Some Reflections on the Emergence of Space-filling Curves: The Way it Could Have Happened and Should Have Happened, but did not Happen

by HANS SAGAN
North Carolina State University, Department of Mathematics, Raleigh, NC 27695-8205, U.S.A.

To the Memory of Isaac J. Schoenberg (1903–1990)

DEFINITION: Φ*:* $[0, 1] \to E^n, n \geq 2$, *represents a space-filling curve if* Φ *is continuous and the image* $\Phi([0, 1])$ *is an* n*-dimensional region with positive content (such as a square for* $n = 2$).

1990 marked the 100th anniversary of the discovery of the first such curve.

I. Introduction

Three significant events preceded the emergence of space filling curves:

(1) In 1878, G. Cantor startled the mathematical world by demonstrating that any two finite-dimensional smooth manifolds, no matter what their dimensions, have the same cardinality (**1**). In particular, the interval $[0, 1]$ and the unit-square $[0, 1]^2$ have the same cardinality, meaning that there exists a bi-jective map ϕ from $[0, 1]$ onto $[0, 1]^2$

$$\phi: [0, 1] \leftrightarrow [0, 1]^2. \tag{1}$$

(2) One year later, E. Netto showed that such a mapping cannot possibly be continuous ["A bi-jective map from a one dimensional manifold onto a two-dimensional manifold is, by necessity, discontinuous" (**2**)]. If ϕ in (1) were continuous, then, with $[0, 1]$ being compact, ϕ^{-1} would also be continuous and map connected sets onto connected sets. The removal of a point t_0 from the interior of $[0, 1]$ and its corresponding image $\Phi(t_0)$ from $[0, 1]^2$ disconnects the interval but not the square and the continuous function ϕ^{-1} would map the connected punctured square onto the disconnected set $[0, t_0) \cup (t_0, 1]$. More generally, manifolds of different dimensions cannot be homeomorphic (**3**).

(3) Finally, in 1883, G. Cantor introduced the set Γ of all points that have the representation

$$\frac{2a_1}{3}+\frac{2a_2}{3^2}+\frac{2a_3}{3^3}+\frac{2a_4}{3^4}+\cdots$$

where $a_i = 0$ or 1 (4, p. 590) which is now called the Cantor set or the "set of the excluded middle thirds". Cantor offers this set in an appendix to his paper on infinite linear point manifolds as an example of a perfect set (a set that is equal to the set of all its accumulation points) which is in no interval, no matter how small, everywhere dense. (In this paper, Cantor was groping for a characterization of the continuum to wind up defining it as a perfect connected set.) While Cantor does not mention it explicitly, it could not possibly have escaped him that this set, being equivalent to the set of all sequences of 0's and 1's, has the same cardinality as the interval [0, 1], i.e. there is a bi-jective map

$$\psi : \Gamma \leftrightarrow [0, 1]. \tag{2}$$

Putting (1) and (2) together, we have

$$\varphi \circ \psi : \Gamma \leftrightarrow [0, 1]^2. \tag{3}$$

It would appear that at this point in time, the discovery of space-filling curves was inevitable. Since a bi-jective map from [0, 1] onto $[0, 1]^2$ cannot be continuous and hence, a continuous map cannot be bi-jective, all that is needed is to extend some convenient continuous surjective map Φ from Γ onto $[0, 1]^2$ into the countably many disjoint open sets that make up the complement of the Cantor set—see Appendix (I)—so that the extension is continuous in [0, 1]. This is how it should have happened but it is not how it came about. What really happened at that time will be related in Section III.

II. The Lebesgue and Schoenberg Curves

It was not until 21 years later, in 1904, that H. Lebesgue finally followed through on this idea. Preparatory to obtaining the required continuous surjective map from Γ onto $[0, 1]^2$, let us adopt the convention that in the binary representation of numbers in [0, 1] every finite binary—such as $0_{\dot{2}}01$—be replaced by the corresponding infinite repeating binary—such as $0_{\dot{2}}00\bar{1}$, where ${}_{\dot{2}}$ represents the binary point. With ${}_{\dot{3}}$ representing the ternary point and $a_i = 0$ or 1, we define the following mapping on Γ:

$$\Phi(0_{\dot{3}}(2a_1)(2a_2)(2a_3)(2a_4)\ldots) = (0_{\dot{2}}a_1a_3a_5\ldots, 0_{\dot{2}}a_2a_4a_6\ldots) \tag{4}$$

