE 61 (1973) 1443.
[35] V.G. Makhankiov, Physics Reports, 35C, no. 3 (1978).
[36] J.D. Barrow and B.J. Carr, Mon. Not. Roy. Astron. Soc. 182 (1978) 537.
[37] E.R. Harrison, in: Cargese lectures in physics, Vol. 6, ed. E. Schatzman (Gordon and Breach, N-Y, 1973).

[38] Y.B. Zeldovich, Adv. Astr. Astrophys. 3 (1965) 241.
[39] P.J.E. Peebles, Astrophys. J. 147 (1967) 859.
[40] G.F. Chapline, Phys. Rev. D 12 (1975) 2949.
[41] S. Aarseth, J.R. Bott and E.L. Turner, Proc. IAU Symposium No. 79 (1977), (Tallin, Estonia) to appear.
[42] S.D.M. White, Mon. Not. Roy. Astr. Soc. 177 (1976) 717.
[43] R.C. Tolman, Proc. Nat. Acad. Sci. 20 (1934) 169.
[44] H. Bondi, Mon. Not. Roy. Astron. Soc. 107 (1947) 410.
[45] P. Szekeres, Commun. Math. Phys. 41 (1975) 55.
[46] R. Kantowski and R.K. Sachs, J. Math. Phys. 7 (1966) 443.
[47] A. Einstein and N. Rosen, J. Franklin Inst. 223 (1937) 43.
[48] R. Gowdy, Phys. Rev. Lett. 27 (1971) 826; Ann. Phys. 83 (1974) 203.
[49] E.P. Liang, Astrophys. J. 204 (1976) 235.
[50] D.M. Eardley, E. Liang and R.K. Sachs, J. Math. Phys. 13 (1972) 99; 13 (1972) 386.
[51] R. Hagedorn, in: Cargese lectures in physics, Vol. 6, ed. E. Schatzman (Gordon and Breach, N-Y, 1973).
[52] J. Collins and M.J. Perry, Phys. Rev. Lett. 34 (1975) 1353.
[53] Y.B. Zeldovich, Sov. Phys. JETP. 14 (1962) 1143.
[54] J.D. Walecka, Ann. Phys. 83 (1974) 491.
[55] J.D. Walecka, Phys. Lett. 59B (1975) 109.
[56] J.D. Barrow, Nature 272 (1978) 211.
[57] P. Szekeres, Phys. Rev. D 12 (1975) 2941.
[58] H. Müller zum Hagen, H.J. Seifert and P. Yodzis, Commun. Math. Phys. 34 (1973) 135.
[59] J.D. Barrow and R.A. Matzner, Mon. Not. Roy. Astr. Soc. 181 (1977) 719.
[60] D.M. Eardley, GR8 Abstract No. 282 (Waterloo, 1977), to be published.
[61] S.W. Hawking and G.F.R. Ellis, The large scale structure of space-time (C.U.P., 1973).
[62] S.C. Siklos, D. Phil Thesis (Univ. Cambridge, 1977) and GR8 abstract no. 30. Comm. Math. Phys. 58 (1978) 255.
[63] This definition is the one used in HE [61], Penrose's Pittsburgh Lectures (p. 1), ref. [72].
[64] F.J. Tipler, J. Math. Phys. 18 (1977) 1568.
[65] S.W. Hawking, Gen. Rel. Grav. 1 (1971) 393.
[66] A.S. Kompanyeets, Sov. Phys. JETP. 7 (1958) 659; 10 (1960) 1218.
[67] C.J.S. Clarke, Commun. Math. Phys. 32 (1973) 205; 41 (1975) 65; 49 (1976) 17.
[68] A.E. Fischer and J.E. Marsden, Proc. Sym. Pure. Math. AMS 27 (2), (1975) 219.
[69] V. Moncrief, J. Math. Phys. 17 (1976) 1893 (the basic ideas for the sufficiency proof are in this paper). The details of the proof are in J.M. Arms and J.E. Marsden, The absence of Killing fields is necessary for linearization stability of Einstein's equations (1978, Berkeley Math. Dept preprint).
[70] J.W. York, Phys. Rev. Lett. 26 (1971) 1656.
[71] R. Schoen and S.T. Yau, Proc. Nat. Acad. Sci. (USA) 75 (1978) 2567.
[72] R. Penrose, Reg. Conf. Ser. Appl. Math. No. 7, Techniques of differential topology in relativity (Siam-Phil, 1972).
[73] I.M. Khalatnikov and E.M. Lifschitz, Phys. Rev. Lett. 24 (1970) 76.
[74] I.M. Khalatnikov, E.M. Lifschitz and V.V. Sudakov, Phys. Rev. Lett. 6 (1961) 311.
