fying points in $\mathscr{M}$ which are equivalent under a discrete isometry without a fixed point (e.g. identifying the point (x^1, x^2, x^3, x^4) with the point $(x^1, x^2, x^3, x^4 + c)$, where c is a constant, changes the topological structure from R^4 to $R^3 \times S^1$, and introduces closed timelike lines into the space–time). Clearly, $(\mathscr{M}, \boldsymbol{\eta})$ is the universal covering space for all such derived spaces, which have been studied in detail by Auslander and Markus (1958).

5.2 De Sitter and anti-de Sitter space–times

The space–time metrics of constant curvature are locally characterized by the condition $R_{abcd} = \frac{1}{12}R(g_{ac}g_{bd} - g_{ad}g_{bc})$. This equation is equivalent to $C_{abcd} = 0 = R_{ab} - \frac{1}{4}Rg_{ab}$; thus the Riemann tensor is determined by the Ricci scalar R alone. It follows at once from the contracted Bianchi identities that R is constant throughout space–time; in fact these space–times are homogeneous. The Einstein tensor is

$$R_{ab} - \tfrac{1}{2}Rg_{ab} = -\tfrac{1}{4}Rg_{ab}.$$

One can therefore regard these spaces as solutions of the field equations for an empty space with $\Lambda = \frac{1}{4}R$, or for a perfect fluid with a constant density $R/32\pi$ and a constant pressure $-R/32\pi$. However the latter choice does not seem reasonable, as in this case one cannot have both the density and the pressure positive; in addition, the equation of motion (3.10) is indeterminate for such a fluid.

The space of constant curvature with $R = 0$ is Minkowski space–time. The space for $R > 0$ is *de Sitter space–time*, which has the topology $R^1 \times S^3$ (see Schrödinger (1956) for an interesting account of this space). It is easiest visualized as the hyperboloid

$$-v^2 + w^2 + x^2 + y^2 + z^2 = \alpha^2$$

in flat five-dimensional space R^5 with metric

$$-\mathrm{d}v^2 + \mathrm{d}w^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 = \mathrm{d}s^2$$

(see figure 16). One can introduce coordinates (t, χ, θ, ϕ) on the hyperboloid by the relations

$$\alpha \sinh(\alpha^{-1}t) = v, \quad \alpha \cosh(\alpha^{-1}t)\cos\chi = w,$$

$$\alpha \cosh(\alpha^{-1}t)\sin\chi\cos\theta = x, \quad \alpha\cosh(\alpha^{-1}t)\sin\chi\sin\theta\cos\phi = y,$$

$$\alpha\cosh(\alpha^{-1}t)\sin\chi\sin\theta\sin\phi = z.$$

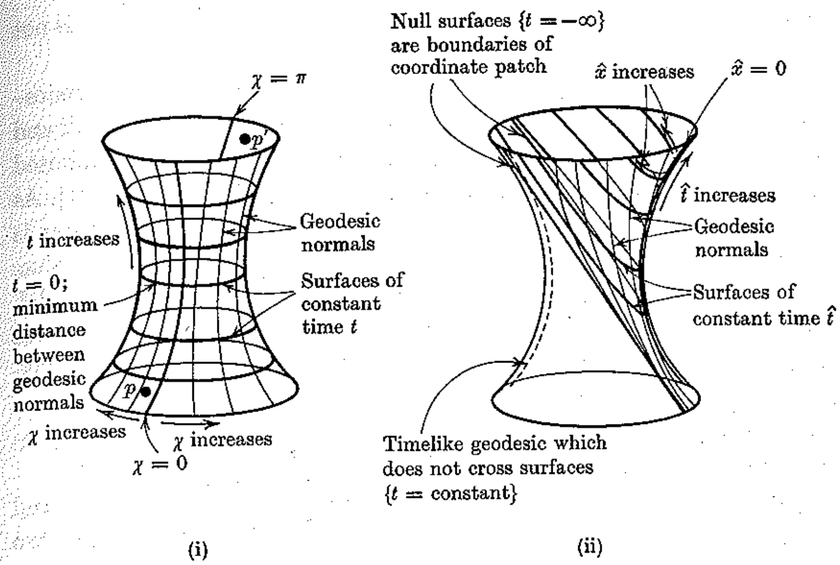


FIGURE 16. De Sitter space–time represented by a hyperboloid imbedded in a five-dimensional flat space (two dimensions are suppressed in the figure).

