BULL. AUSTRAL. MATH. SOC.
VOL. 43 (1991) [355–363]

SIGNAL SPACES—AN AXIOMATIC APPROACH TO SPACE-TIME

PETER SZEKERES

Dedicated to my father, George Szekeres, on his eightieth birthday

An axiomatic approach to relativity theory is proposed which is based entirely on a single reflexive relation called signalling. This generalises and simplifies earlier schemes of Kronheimer and Penrose and of Carter which are based on causality relations. Many of the features of space-times, such as photon paths and topology, have their counterparts in general signal spaces.

1. CASUAL SPACES

There is a school of thought [7] which regards Euclidean geometry as an excellent piece of mathematical physics, that is, an axiomatic system which applies to the real world. It was in something of this spirit in 1968 that George Szekeres, following a suggestion of M.L. Urquhart, attempted to give an axiomatic foundation of the kinematic part of Einstein's special theory of relativity [9]. This work was subsequently enlarged on by John Schutz [8].

In a related vein there is Kronheimer and Penrose's axiomatic approach to space and time based entirely on causal relations [6]. This work is motivated more by general relativity than special, and attempts to encompass some of the unusual causal properties which can arise in that theory. However it does allow for structures more general than manifolds, and may eventually pave the way to a theory of discrete space-times which could be reconciled with the paradoxes of quantum theory.

One thing which is not permitted in the Kronheimer-Penrose theory (hereinafter referred to as KP) is closed timelike lines such as occur in the Gödel universe [3]. Carter [1] has extended the idea of a causal space to include the possibility of such causal loops. The resulting structure is one he calls an *etiological space*. In this paper I shall adopt a less arcane terminology.

By a *pre-causal space* we shall mean a triple $(X, \prec, \ll)$ where X is a set endowed with two relations $\prec$ (*causality*) and $\ll$ (*chronology*) satisfying the following axioms:

(C1) $x \prec x$.

Received 9 January 1991

(C2) If $x \prec y$ and $y \prec z$ then $x \prec z$.

(C3) If $x \ll y$ then there exists $z \in X$ such that $z \neq x$, $z \neq y$ and $x \prec z \prec y$.

(C4) If $x \ll y$, $y \prec z$ or $x \prec y$, $y \ll z$ then $x \ll z$.

Axioms (C1) and (C2) say that causality is reflexive and transitive. (C3) says that chronology only operates through distinct points. Condition (C4), called *strong transitivity*, applies in general to spacetimes (where $x \prec y$ means there is a future-pointing causal (timelike or null) curve from x to y, and $x \ll y$ means there is a timelike curve from x to y), although Carter points out that it may be violated if the usual requirement of C^1 differentiability is relaxed. In place of (C3) Carter uses the slightly weaker requirement that a point x can only be chronologically related to itself through a distinct point, but apart from this minor difference our definition of a pre-causal space is identical with Carter's etiological space. Our axiom makes an important fundamental difference between chronology and causality in that the latter can be "atomic" (that is, not broken down into finer connections) while the former can never be. Using (C2) and (C3) it follows at once that chronology implies causality, that is,

$$x \ll y \text{ implies } x \prec y.$$

Finally, the space will be called a *causal space* if it satisfies the further axiom

(C5) If $x \prec y$ and $y \prec x$ then $x = y$.

With a minor modification, this is essentially the same definition as used by KP. (C3) and (C5) imply immediately that chronology never reflects, that is,

$$x \ll x \text{ for no } x \in X.$$

KP also include a third relation $\rightarrow$ called *horismos*, in their definition. It is defined entirely in terms of the other two by

(H) $x \rightarrow y$ if and only if $x \prec y$ and not $x \ll y$.

Although it adds nothing fundamentally new to the space, its importance lies in that it demarcates the border between causality and chronology. In space-times chronology is defined as connectibility by a timelike curve (path of a material particle) while causality allows null paths (photons) as well. In Minkowski space (special relativity) horismos is simply the relation of being connected by a light signal, but in curved space-times null cones can develop all sorts of pathologies such as caustics and focal points, making the situation much more complicated—in general $x \rightarrow y$ does imply that there is a light signal from x to y, but the converse is not true. The following statement, called *horismoidal completeness*, follows immediately from strong transitivity (C4):

(HC) If $x \prec y \prec z$ and $x \rightarrow z$ then $x \rightarrow y$ and $y \rightarrow z$.

https://doi.org/10.1017/S0004972700029191 Published online by Cambridge University Press

Following Carter [1] with minor modifications, we define a subset Y of a pre-causal space X to be *vicious* if it contains at least two points and both $x \prec y$ and $y \prec x$ hold for all pairs of points $x, y \in Y$. A subset Z will be called *virtuous* if the causality condition (C5) holds for all pairs of points $x, y \in Z$.