which satisfies our requirements. [For the proof of continuity, see (5), p. 363. More generally, every compact set is the continuous image of a dyadic discontinuum, as, for example, the Cantor set (6).] Lebesgue extended the definition simply by linear interpolation (7) to obtain an extension Φ_l of Φ into [0, 1]. Take, for example, $t \in (1/9, 2/9)$. Since $1/9 = 0_{\dot{3}}00\bar{2}$ and $2/9 = 0_{\dot{3}}02$, we have from (4) that

$$\Phi(1/9) = \left(0_2 0\bar{1}, 0_2 0\bar{1}\right) = (1/2, 1/2)$$

and

$$\Phi(2/9) = \left(0_2 0, 0_2 1\right) = (0, 1/2).$$

Hence, for $t \in (1/9, 2/9)$,

$$\Phi_l(t) = (-9t/2+1, 1/2).$$

Since the restriction of Φ_l to Γ is already onto $[0,1]^2$, it follows that

$$\Phi_l : [0,1] \xrightarrow{\text{onto}} [0,1]^2.$$

By construction, the restriction of Φ_l to the complement of the Cantor set is continuous, and its restriction to the Cantor set is also continuous. To show that it is continuous on [0, 1] requires some insight into the structure of the Cantor set. For a proof, we refer the reader to (8). Since the Cantor set has Lebesgue measure zero [see Appendix (II)], Φ_l is differentiable a.e.

In 1938, while proctoring a 2-hour mechanics examination at Colby College (9), I. J. Schoenberg, by contrast, extended the definition of Φ in such a manner that the continuity of the mapping became obvious (10). He defined the components (f_s, g_s) of a function Φ_s by

$$f_s(t) = \sum_{k=1}^{\infty} p(3^{2k-2}t)/2^k, \quad g_s(t) = \sum_{k=1}^{\infty} p(3^{2k-1}t)/2^k \tag{5}$$

where

$$p(t) = \begin{cases} 0 & \text{for } 0 \leqslant t < 1/3 \\ 3t-1 & \text{for } 1/3 \leqslant t < 2/3 \\ 1 & \text{for } 2/3 \leqslant t \leqslant 1 \end{cases} \tag{6}$$

with the provision that $p(-t) = p(t)$ and $p(t+2) = p(t)$ (see Fig. 1).

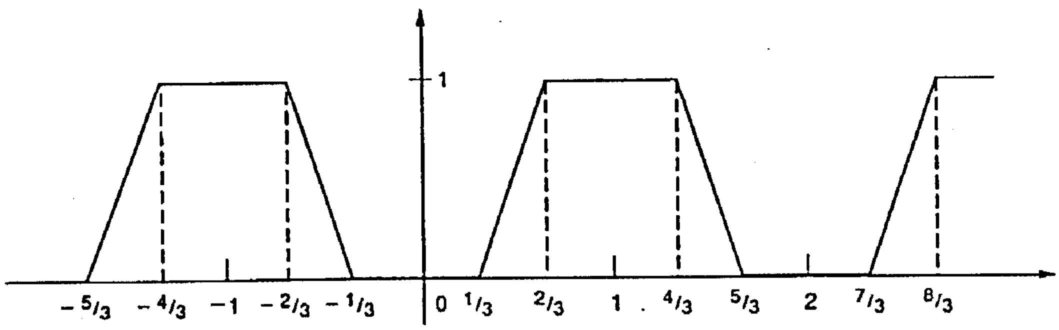

FIG. 1. Graph of generating function p.

f_s, g_s, being represented by uniformly convergent series of continuous functions, are continuous. Hence, Φ_s is continuous. If one restricts Φ_s to the Cantor set Γ, one obtains precisely the mapping in (4) [see Appendix (III)] and hence, Φ_s is also surjective and represents a space-filling curve. This is almost too easy and there is a penalty to be paid: While it is obvious that the Lebesgue curve is differentiable a.e., the differentiability properties of the Schoenberg curve are not that easy to discover. While it was assumed all along—at least by this author—that Schoenberg's Φ_s is nowhere differentiable, this was not proved until 1981 by J. Alsina (11). Alsina's proof is fairly complicated and laborious. A somewhat simpler proof [that $3f_s+g_s$ is nowhere differentiable whence the desired result follows in view of $g_s(t) = f_s(3t)$] can be found in (12) but a simple and elementary proof did not appear until very recently (13).