(i) Coordinates (t, χ, θ, ϕ) cover the whole hyperboloid; the sections $\{t = \text{constant}\}$ are surfaces of curvature $k = +1$.

(ii) Coordinates $(\hat{t}, \hat{x}, \hat{y}, \hat{z})$ cover half the hyperboloid; the surfaces $\{\hat{t} = \text{constant}\}$ are flat three-spaces, their geodesic normals diverging from a point in the infinite past.

In these coordinates, the metric has the form

$$ds^2 = -dt^2 + \alpha^2 . \cosh^2(\alpha^{-1}t) . \{d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)\}.$$

The singularities in the metric at $\chi = 0$, $\chi = \pi$ and at $\theta = 0$, $\theta = \pi$, are simply those that occur with polar coordinates. Apart from these trivial singularities, the coordinates cover the whole space for $-\infty < t < \infty$, $0 \leqslant \chi \leqslant \pi$, $0 \leqslant \theta \leqslant \pi$, $0 \leqslant \phi \leqslant 2\pi$. The spatial sections of constant t are spheres S^3 of constant positive curvature and are Cauchy surfaces. Their geodesic normals are lines which contract monotonically to a minimum spatial separation and then re-expand to infinity (see figure 16 (i)).

One can also introduce coordinates

$$\hat{t} = \alpha \log \frac{w+v}{\alpha}, \quad \hat{x} = \frac{\alpha x}{w+v}, \quad \hat{y} = \frac{\alpha y}{w+v}, \quad \hat{z} = \frac{\alpha z}{w+v}$$

on the hyperboloid. In these coordinates, the metric takes the form

$$ds^2 = -d\hat{t}^2 + \exp(2\alpha^{-1}\hat{t})(d\hat{x}^2 + d\hat{y}^2 + d\hat{z}^2).$$

However these coordinates cover only half the hyperboloid as $\hat{t}$ is not defined for $w+v \leqslant 0$ (see figure 16 (ii)).

The region of de Sitter space for which $v+w > 0$ forms the space–time for the *steady state* model of the universe proposed by Bondi and Gold (1948) and Hoyle (1948). In this model, the matter is supposed to move along the geodesic normals to the surfaces $\{\hat{t} = \text{constant}\}$. As the matter moves further apart, it is assumed that more matter is continuously created to maintain the density at a constant value. Bondi and Gold did not seek to provide field equations for this model, but Pirani (1955), and Hoyle and Narlikar (1964) have pointed out that the metric can be considered as a solution of the Einstein equations (with $\Lambda = 0$) if in addition to the ordinary matter one introduces a scalar field of negative energy density. This 'C'-field would also be responsible for the continual creation of matter.

The steady state theory has the advantage of making simple and definite predictions. However from our point of view there are two unsatisfactory features. The first is the existence of negative energy, which was discussed in § 4.3. The other is the fact that the space–time is extendible, being only half of de Sitter space. Despite these aesthetic objections, the real test of the steady state theory is whether its predictions agree with observations or not. At the moment it seems that they do not, though the observations are not yet quite conclusive.

de Sitter space is geodesically complete; however, there are points in the space which cannot be joined to each other by any geodesic. This is in contrast to spaces with a positive definite metric, when geodesic completeness guarantees that any two points of a space can be joined by at least one geodesic. The half of de Sitter space which represents the steady state universe is not complete in the past (there are geodesics which are complete in the full space, and cross the boundary of the steady state region; they are therefore incomplete in that region).

