THE LARGE SCALE STRUCTURE OF SPACE–TIME

S. W. HAWKING, F.R.S.
Lucasian Professor of Mathematics in the University of Cambridge and
Fellow of Gonville and Caius College

AND

G. F. R. ELLIS
Professor of Applied Mathematics, University of Cape Town

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain

© Cambridge University Press 1973

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 1973
First paperback edition 1974
Reprinted 1976, 1977, 1979, 1980, 1984, 1986, 1987, 1989, 1991,
1993 (twice), 1994, 1995, 1997, 1999

Printed in the United States of America

Typeset in Caledonia

A catalog record for this book is available from the British Library

Library of Congress Catalog card number: 72-93671

ISBN 0 521 09906-4 paperback

Contents

Preface *page* xi

1 The role of gravity 1

2 Differential geometry 10
2.1 Manifolds 11
2.2 Vectors and tensors 15
2.3 Maps of manifolds 22
2.4 Exterior differentiation and the Lie derivative 24
2.5 Covariant differentiation and the curvature tensor 30
2.6 The metric 36
2.7 Hypersurfaces 44
2.8 The volume element and Gauss' theorem 47
2.9 Fibre bundles 50

3 General Relativity 56
3.1 The space–time manifold 56
3.2 The matter fields 59
3.3 Lagrangian formulation 64
3.4 The field equations 71

4 The physical significance of curvature 78
4.1 Timelike curves 78
4.2 Null curves 86
4.3 Energy conditions 88
4.4 Conjugate points 96
4.5 Variation of arc-length 102

5 Exact solutions 117
5.1 Minkowski space–time 118
5.2 De Sitter and anti-de Sitter space–times 124
5.3 Robertson–Walker spaces 134
5.4 Spatially homogeneous cosmological models 142

5.5 The Schwarzschild and Reissner–Nordström solutions *page* 149
5.6 The Kerr solution 161
5.7 Gödel's universe 168
5.8 Taub–NUT space 170
5.9 Further exact solutions 178

6 Causal structure 180
6.1 Orientability 181
6.2 Causal curves 182
6.3 Achronal boundaries 186
6.4 Causality conditions 189
6.5 Cauchy developments 201
6.6 Global hyperbolicity 206
6.7 The existence of geodesics 213
6.8 The causal boundary of space–time 217
6.9 Asymptotically simple spaces 221

7 The Cauchy problem in General Relativity 226
7.1 The nature of the problem 227
7.2 The reduced Einstein equations 228
7.3 The initial data 231
7.4 Second order hyperbolic equations 233
7.5 The existence and uniqueness of developments for the empty space Einstein equations 244
7.6 The maximal development and stability 249
7.7 The Einstein equations with matter 254

8 Space–time singularities 256
8.1 The definition of singularities 256
8.2 Singularity theorems 261
8.3 The description of singularities 276
8.4 The character of the singularities 284
8.5 Imprisoned incompleteness 289

9 Gravitational collapse and black holes 299
9.1 Stellar collapse 299
9.2 Black holes 308
9.3 The final state of black holes 323

10 The initial singularity in the universe *page* 348
10.1 The expansion of the universe 348
10.2 The nature and implications of singularities 359

Appendix A:
Translation of an essay by P. S. Laplace 365

Appendix B:
Spherically symmetric solutions and Birkhoff's theorem 369

References 373

Notation 381

Index 385

1
The role of gravity

The view of physics that is most generally accepted at the moment is that one can divide the discussion of the universe into two parts. First, there is the question of the local laws satisfied by the various physical fields. These are usually expressed in the form of differential equations. Secondly, there is the problem of the boundary conditions for these equations, and the global nature of their solutions. This involves thinking about the edge of space–time in some sense. These two parts may not be independent. Indeed it has been held that the local laws are determined by the large scale structure of the universe. This view is generally connected with the name of Mach, and has more recently been developed by Dirac (1938), Sciama (1953), Dicke (1964), Hoyle and Narlikar (1964), and others. We shall adopt a less ambitious approach: we shall take the local physical laws that have been experimentally determined, and shall see what these laws imply about the large scale structure of the universe.

There is of course a large extrapolation in the assumption that the physical laws one determines in the laboratory should apply at other points of space–time where conditions may be very different. If they failed to hold we should take the view that there was some other physical field which entered into the local physical laws but whose existence had not yet been detected in our experiments, because it varies very little over a region such as the solar system. In fact most of our results will be independent of the detailed nature of the physical laws, but will merely involve certain general properties such as the description of space–time by a pseudo-Riemannian geometry and the positive definiteness of energy density.