THEOREM 1. *A pre-causal space is a disjoint union of maximal vicious sets and a remainder which is virtuous.*

PROOF: Given two points $x, y \in X$ we write $x \equiv y$ if $x \prec y$ and $y \prec x$. This is an equivalence relation on X. The equivalence classes consisting of two or more points are disjoint maximal vicious sets, while the remainder consisting of all those equivalence classes which are singletons forms a virtuous set. □

Finally a few words concerning order relations. A transitive relation $<$ on a space X is sometimes called a *pre-ordering*, while if it further satisfies the anti-symmetric condition (C5) it is called a *partial order*. If for any pair of points $x, y \in X$ one has either $x < y$ or $y < x$ then the relation is called a *total pre-order*, while if it is a partial order and total it is called a *simple order*. If $<$ is a non-reflexive partial order (that is, for no x is $x < x$), the *order topology* is the coarsest topology such that the sets $I^+(x) = \{y; y > x\}$ and $I^-(x) = \{y; y < x\}$ are open. If $<$ is a simple order the topology is called the *interval topology* and is Hausdorff.

2. SIGNAL SPACES

In the foregoing two relations $\prec, \ll$ invoking four or five axioms were required. Furthermore, despite this complexity, the horismos relationship does not provide a suitable basis in general relativity for the concept of a light signal. One is thus led to enquire whether a much simpler scheme could not be found in which the physical notion of a light signal plays the central role. To this end we define a *signal space* as a pair (X, S) consisting of a set X and a reflexive binary relation S. S is called the *signalling relation* and we write xSy to signify that x signals y with a light ray. Note that as in relativity the points x and y should be called *events*, and should not be confused with observers. The single axiom is the requirement of reflexivity

(R) xSx for all $x \in X$.

It is hard to conceive of a simpler mathematical structure than this, yet it immediately imposes on X a pre-causal structure as follows.

First define a causality relation $\prec_S$ by

> $x \prec_S y$ if and only if there exists a sequence of events $x_1 = x, x_2, \ldots, x_n = y$ such that
>
> $$x_i S x_{i+1} \qquad (i = 1, 2, \ldots, n-1).$$

https://doi.org/10.1017/S0004972700029191 Published online by Cambridge University Press

Such a sequence is called a *signal sequence* from x to y and n is its *length.* A signal sequence is called *complete* if $x_i S x_j$ for all $1 \leqslant i \leqslant j \leqslant n$. We define a relation $\rightarrow_S$ by

$x \rightarrow_S y$ if and only if $x \prec_S y$ and every signal sequence from x to y is complete.

Note how the horismoidal completeness relation (HC) follows immediately from this definition, and also that $x \rightarrow_S y$ implies xSy as required in space-times.

Finally define

$$x \ll_S y \text{ if and only if } x \prec_S y \text{ and not } x \rightarrow_S y.$$

This chronology relation is equivalent to saying that there exists an incomplete signal sequence from x to y. The space $(X, \prec_S, \ll_S)$ is readily shown to satisfy axioms (C1–C4) and therefore forms a pre-causal space, said to be *generated* by S. If this space is causal, that is, satisfies (C5), then the signal relation S is said to be *causal*—it is equivalent to the existence of no non-trivial signal loops, that is,

$$x = x_1 S x_2 S \ldots S x_n = x \text{ implies } x_i = x_j \text{ for all } i, j.$$

How much is lost by restricting oneself to signal-generated pre-causal spaces? Very little it would seem. All space-times are signal-generated, as may be seen by the fact that any timelike line can be approximated by a zig-zagging sequence of null geodesics (known by the descriptive term "zitterbewegung"). Also all $\mathcal{A}$-spaces (defined by KP to be those causal spaces in which every pair of points $x \prec y$ can be connected by a horismoidal chain $x = x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow x_n = y$) are signal-generated by setting xSy if and only if $x \rightarrow y$. Furthermore, every pre-causal space can be "embedded" in a signal-generated space as the following theorem shows.