To study infinity in de Sitter space–time, we define a time coordinate t' by

$$t' = 2 \arctan(\exp \alpha^{-1} t) - \tfrac{1}{2}\pi,$$

where

$$-\tfrac{1}{2}\pi < t' < \tfrac{1}{2}\pi. \tag{5.8}$$

Then

$$ds^2 = \alpha^2 \cosh^2(\alpha^{-1}t') \,.\, d\bar{s}^2,$$

where $d\bar{s}^2$ is given by (5.7) on identifying $r' = \chi$. Thus the de Sitter space is conformal to that part of the Einstein static universe defined by (5.8) (see figure 17 (i)). The Penrose diagram of de Sitter space is accordingly as in figure 17 (ii). One half of this figure gives the Penrose

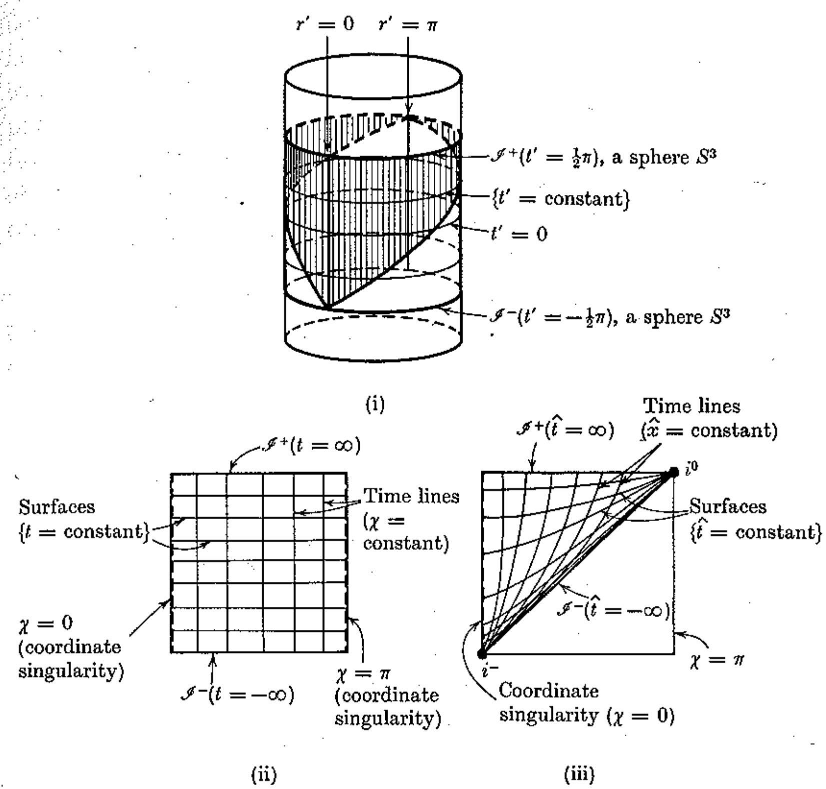


FIGURE 17

(i) De Sitter space–time is conformal to the region $-\frac{1}{2}\pi < t' < \frac{1}{2}\pi$ of the Einstein static universe. The steady state universe is conformal to the shaded region.

(ii) The Penrose diagram of de Sitter space–time.

(iii) The Penrose diagram of the steady state universe.

In (ii), (iii) each point represents a two-sphere of area $2\pi \sin^2 \chi$; null lines are at 45°. $\chi = 0$ and $\chi = \pi$ are identified.

diagram of the half of de Sitter space–time which constitutes the steady state universe (figure 17 (iii)).

One sees that de Sitter space has, in contrast to Minkowski space, a spacelike infinity for timelike and null lines, both in the future and the past. This difference corresponds to the existence in de Sitter space–time of both particle and event horizons for geodesic families of observers.

In de Sitter space, consider a family of particles whose histories are timelike geodesics; these must originate at the spacelike infinity $\mathscr{I}^-$ and end at the spacelike infinity $\mathscr{I}^+$. Let p be some event on the world-