The fundamental interactions at present known to physics can be divided into four classes: the strong and weak nuclear interactions, electromagnetism, and gravity. Of these, gravity is by far the weakest (the ratio Gm^2/e^2 of the gravitational to electric force between two electrons is about 10^{-40}). Nevertheless it plays the dominant role in shaping the large scale structure of the universe. This is because the

strong and weak interactions have a very short range ($\sim 10^{-13}$ cm or less), and although electromagnetism is a long range interaction, the repulsion of like charges is very nearly balanced, for bodies of macroscopic dimensions, by the attraction of opposite charges. Gravity on the other hand appears to be always attractive. Thus the gravitational fields of all the particles in a body add up to produce a field which, for sufficiently large bodies, dominates over all other forces.

Not only is gravity the dominant force on a large scale, but it is a force which affects every particle in the same way. This universality was first recognized by Galileo, who found that any two bodies fell with the same velocity. This has been verified to very high precision in more recent experiments by Eotvos, and by Dicke and his collaborators (Dicke (1964)). It has also been observed that light is deflected by gravitational fields. Since it is thought that no signals can travel faster than light, this means that gravity determines the causal structure of the universe, i.e. it determines which events of space–time can be causally related to each other.

These properties of gravity lead to severe problems, for if a sufficiently large amount of matter were concentrated in some region, it could deflect light going out from the region so much that it was in fact dragged back inwards. This was recognized in 1798 by Laplace, who pointed out that a body of about the same density as the sun but 250 times its radius would exert such a strong gravitational field that no light could escape from its surface. That this should have been predicted so early is so striking that we give a translation of Laplace's essay in an appendix.

One can express the dragging back of light by a massive body more precisely using Penrose's idea of a closed trapped surface. Consider a sphere $\mathscr{T}$ surrounding the body. At some instant let $\mathscr{T}$ emit a flash of light. At some later time t, the ingoing and outgoing wave fronts from $\mathscr{T}$ will form spheres $\mathscr{T}_1$ and $\mathscr{T}_2$ respectively. In a normal situation, the area of $\mathscr{T}_1$ will be less than that of $\mathscr{T}$ (because it represents ingoing light) and the area of $\mathscr{T}_2$ will be greater than that of $\mathscr{T}$ (because it represents outgoing light; see figure 1). However if a sufficiently large amount of matter is enclosed within $\mathscr{T}$, the areas of $\mathscr{T}_1$ and $\mathscr{T}_2$ will *both* be less than that of $\mathscr{T}$. The surface $\mathscr{T}$ is then said to be a closed trapped surface. As t increases, the area of $\mathscr{T}_2$ will get smaller and smaller provided that gravity remains attractive, i.e. provided that the energy density of the matter does not become negative. Since the matter inside $\mathscr{T}$ cannot travel faster than light, it will be

trapped within a region whose boundary decreases to zero within a finite time. This suggests that something goes badly wrong. We shall in fact show that in such a situation a space–time singularity must occur, if certain reasonable conditions hold.

One can think of a singularity as a place where our present laws of physics break down. Alternatively, one can think of it as representing part of the edge of space–time, but a part which is at a finite distance instead of at infinity. On this view, singularities are not so bad, but one still has the problem of the boundary conditions. In other words, one does not know what will come out of the singularity.

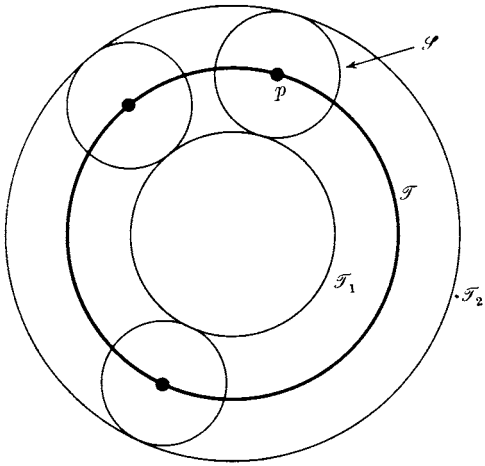

FIGURE 1. At some instant, the sphere $\mathscr{T}$ emits a flash of light. At a later time, the light from a point p forms a sphere $\mathscr{S}$ around p, and the envelopes $\mathscr{T}_1$ and $\mathscr{T}_2$ form the ingoing and outgoing wavefronts respectively. If the areas of *both* $\mathscr{T}_1$ and $\mathscr{T}_2$ are less than the area of $\mathscr{T}$, then $\mathscr{T}$ is a closed trapped surface.