III. The Peano and Hilbert Curves

Let us now return to the end of the last century and relate how space-filling curves really entered the world of mathematics. G. Peano, who was familiar with Cantor's result in (1) on the equivalence of finite-dimensional smooth manifolds and E. Netto's result in (2) on the impossibility of a homeomorphic relationship between $[0,1]$ and $[0,1]^2$, was not cognizant of the Cantor set and its relationship to the interval $[0,1]$. This is not at all surprising. This author, knowing what to look for, had a very difficult time locating the Cantor set in Cantor's papers, buried as it is in an appendix to a lengthy article and in a context that is unrelated to mappings and such. So, it is understandable that Peano did not do the obvious and extend the mapping in (4) when looking for a continuous map from $[0,1]$ onto $[0,1]^2$ because he simply was not aware of it. In spite of all this, he only narrowly missed the mapping in (4) and, with it, an opportunity to re-discover the Cantor set, as we will see from his construction of the first space-filling curve ever.

Peano defined a mapping Φ_p from $[0,1]$ onto $[0,1]^2$ in the following manner: With the operator k defined on $\{0,1,2\}$ by $k0 = 2, k1 = 1, k2 = 0$ and with k^n denoting the nth iterate, and with $t \in [0,1]$ in ternary representation

$$t = 0_3 a_1 a_2 a_3 a_4 \ldots$$

he demonstrated that

$$\Phi_p(t) = (0_3 a_1 (k^{a_2} a_3)(k^{a_2+a_4} a_5)\ldots,\quad 0_3 (k^{a_1} a_2)(k^{a_1+a_3} a_4)\ldots) \tag{7}$$

maps $[0,1]$ continuously onto $[0,1]^2$ (14).

In order to show that Φ_p is surjective, we will demonstrate how to find for any point $P = (0_3 \beta_1 \beta_2 \beta_3 \ldots,\ 0_3 \gamma_1 \gamma_2 \gamma_3 \ldots) \in [0,1]^2$ a $t = 0_3 a_1 a_2 a_3 \ldots \in [0,1]$ such that $\Phi_p(t) = P$. By (7), we have to have

$$(0_3 a_1 (k^{a_2} a_3)(k^{a_2+a_4} a_5)\ldots,\quad 0_3 (k^{a_1} a_2)(k^{a_1+a_3} a_4)(k^{a_1+a_3+a_5} a_6)\ldots)$$
$$= (0_3 \beta_1 \beta_2 \beta_3 \ldots,\quad 0_3 \gamma_1 \gamma_2 \gamma_3 \ldots).$$

Comparing corresponding ternary places and noting that k is its own inverse, we obtain

$$\begin{aligned} a_1 &= \beta_1 \\ k^{a_1}a_2 &= \gamma_1 \quad \text{and hence} \quad a_2 = k^{a_1}\gamma_1 \\ k^{a_2}a_3 &= \beta_2 \quad \text{and hence} \quad a_3 = k^{a_2}\beta_2 \\ k^{a_1+a_3}a_4 &= \gamma_2 \quad \text{and hence} \quad a_4 = k^{a_1+a_3}\gamma_2, \quad \text{etc.} \end{aligned}$$

To show that Φ_p is also continuous, we proceed as follows: First, note that $|t_1-t_2| < 1/3^{n+1}$, where we assume without loss of generality that n is even, implies that t_1, t_2 have to agree up to and including the nth ternary place, i.e.

$$t_1 = 0_{\dot{3}}a_1a_2a_3\ldots a_n\beta_{n+1}\beta_{n+2}\ldots, \quad t_2 = 0_{\dot{3}}a_1a_2a_3\ldots a_n\gamma_{n+1}\gamma_{n+2}\ldots$$

With $a_2+a_4+\cdots+a_{2n} = a$,

$$\begin{aligned} |f_p(t_1)-f_p(t_2)| &= |0_{\dot{3}}a_1(k^{a_2}a_3)\ldots(k^{a}\beta_{n+1})\ldots - 0_{\dot{3}}a_1(k^{a_2}a_3)\ldots(k^{a}\gamma_{n+1})\ldots| \\ &\leqslant |k^{a}\beta_{n+1}-k^{a}\gamma_{n+1}|/3^{(n/2)+1} + |k^{a+\beta_{n+2}}\beta_{n+3}-k^{a+\gamma_{n+2}}\gamma_{n+3}|/3^{(n/2)+2}+\cdots \\ &\leqslant (2/3^{(n/2)+1})(1+1/3+1/9+\cdots) = 1/3^{n/2}. \end{aligned}$$

An analogous argument leads to the continuity of g_p, the second component of Φ_p.