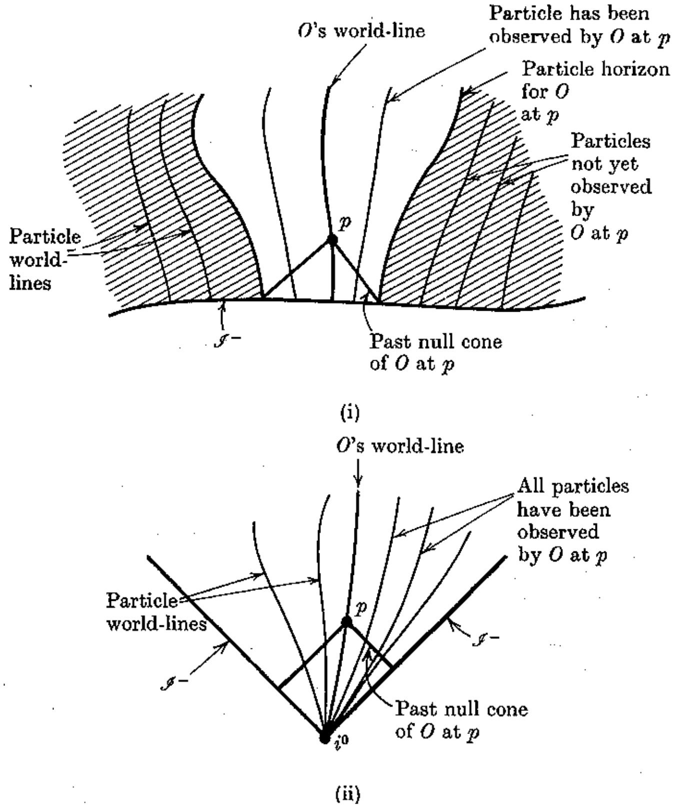


FIGURE 18

(i) The particle horizon defined by a congruence of geodesic curves when past null infinity $\mathscr{I}^-$ is spacelike.

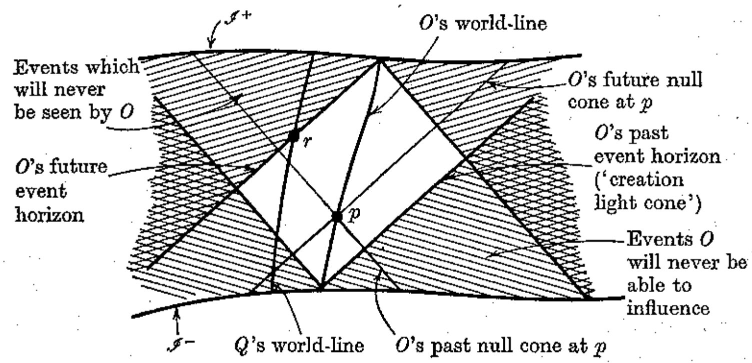
(ii) Lack of such a horizon if $\mathscr{I}^-$ is null.

line of a particle O in this family, i.e. some time in its history (proper time measured along O's world-line). The past null cone of p is the set of events in space–time which can be observed by O at that time. The world-lines of some other particles may intersect this null cone; these particles are visible to O. However, there can exist particles whose world-lines do not intersect this null cone, and so are not yet visible to O. At a later time O can observe more particles, but there still exist particles not visible to O at that time. We say that the division of particles into those seen by O at p and those not seen by O at p, is the *particle horizon* for the observer O at the event p; it represents the history of those particles lying at the limits of O's vision. Note that it is determined only when the world-lines of all the particles in the family are known. If some particle lies on the horizon, then the event p is the event at which the particle's creation light cone intersects O's world-line. In Minkowski space, on the other hand, all the other particles are visible at any event p on O's world-line if they move on timelike geodesics. As long as one considers only families of geodesic observers, one may think of the existence of the particle horizon as a consequence of past null infinity being spacelike (see figure 18).