There are two situations in which we expect there to be a sufficient concentration of matter to cause a closed trapped surface. The first is in the gravitational collapse of stars of more than twice the mass of the sun, which is predicted to occur when they have exhausted their nuclear fuel. In this situation, we expect the star to collapse to a singularity which is not visible to outside observers. The second situation is that of the whole universe itself. Recent observations of the microwave background indicate that the universe contains enough matter to cause a time-reversed closed trapped surface. This implies the existence of a singularity in the past, at the beginning of the present epoch of expansion of the universe. This singularity is in principle visible to us. It might be interpreted as the beginning of the universe.

In this book we shall study the large scale structure of space–time on the basis of Einstein's General Theory of Relativity. The predictions of this theory are in agreement with all the experiments so far performed. However our treatment will be sufficiently general to cover modifications of Einstein's theory such as the Brans–Dicke theory.

While we expect that most of our readers will have some acquaintance with General Relativity, we have endeavoured to write this book so that it is self-contained apart from requiring a knowledge of simple calculus, algebra and point set topology. We have therefore devoted chapter 2 to differential geometry. Our treatment is reasonably modern in that we have formulated our definitions in a manifestly coordinate independent manner. However for computational convenience we do use indices at times, and we have for the most part avoided the use of fibre bundles. The reader with some knowledge of differential geometry may wish to skip this chapter.

In chapter 3 a formulation of the General Theory of Relativity is given in terms of three postulates about a mathematical model for space–time. This model is a manifold $\mathscr{M}$ with a metric $\mathbf{g}$ of Lorentz signature. The physical significance of the metric is given by the first two postulates: those of local causality and of local conservation of energy–momentum. These postulates are common to both the General and the Special Theories of Relativity, and so are supported by the experimental evidence for the latter theory. The third postulate, the field equations for the metric $\mathbf{g}$, is less well experimentally established. However most of our results will depend only on the property of the field equations that gravity is attractive for positive matter densities. This property is common to General Relativity and some modifications such as the Brans–Dicke theory.

In chapter 4, we discuss the significance of curvature by considering its effects on families of timelike and null geodesics. These represent the paths of small particles and of light rays respectively. The curvature can be interpreted as a differential or tidal force which induces relative accelerations between neighbouring geodesics. If the energy–momentum tensor satisfies certain positive definite conditions, this differential force always has a net converging effect on non-rotating families of geodesics. One can show by use of Raychaudhuri's equation (4.26) that this then leads to focal or conjugate points where neighbouring geodesics intersect.

To see the significance of these focal points, consider a one-dimensional surface $\mathscr{S}$ in two-dimensional Euclidean space (figure 2). Let p

be a point not on $\mathscr{S}$. Then there will be some curve from $\mathscr{S}$ to p which is shorter than, or as short as, any other curve from $\mathscr{S}$ to p. Clearly this curve will be a geodesic, i.e. a straight line, and will intersect $\mathscr{S}$ orthogonally. In the situation shown in figure 2, there are in fact three geodesics orthogonal to $\mathscr{S}$ which pass through p. The geodesic through the point r is clearly not the shortest curve from $\mathscr{S}$ to p. One way of recognizing this (Milnor (1963)) is to notice that the neighbouring

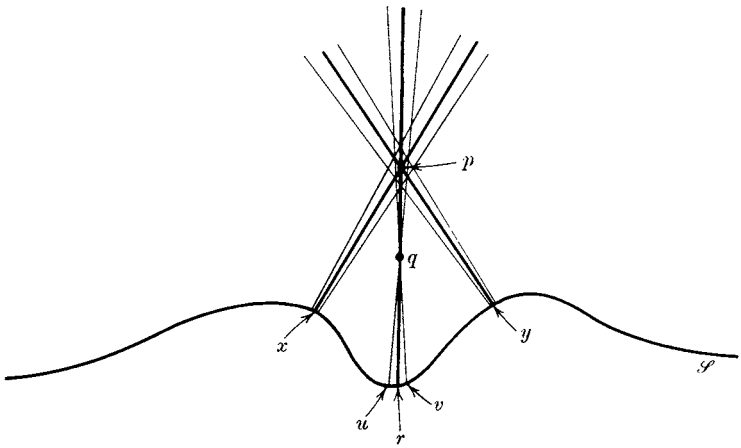

FIGURE 2. The line pr cannot be the shortest line from p to $\mathscr{S}$, because there is a focal point q between p and r. In fact either px or py will be the shortest line from p to $\mathscr{S}$.