[75] see p. 289 of ref. [14].
[76] A.E. Fischer and J.E. Marsden, The space of gravitational degrees of freedom (Berkeley Math. Dept. preprint, 1978).
[77] A.E. Fischer, J.E. Marsden and V. Moncrief, The structure of the set of solutions to Einstein's equations (1978) in preparation
[78] J.M. Arms, J. Math. Phys. 18 (1977) 830; Ph.D. Thesis, Univ. California, Berkeley (1977).
[79] V. Moncrief, J. Math. Phys. 16 (1974) 493.
[80] Ref. [10] p. 1174.
[81] Ref. [9] p. 557.
[82] Ref. [6] p. 497.
[83] Ref. [8] p. 911.
[84] Ref. [6] p. 505.
[85] Ref. [6] p. 500.
[86] Ref. [6] p. 499.
[87] Ref. [10] p. 1174.
[88] Ref. [14] p. 396.
[89] Ref. [11] p. 172.
[90] Ref. [12] p. 759.
[91] A.A. Peresetskii, Math. Notes. Acad. Sci (USSR) 21 (1976) 71 see also
O.I. Bogoyavlenskii and S.P. Novikov, Russian Math. Survey 31 (5) (1977) 31. (These results rely on the 'separatrix approximation'.)
[92] C.B. Collins and S.W. Hawking, Astrophys. J. 180 (1973) 317.

[93] Ref. [10] p. 1179.
[94] V.A. Belinskii and I.M. Khalatnikov, Sov. Phys. JETP. 63 (1975) 1121.
[95] J.D. Barrow, Mon. Not. Roy. Astron. Soc. 188 (1979).
[96] A.R. King and G.F.R. Ellis, Commun. Math. Phys. 31 (1973) 209.
[97] R. Penrose, Proc. 1st. Marcel Grossman Meeting on General Relativity, ed. R. Ruffini (North-Holland, Amsterdam, 1977) p. 173.
[98] J.D. Barrow and F.J. Tipler, Nature 276 (1978) 453.
[99] R. Penrose, IAU Symposium No. 64, ed. C. DeWitt-Moreite (Reidel, Dordrecht 1974) p. 87.
[100] T.J.R. Hughes, T. Kato and J.E. Marsden, Arch. Rat. Mech. Anal. 63 (1977) 273.
[101] F.J. Tipler, Phys. Lett. 64A (1977) 8.
[102] J.A. Wheeler, The physicist's conception of nature, ed. J. Mehra (Reidel, Dordrecht) p. 202.
[103] F.J. Tipler, Phys. Rev. D 15 (1977) 942.
[104] D.R. Brill and F.J. Flaherty, Comm. Math. Phys. 50 (1976) 157, see proof of theorem 3.3.
[105] A.E. Fischer and J.E. Marsden, Nuovo Cimento (in press), (1978).
[106] F.J. Tipler, Nature 270 (1977) 500.
[107] R. Penrose, in: Theoretical principles in astrophysics and relativity, eds. N.R. Lebovitz, W.M. Reid and P.O. Vandervoort (University of Chicago Press, Chacago, 1978).
[108] A. Lichnerowicz, Theories relativistes de la gravitation et de l'electromagnétisme (Masson, Paris, 1955).
[109] Ref. [72] section 7.
[110] R. Budic, M. Gotay, L. Lindblom and P. Yasskin, Comm. Math. Phys. 61 (1978) 87.
[111] S. Lange, Analysis II (Addison-Wesley, Reading, 1969) p. 95; for the relation with the Jacobian see the implicit function theorem.
[112] P. D'Eath, Ann. Phys. 98 (1976) 237.
[113] B.F. Schutz Jr., Phys. Rev. D 4 (1971) 3559.
[114] J.M. Arms, Does matter break the link between symmetry and linearization stability?, preprint, University of Utah Physics Department (1979).
[115] J. Winicour, A.I. Janis and E.T. Newman, Phys. Rev. Lett. 20 (1968) 878.
[116] W. Rindler, Phys. Rev. Lett. 15 (1965) 1001.
[117] W. Israel, Phys. Rev. 153 (1967) 1388.
[118] F.J. Belinfante, Phys. Lett. 20 (1966) 25.
[119] J.L. Anderson and J. Gautreau, Phys. Lett. 20 (1966) 24.
[120] B. Schutz, J. Math. Phys. 16 (1975) 855.
[121] A. Einstein, Meaning of relativity, 5th. (Princeton U.P., 1955).
[122] S. Weinberg, Phys. Rev. Lett. 42 (1979) 850.
[123] J.D. Barrow and R.A. Matzner, The isotropy of the universe and the proton half-life, preprint Dept. of Astrophysics, Oxford (1979).