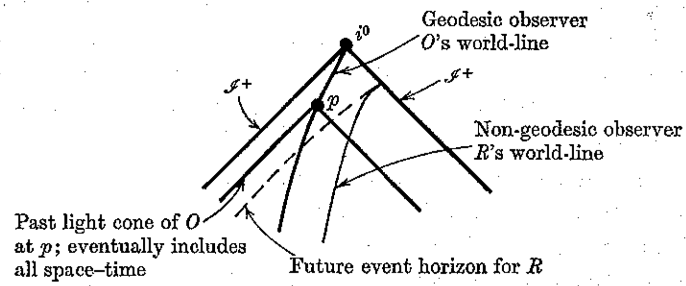
All events outside the past null cone of p are events which are not, and never have been, observable by O up to the time represented by the event p. There is a limit to O's world-line on $\mathscr{I}^+$. In de Sitter space–time, the past null cone of this point (obtained by a limiting process in the actual space–time, or directly from the conformal space–time) is a boundary between events which will at some time be observable by O, and those that will never be observable by O. We call this surface the *future event horizon* of the world-line. It is the boundary of the past of the world-line. In Minkowski space–time, on the other hand, the limiting null cone of any geodesic observer includes the whole of space–time, so there are no events which a geodesic observer will never be able to see. However if an observer moves with uniform acceleration his world-line may have a future event horizon. One may think of the existence of a future event horizon for a geodesic observer as being a consequence of $\mathscr{I}^+$ being spacelike (see figure 19).

Consider the event horizon for the observer O in de Sitter space–time and suppose that at some proper time (event p) on his world-line, his light cone intersects the world-line of the particle Q. Then Q is always visible to O at times after p. However there is on Q's world-line an event r which lies on O's future event horizon; O can never see later events on Q's world-line than r. Moreover an infinite proper time elapses on O's world-line from any given point till he observes r, but a finite proper time elapses along Q's world-line from any given event to r, which is a perfectly ordinary event on his world-line. Thus O sees a finite part of Q's history in an infinite time; expressed more physically, as O observes Q he sees a redshift which approaches infinity as O observes points on Q's world-line which approach r. Correspondingly, Q never sees beyond some point on O's world-line, and sees nearby points on O's world-line only with a very large redshift.

At any point on O's world-line, the future null cone is the boundary of the set of events in space–time which O can influence at and after that time. To obtain the maximal set of events in space–time that O could at any time influence, we take the future light cone of the limit



(i)



(ii)

FIGURE 19

(i) The future event horizon for a particle O which exists when future infinity $\mathscr{I}^+$ is spacelike; also the past event horizon which exists when past infinity $\mathscr{I}^-$ is spacelike.

(ii) If future infinity consists of a null $\mathscr{I}^+$ and i^0, there is no future event horizon for a geodesic observer O. However an accelerating observer R may have a future event horizon.

point of O's world-line on past infinity $\mathscr{I}^-$; that is, we take the boundary of the future of the world-line (which can be regarded as O's creation light cone). This has a non-trivial existence for a geodesic observer only if the past infinity $\mathscr{I}^-$ is spacelike (and is in fact then O's past event horizon). It is clear from the above discussion that in the steady state universe, which has a null past infinity for timelike and null geodesics and a spacelike future infinity, any fundamental observer has a future event horizon but no past particle horizon.

One can obtain other spaces which are locally equivalent to the de Sitter space, by identifying points in de Sitter space. The simplest such identification is to identify antipodal points p, p' (see figure 16) on the hyperboloid. The resulting space is not time orientable; if time increases in the direction of the arrow at p, the antipodal identification implies it must increase in the direction of the arrow at p', but one cannot continuously extend this identification of future and past half null cones over the whole hyperboloid. Calabi and Markus (1962) have studied in detail the spaces resulting from such identifications; they show in particular that an arbitrary point in the resulting space can be joined to any other point by a geodesic if and only if it is not time orientable.

The space of constant curvature with $R < 0$ is called *anti-de Sitter space*. It has the topology $S^1 \times R^3$, and can be represented as the hyperboloid

$$-u^2 - v^2 + x^2 + y^2 + z^2 = 1$$

in the flat five-dimensional space R^5 with metric

$$\mathrm{d}s^2 = -(\mathrm{d}u)^2 - (\mathrm{d}v)^2 + (\mathrm{d}x)^2 + (\mathrm{d}y)^2 + (\mathrm{d}z)^2.$$

There are closed timelike lines in this space; however it is not simply connected, and if one unwraps the circle S^1 (to obtain its covering space R^1) one obtains the universal covering space of anti-de Sitter space which does not contain any closed timelike lines. This has the topology of R^4. We shall in future mean by 'anti-de Sitter space', this universal covering space.