When restricting Φ_p to the Cantor set Γ, we obtain in view of

$$k^{2n}0 = 0, \quad k^{2n}2 = 2, \quad \text{for} \quad n = 0, 1, 2, 3, \ldots$$

that

$$\Phi_p(0_{\dot{3}}a_1a_2a_3a_4\ldots) = (0_{\dot{3}}a_1a_3a_5\ldots, \quad 0_{\dot{3}}a_2a_4a_6\ldots)$$

because all the a_i are now 0 or 2 and hence, all iterates of k are even. So close—and still so far apart from (4)!

One year after Peano's discovery, D. Hilbert came forth with his own version of a space-filling curve (15). He was led to it by a repeated application of the following heuristic principle: If an interval I can be mapped onto a square, then, after partitioning I into four congruent subintervals, and the square into four congruent subsquares, each of the subintervals can be mapped onto one of the subsquares with adjacent subintervals mapped onto adjacent subsquares with a common edge. In Fig. 2, we have indicated the first three steps in this process where the polygonal lines indicate the order in which the subsquares may be taken in order to satisfy our requirement.

The mapping $\Phi_h: [0, 1] \to [0, 1]^2$ is now to be defined as follows: Every $t \in [0, 1]$ lies in a sequence of nested intervals the lengths of which shrink to 0. With this sequence corresponds a sequence of nested squares the diagonals of which shrink to 0 and hence, define a unique point in $[0, 1]^2$, the image $\Phi_h(t)$ of t. (If t is the endpoint of one of the subintervals—other than 0 or 1—it belongs to two different sequences of nested intervals. This leads to the same image, however, since adjacent subintervals are mapped onto adjacent subsquares.)