geodesics orthogonal to $\mathscr{S}$ through u and v intersect the geodesic through r at a focal point q between $\mathscr{S}$ and p. Then joining the segment uq to the segment qp, one could obtain a curve from $\mathscr{S}$ to p which had the same length as a straight line rp. However as uqp is not a straight line, one could round off the corner at q to obtain a curve from $\mathscr{S}$ to p which was shorter than rp. This shows that rp is not the shortest curve from $\mathscr{S}$ to p. In fact the shortest curve will be either xp or yp.

One can carry these ideas over to the four-dimensional space–time manifold $\mathscr{M}$ with the Lorentz metric $\mathbf{g}$. Instead of straight lines, one considers geodesics, and instead of considering the shortest curve one considers the longest timelike curve between a point p and a spacelike surface $\mathscr{S}$ (because of the Lorentz signature of the metric, there will be no shortest timelike curve but there may be a longest such curve). This longest curve must be a geodesic which intersects $\mathscr{S}$ orthogonally, and there can be no focal point of geodesics orthogonal to $\mathscr{S}$ between

$\mathscr{S}$ and p. Similar results can be proved for null geodesics. These results are used in chapter 8 to establish the existence of singularities under certain conditions.

In chapter 5 we describe a number of exact solutions of Einstein's equations. These solutions are not realistic in that they all possess exact symmetries. However they provide useful examples for the succeeding chapters and illustrate various possible behaviours. In particular, the highly symmetrical cosmological models nearly all possess space–time singularities. For a long time it was thought that these singularities might be simply a result of the high degree of symmetry, and would not be present in more realistic models. It will be one of our main objects to show that this is not the case.

In chapter 6 we study the causal structure of space–time. In Special Relativity, the events that a given event can be causally affected by, or can causally affect, are the interiors of the past and future light cones respectively (see figure 3). However in General Relativity the metric $\mathbf{g}$ which determines the light cones will in general vary from point to point, and the topology of the space–time manifold $\mathscr{M}$ need not be that of Euclidean space R^4. This allows many more possibilities. For instance one can identify corresponding points on the surfaces $\mathscr{S}_1$ and $\mathscr{S}_2$ in figure 3, to produce a space–time with topology $R^3 \times S^1$. This would contain closed timelike curves. The existence of such a curve would lead to causality breakdowns in that one could travel into one's past. We shall mostly consider only space–times which do not permit such causality violations. In such a space–time, given any spacelike surface $\mathscr{S}$, there is a maximal region of space–time (called the Cauchy development of $\mathscr{S}$) which can be predicted from knowledge of data on $\mathscr{S}$. A Cauchy development has a property ('Global hyperbolicity') which implies that if two points in it can be joined by a timelike curve, then there exists a longest such curve between the points. This curve will be a geodesic.

The causal structure of space–time can be used to define a boundary or edge to space–time. This boundary represents both infinity and the part of the edge of space–time which is at a finite distance, i.e. the singular points.

In chapter 7 we discuss the Cauchy problem for General Relativity. We show that initial data on a spacelike surface determines a unique solution on the Cauchy development of the surface, and that in a certain sense this solution depends continuously on the initial data. This chapter is included for completeness and because it uses a number