It can be represented by the metric

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \cos^2 t\,\{\mathrm{d}\chi^2 + \sinh^2\chi(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\}. \qquad (5.9)$$

This coordinate system covers only part of the space, and has apparent singularities at $t = \pm\frac{1}{2}\pi$. The whole space can be covered by coordinates $\{t', r, \theta, \phi\}$ for which the metric has the static form

$$\mathrm{d}s^2 = -\cosh^2 r\,\mathrm{d}t'^2 + \mathrm{d}r^2 + \sinh^2 r(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2).$$

In this form, the space is covered by the surfaces $\{t' = \text{constant}\}$ which have non-geodesic normals.

To study the structure at infinity, define the coordinate r' by

$$r' = 2\arctan(\exp r) - \tfrac{1}{2}\pi, \quad 0 \leqslant r' < \tfrac{1}{2}\pi.$$

Then one finds $\mathrm{d}s^2 = \cosh^2 r\,\mathrm{d}\bar{s}^2$, where $\mathrm{d}\bar{s}^2$ is given by (5.7); that is, the whole of anti-de Sitter space is conformal to the region $0 \leqslant r' < \frac{1}{2}\pi$ of the Einstein static cylinder. The Penrose diagram is shown in figure 20; null and spacelike infinity can be thought of as a timelike surface in this case. This surface has the topology $R^1 \times S^2$.

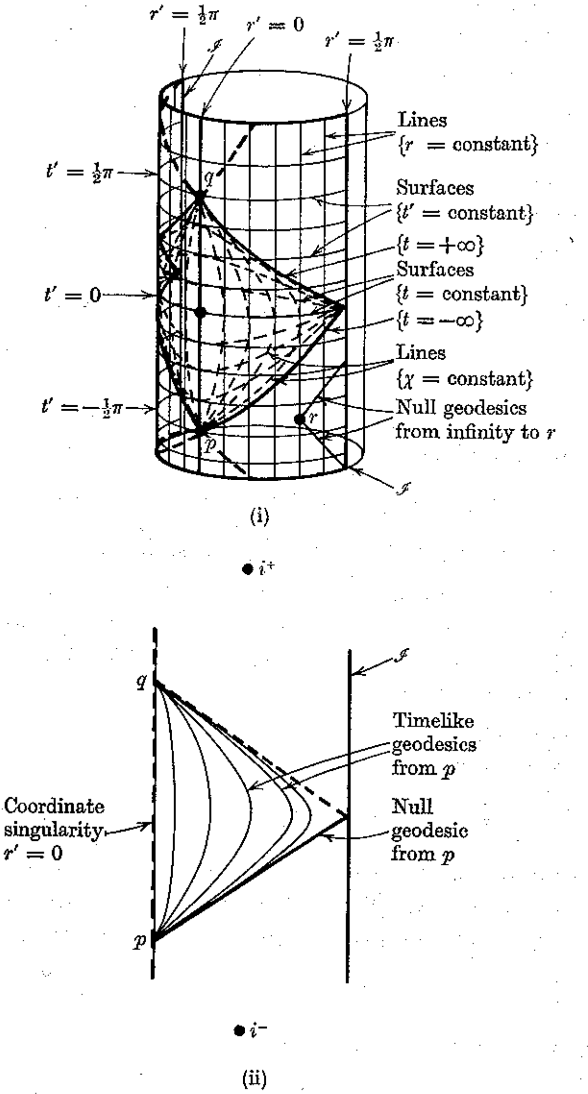


FIGURE 20

(i) Universal anti-de Sitter space is conformal to one half of the Einstein static universe. While coordinates (t', r, θ, ϕ) cover the whole space, coordinates (t, χ, θ, ϕ) cover only one diamond-shaped region as shown. The geodesics orthogonal to the surfaces $\{t = \text{constant}\}$ all converge at p and q, and then diverge out into similar diamond-shaped regions.

(ii) The Penrose diagram of universal anti-de Sitter space. Infinity consists of the timelike surface $\mathscr{I}$ and the disjoint points i^+, i^-. The projection of some timelike and null geodesics is shown.