THEOREM 2. *Let $(X, \prec, \ll)$ be a pre-causal space. Then there exists a signal space (Y, S) such that $X \subseteq Y$ and $\prec, \ll$ are precisely the relations $\prec_S$ and $\ll_S$ restricted to the subset X.*

PROOF: Set xSy for $x, y \in X$ if and only if $x \rightarrow y$. If $x \ll y$ and there is no horismoidal chain $x = x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow x_n = y$ then define a new point P_{xy} and set

$$xSP_{xy},\ ySP_{xy} \text{ and } P_{xy}SP_{xy}.$$

Let Y be the set consisting of the union of X and all such added points P_{xy}. Then (Y, S) is clearly a signal space and it is easily shown for any pair of points $x, y \in X$ that

$$x \prec_S y \text{ if and only if } x \prec y, \text{ and } x \ll_S y \text{ if and only if } x \ll y.$$

□

Quite possibly the more powerful statement that all pre-causal spaces can be signal-generated is true. It may be shown equivalent to the following set-theoretical

statement—let $(X, \leqslant)$ be a partially ordered set which is dense in itself, then there exists a subset D such that D and $\sim D$ are dense in each other. [A subset Y is said to be *dense* in a subset Z if for all $x, z \in Z$ there exists $y \in Y$ such that $x \leqslant y \leqslant z$.]

3. Rays

Let (X, S) be a signal space. A subset Y of X will be called *signal complete* if for every pair of points $x, y \in Y$ either xSy or ySx. A maximal signal complete set (that is, one which is not a proper subset of any other signal complete set) will be called a *ray*.

Theorem 3. *Every signal complete set is contained in a ray.*

Proof: Let C be a signal complete set. Let $\mathcal{C}$ be the set of all signal complete supersets of C, partially ordered by set inclusion. Every linearly ordered subset $\mathcal{A}$ of $\mathcal{C}$ is bounded above by $A = \bigcup \mathcal{A}$, which is clearly signal complete. Hence by Zorn's lemma, $\mathcal{C}$ contains a maximal element, which will be a ray containing C. □

Define the *future* and *past signal cones* of $x \in X$ to be

$$S^+(x) = \{y \in X; xSy\} \text{ and } S^-(x) = \{y \in X; ySx\}$$

respectively. The *signal cone* of x is their union $S(x) = S^+(x) \cup S^-(x)$. If Y is any signal complete subset of X we set

$$S(Y) = \bigcap_{y \in Y} S(y).$$

That is, $S(Y)$ is the set of points which are signal connected to all points of Y.

Theorem 4. *A signal complete set Y is contained in a unique ray if and only if $S(Y)$ is signal complete.*

Proof: If $S(Y)$ is signal complete then it is clearly maximal, and since every ray containing Y must be contained in $S(Y)$, it is the unique ray containing Y. Conversely if $S(Y)$ is not signal complete then there exists $x, y \in S(Y) - Y$ such that neither xSy or ySx. $Y \cup \{x\}$ and $Y \cup \{y\}$ are signal complete sets which are contained in rays R_1 and R_2. Clearly $R_1 \neq R_2$ since $x \in R_1$ and $x \notin R_2$. □

If $S(Y)$ is signal complete we say the ray $R = S(Y)$ is *generated* by Y. We set

$$\langle a, b\rangle = \{x \in X; aSx \text{ and } xSb\}.$$

If R is a ray and $a, b \in R$ with aSb then we define the *ray segment* from a to b to be

$$R(a, b) = R \cap \langle a, b\rangle.$$

https://doi.org/10.1017/S0004972700029191 Published online by Cambridge University Press

We will call the ray R *regular* if for any pair of points $a, b \in R$, R is either generated by one of the ray segments $R(a, b)$ or $R(b, a)$.

The notion of a regular ray lends itself to the physical interpretation of a ray being the path of a signalling particle (photon), a role occupied by null geodesics in general relativity. It is not hard to construct space-times having non-regular rays. For example if there exist three events a, b, c such that there are separate null geodesics from a to b, b to c, and a to c respectively (such a set of three points is easily found for example in the Einstein static universe), then the three points $R = \{a, b, c\}$ will form a ray (if no other event is signal-connected simultaneously to each of them), but it will not be regular since it is not the unique ray generated by $\{a, b\} = R(a, b)$.