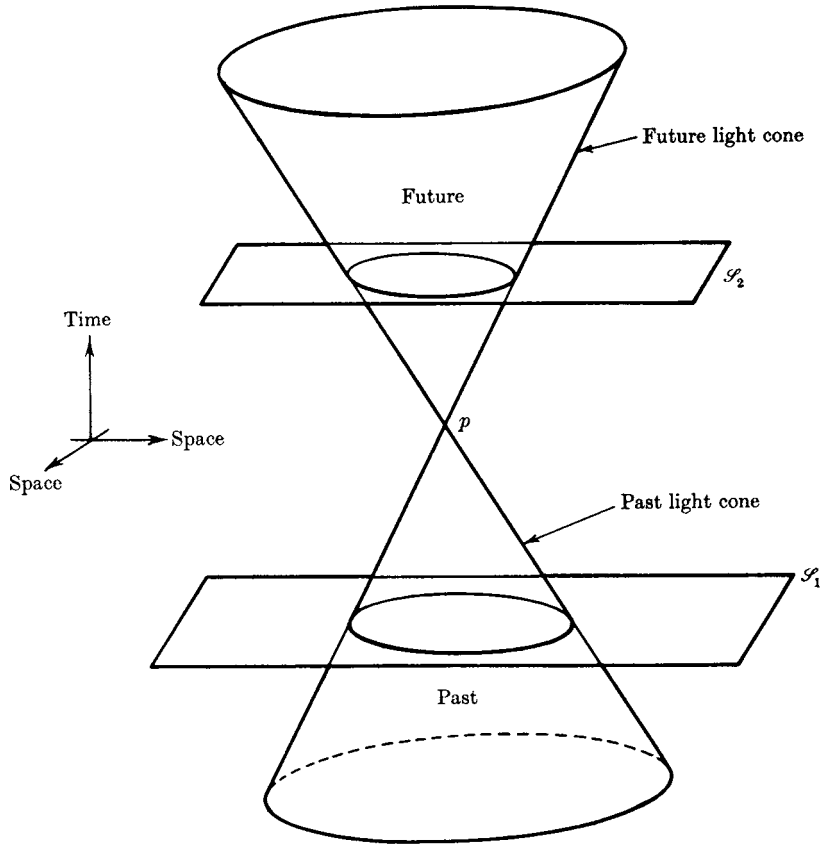

FIGURE 3. In Special Relativity, the light cone of an event p is the set of all light rays through p. The past of p is the interior of the past light cone, and the future of p is the interior of the future light cone.

of results of the previous chapter. However it is not necessary to read it in order to understand the following chapters.

In chapter 8 we discuss the definition of space–time singularities. This presents certain difficulties because one cannot regard the singular points as being part of the space–time manifold $\mathscr{M}$.

We then prove four theorems which establish the occurrence of space–time singularities under certain conditions. These conditions fall into three categories. First, there is the requirement that gravity shall be attractive. This can be expressed as an inequality on the energy–momentum tensor. Secondly, there is the requirement that there is enough matter present in some region to prevent anything escaping from that region. This will occur if there is a closed trapped

surface, or if the whole universe is itself spatially closed. The third requirement is that there should be no causality violations. However this requirement is not necessary in one of the theorems. The basic idea of the proofs is to use the results of chapter 6 to prove there must be longest timelike curves between certain pairs of points. One then shows that if there were no singularities, there would be focal points which would imply that there were no longest curves between the pairs of points.

We next describe a procedure suggested by Schmidt for constructing a boundary to space–time which represents the singular points of space–time. This boundary may be different from that part of the causal boundary (defined in chapter 6) which represents singularities.

In chapter 9, we show that the second condition of theorem 2 of chapter 8 should be satisfied near stars of more than $1\frac{1}{2}$ times the solar mass in the final stages of their evolution. The singularities which occur are probably hidden behind an event horizon, and so are not visible from outside. To an external observer, there appears to be a 'black hole' where the star once was. We discuss the properties of such black holes, and show that they probably settle down finally to one of the Kerr family of solutions. Assuming this to be the case, one can place certain upper bounds on the amount of energy which can be extracted from black holes. In chapter 10 we show that the second conditions of theorems 2 and 3 of chapter 8 should be satisfied, in a time-reversed sense, in the whole universe. In this case, the singularities are in our past and constitute a beginning for all or part of the observed universe.

The essential part of the introductory material is that in § 3.1, § 3.2 and § 3.4. A reader wishing to understand the theorems predicting the existence of singularities in the universe need read further only chapter 4, § 6.2–§ 6.7, and § 8.1 and § 8.2. The application of these theorems to collapsing stars follows in § 9.1 (which uses the results of appendix B); the application to the universe as a whole is given in § 10.1, and relies on an understanding of the Robertson–Walker universe models (§ 5.3). Our discussion of the nature of the singularities is contained in § 8.1, § 8.3–§ 8.5, and § 10.2; the example of Taub–NUT space (§ 5.8) plays an important part in this discussion, and the Bianchi I universe model (§ 5.4) is also of some interest.

A reader wishing to follow our discussion of black holes need read only chapter 4, § 6.2–§ 6.6, § 6.9, and § 9.1, § 9.2 and § 9.3. This discussion relies on an understanding of the Schwarzschild solution (§ 5.5) and of the Kerr solution (§ 5.6).

Finally a reader whose main interest is in the time evolution properties of Einstein's equations need read only § 6.2–§ 6.6 and chapter 7. He will find interesting examples given in § 5.1, § 5.2 and § 5.5.

We have endeavoured to make the index a useful guide to all the definitions introduced, and the relations between them.