One cannot find a conformal transformation which makes timelike infinity finite without pinching off the Einstein static universe to a point (if a conformal transformation makes the time coordinate finite it also scales the space sections by an infinite factor), so we represent timelike infinity by the disjoint points i^+, i^-.

The lines $\{\chi, \theta, \phi \text{ constant}\}$ are the geodesics orthogonal to the surfaces $\{t = \text{constant}\}$; they all converge to points q (respectively, p) in the future (respectively, past) of the surface, and this convergence is the reason for the apparent (coordinate) singularities in the original metric form. The region covered by these coordinates is the region between the surface $t = 0$ and the null surfaces on which these normals become degenerate.

The space has two further interesting properties. First, as a consequence of the timelike infinity, there exists no Cauchy surface whatever in the space. While one can find families of spacelike surfaces (such as the surfaces $\{t' = \text{constant}\}$) which cover the space completely, each surface being a complete cross-section of the space–time, one can find null geodesics which never intersect any given surface in the family. Given initial data on any such surface, one cannot predict beyond the Cauchy development of the surface; thus from the surface $\{t = 0\}$, one can predict only in the region covered by the coordinates t, χ, θ, ϕ. Any attempt to predict beyond this region is prevented by fresh information coming in from the timelike infinity.

Secondly, corresponding to the fact that the geodesic normals from $t = 0$ all converge at p and q, all the past timelike geodesics from p expand out (normal to the surfaces $\{t = \text{constant}\}$) and reconverge at q. In fact, all the timelike geodesics from any point in this space (to either the past or future) reconverge to an image point, diverging again from this image point to refocus at a second image point, and so on. The future timelike geodesics from p therefore never reach $\mathscr{I}$, in contrast to the future null geodesics which go to $\mathscr{I}$ from p and form the boundary of the future of p. This separation of timelike and null geodesics results in the existence of regions in the future of p (i.e. which can be reached from p by a future-directed timelike line) which cannot be reached from p by any geodesic. The set of points which can be reached by future-directed timelike lines from p is the set of points lying beyond the future null cone of p; the set of points which can be reached from p by future-directed timelike geodesics is the interior of the infinite chain of diamond-shaped regions similar to that covered by coordinates (t, χ, θ, ϕ). One notes that all points in the Cauchy development of the surface $t = 0$ can be reached from this surface by a unique geodesic normal to this surface, but that a general point outside this Cauchy development cannot be reached by any geodesic normal to the surface.

5.3 Robertson–Walker spaces

So far, we have not considered the relation of exact solutions to the physical universe. Following Einstein, we can ask: can one find space-times which are exact solutions for some suitable form of matter and which give a good representation of the large scale properties of the observable universe? If so, we can claim to have a reasonable 'cosmological model' or model of the physical universe.

However we are not able to make cosmological models without some admixture of ideology. In the earliest cosmologies, man placed himself in a commanding position at the centre of the universe. Since the time of Copernicus we have been steadily demoted to a medium sized planet going round a medium sized star on the outer edge of a fairly average galaxy, which is itself simply one of a local group of galaxies. Indeed we are now so democratic that we would not claim that our position in space is specially distinguished in any way. We shall, following Bondi (1960), call this assumption the *Copernican principle*.

A reasonable interpretation of this somewhat vague principle is to understand it as implying that, when viewed on a suitable scale, the universe is approximately spatially homogeneous.

By spatially homogeneous, we mean there is a group of isometries which acts freely on $\mathscr{M}$, and whose surfaces of transitivity are spacelike three-surfaces; in other words, any point on one of these surfaces is equivalent to any other point on the same surface. Of course, the universe is not exactly spatially homogeneous; there are local irregularities, such as stars and galaxies. Nevertheless it might seem reasonable to suppose that the universe is spatially homogeneous on a large enough scale.

While one can build mathematical models fulfilling this requirement of homogeneity (see next section), it is difficult to test homogeneity directly by observation, as there is no simple way of measuring the separation between us and distant objects. This difficulty is eased by the fact that we can, in principle, fairly easily observe *isotropies* in extragalactic observations (i.e. we can see if these observations are the same in different directions, or not), and isotropies are closely con-