One of the key properties of space-times is that if two signal-connected points are close enough ("proximate" in a sense to be defined below) then there is a unique geodesic connecting them. We accordingly define a signal space as *regular* if for any two points a and b such that aSb there exists a regular ray R which is generated by $R(a, b)$. Our definition of regularity discounts certain rays which arise in a more or less accidental fashion, such as the above example of a non-regular ray. Also, in the event of both aSb and bSa the ray may only be regular for one of these pairs—to achieve the reverse direction a new regular ray may have to be found.

A comparable notion of regularity is provided by KP entirely in terms of horismos. It is not hard to show that our notion of regularity implies KP regularity, but in a sense our concept is more general since it applies to pre-causal spaces, whereas KP regularity is only applicable to causal spaces.

From KP we take the following definition: two points x and y will be called *proximate* if $x \to y$ and there exists a point u such that either ($x \to u$ and $y \to u$) or ($u \to x$ and $u \to y$), that is, x and y form the "short side" of a horismoidal triangle. This is just another way of saying that the points are close enough that their horismoidal connection can be continued at least in one direction.

THEOREM 5. *Let (X, S) be a regular signal space, and $a, b \in X$ a pair of proximate points. Then there exists a unique regular ray through a and b.*

PROOF: Suppose $a \to b$ and there is a point c such that $a \to c$ and $b \to c$ (the proof is identical if c horismoidally precedes a and b). By regularity there exists a ray R which is generated by $R(b, c)$. Let x_1, x_2 be any pair of points in $\langle a, b \rangle$. Then

$$a \prec x_i \prec c \text{ for } i = 1, 2$$

and since $a \to c$ we have by (HC) that

$$a \to x_i \to c \text{ for } i = 1, 2.$$

https://doi.org/10.1017/S0004972700029191 Published online by Cambridge University Press

Now let w be any point in $R(b, c)$. Then for each $i = 1, 2$ we have $x_i \prec b \prec w \prec c$, and since $x_i \rightarrow c$ we have by (HC) that $x_i \rightarrow w$. Thus $x_i S w$ and it follows that both x_1 and x_2 belong to $S(R(b, c)) = R$. Since R is ray signal complete we must have that either $x_1 S x_2$ or $x_2 S x_1$. Hence $\langle a, b \rangle$ is signal complete and R is the unique regular ray through a and b. □

4. Topology

Various possibilities suggest themselves for defining a topology on a signal space in a natural way.

Alexandroff topology.

In a causal space X the *Alexandroff topology* is defined as the coarsest (smallest) topology such that the sets $I^+(x) = \{y; y \gg x\}$ and $I^- = \{y \ll x\}$ are open. This is essentially the interval topology defined at the end of Section 1 with respect to chronology as a partial order. This topology is quite unsuitable for general pre-causal spaces, and even in causal spaces it need not be Hausdorff. In space-times it leads to the manifold topology if and only if it is Hausdorff [4]. In such cases the space-time is called *strongly causal.*

Zeeman topology.

A different approach to the topology of Minkowski space has been given by Zeeman [10], based purely on its causal properties. Zeeman type topologies are defined by the topologies they induce on certain subsets, for example, the interval topology on straight timelike lines. This idea can, with modifications, be extended to space-times in general [2, 5]. This type of topology does not generally reproduce the manifold topology in space-times (usually it is too fine), but has interesting properties in its own right.

For signal spaces, particularly regular signal spaces, the construct might proceed as follows. On any regular ray R one can define a relation $a \leqslant_R b$ as holding if and only if aSb and R is generated by $R(a, b)$. We say the ray is *oriented* if

$$a \leqslant_R b \text{ and } b \leqslant_R c \text{ implies } a \leqslant_R c.$$

If this holds then $\leqslant_R$ is a pre-order on R. In causal spaces all rays are orientable and the pre-order is a partial order. When the space is regular the partial order becomes a simple order. In general a Zeeman type topology is available when the space is regular and orientable. It can be defined as the finest topology which induces a subtopology of the interval topology on all regular rays. Although this topology is frequently Hausdorff even in non-causal signal spaces, problems arise when a regular ray R exists such that aSb and bSa for all $a, b \in R$. Such "signal loops" have no preferred orientations and

every open set would be required to contain the whole ray if it contained a single point of it.

Such topologies are in general finer than the manifold topology in the case of space-times. For example in Minkowski space any open neighbourhood U of the origin with all events of the form $(x = 0, y = 0, z = 0, t > 0)$ removed is open in our Zeeman topology but is clearly not open in the manifold topology.

INFRASTRUCTURE TOPOLOGIES.

Carter [1] has introduced the notion of a causal infrastructure. This is a way of providing a causal structure $(U, \prec_U, \ll_U)$ for every subset U which is a substructure of $(X, \prec, \ll)$. We define a *signal infrastructure* on a signal space (X, S) to consist of a signal relation S_U for every subset $U \subseteq X$ such that

(IS1) If $U \subseteq V$ then aS_Ub implies aS_Vb.

(IS2) If aS_Vb and $V \supseteq U \supseteq <a, b>_V = \{x; aS_Vx \text{ and } xS_Vb\}$ then aS_Ub.

These two axioms imply, by Zorn's lemma, that any pair of points such that aSb are contained in a minimal set M (not necessarily unique) such that aS_Mb. We impose the following axiom on such minimal sets:

(IS3) If aSb and M is a minimal set connecting a and b then M is S_M-complete.

Hence minimal sets are always part of a ray connecting a and b, and (IS3) avoids the need for imposing regularity as a way of modelling light signals. The path of a photon from a to b can be regarded as proceeding along a minimal set.

A Zeeman type topology can be defined as the finest topology which induces a subtopology of the interval topology (with respect to the simple order S) on all causal minimal sets. Whenever regular rays can be covered by causal minimal sets, which is certainly true in space-times for example, then we essentially recover the Zeeman topology given above.

To obtain a topology closer to the manifold topology in the case of space-times it seems necessary to impose a number of restrictions on the signal space and its infrastructure. We shall call a subset U *normal* if it has the following properties:

(N1) The infrastructure signal relation S_U is causal.

(N2) If $a \in U$ then every set of the form $R^{\pm}(a) = R \cap S^{\pm}(a)$ where R is a regular ray through a (that is, a ray "starting" or "ending" at a) intersects U in a point distinct from a.

(N3) If $a \ll_U b$ then every regular S_U-ray through a meets $S(b)$ in a unique point.

These properties all hold for suitable normal geodesic neighbourhoods of points in space-times and are therefore minimal requirements. We will say the signal space is *normal*

if every point a has a normal neighbourhood U. We can define a topology having a subbase consisting of all subsets of the form

$$\ll a, b \gg_U = \{x; a \ll_U x \text{ and } x \ll_U b\}$$

where U is any normal subset. This topology is clearly a refinement of the Alexandroff topology. It can be shown to be Hausdorff when the signal space is normal, and reduces to the manifold topology in the case of space-times (even when not causal).

It still remains an open question exactly what conditions must be imposed on a signal space in order to recover completely the notion of a space-time, and indeed the whole of general relativity.

References

[1] B. Carter, 'Causal structure in space-time', *Gen. Relativity Gravitation* **1** (1971), 349–391.

[2] R. Göbel, 'Zeeman topologies on space-times of general relativity theory', *Comm. Math. Phys.* **46** (1976), 289-307.

[3] K. Gödel, 'An example of a new type of cosmological solution of Einstein's field equations of gravitation', *Rev. Modern Phys.* **21** (1949), 447–450.

[4] S.W. Hawking and G.F.R. Ellis, *The large scale structure of space-time* (Cambridge University Press, 1973).

[5] S.W. Hawking, A.R. King and P.J. McCarthy, 'A new topology for curved space-time which incoroprates the Causal differential and conformal structures', *J. Math. Phys.* **17** (1976), 174–181.

[6] E.H. Kronheimer and R. Penrose, 'On the structure of Causal spaces', in *Proc. Camb. Phil. Soc.* **63**, pp. 481–501, 1967.

[7] R. Penrose, *The Emperor's new mind*, pp. 204 (Oxford University Press, 1989).

[8] J.W. Schutz, *Foundations of special relativity: kinematic axioms for Minkowski space-time*: *Lecture Notes in Mathematics* 361 (Springer-Verlag, Berlin, Heidelberg, New York, 1973).

[9] G.Szekeres, 'Kinematic geometry; and axiomatic system for Minkowski space-time', *J. Austral. Math. Soc.* **8** (1968), 134–160.

[10] E.C. Zeeman, 'The topology of Minkowski space', *Topology* **6** (1967), 161–170.

Department of Physics
and Mathematical Physics
University of Adelaide
GPO Box 498
Adelaide SA 5001
Australia