Vol. 328, No. 4, pp. 419–430, 1991
Printed in Great Britain

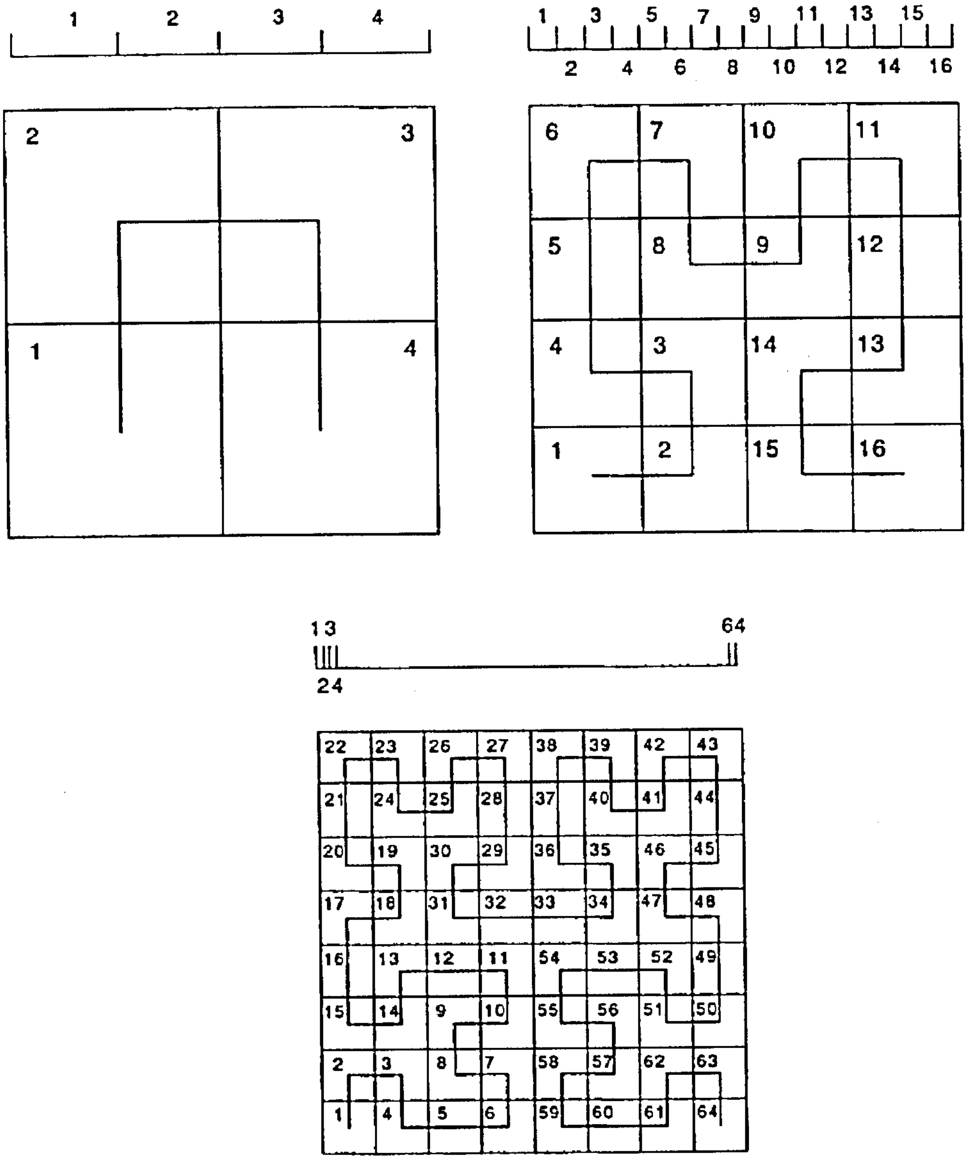

FIG. 2. Generating Hilbert's space-filling curve.

This mapping is surjective. Every point in $[0, 1]^2$ lies in a sequence of nested squares whose diameters shrink to zero and is, by construction, the image of that point in $[0, 1]$ that is defined by the corresponding sequence of nested intervals. (Note: if the point lies in a corner of a square, then it will, and if it lies on the edge of a square, then it may belong to at least two squares that are not consecutively numbered and it may be viewed as belonging to two or more sequences of nested squares with distinct pre-images in $[0, 1]$, leading to at least two values of t with the same image. The center of $[0, 1]^2$ lies in the corner of three non-adjacent squares!)

The mapping is continuous. Choose $t_1, t_2 \in [0, 1]$ so that $t_1 < t_2$ and $t_2 - t_1 < 1/4^n$. At the nth step, the interval $[0, 1]$ is partitioned into 4^n subintervals of length

$1/4^n$. Hence, the interval $[t_1, t_2]$ overlaps with at most two adjacent subintervals. Therefore, the images $\Phi_h(t_1)$ and $\Phi_h(t_2)$ lie, at worst, in two different, but adjacent squares of sidelength $1/2^n$ and their distance cannot exceed the length of the diagonal of the rectangle formed from two such squares:

$$|f_h(t_2)-f_h(t_1)|^2+|g_h(t_2)-g_h(t_1)|^2 \leqslant \left(\frac{2}{2^n}\right)^2+\left(\frac{1}{2^n}\right)^2 = \frac{5}{4^n}.$$

[If, in the generating process, one partitions at every step into nine congruent parts instead of four, and takes the subsquares in the order that is indicated in Fig. 4, one is led to Peano's curve (16, p. 5).]

Peano notes at the very end of his paper (14) without proof that "Ces x et y, fonctions continues de la variable t, manquent toujours de dérivée." ("These x and y,"—we call them f_p and g_p in this article—"continuous functions of the variable t, completely lack a derivative.") It was not until 10 years later that E. H. Moore published a proof in (17). It is, of course, impossible for us to guess how Peano proved this to his own satisfaction, but his proof could not possibly have been any simpler than the one we have to offer:

For $\quad t = 0._3 a_1a_2a_3\ldots a_{2n}a_{2n+1}a_{2n+2}\ldots \in [0,1]$ we define

$$t_n = 0._3 a_1a_2a_3\ldots a_{2n}\beta_{2n+1}a_{2n+2}\ldots$$

where $\beta_{2n+1} = a_{2n+1}+1 \pmod 2$. Then, $|t-t_n| = 1/3^{2n+1}$. By (7), $f_p(t)$ and $f_p(t_n)$ only differ in the $(n+1)$th ternary place and we have

$$|f_p(t)-f_p(t_n)| = |k^{a_2+\cdots+a_{2n}}a_{2n+1}-k^{a_2+\cdots+a_{2n}}\beta_{2n+1}|/3^{n+1} = 1/3^{n+1}.$$

Hence, $|(f_p(t)-f_p(t_n))/(t-t_n)| = 3^n \to \infty$. A similar argument applies to g_p.

Hilbert makes the same claim without proof in his paper (15): "Die ... Funktionen sind zugleich einfache Beispiele fuer ueberall stetige und nirgends differentiirbare Funktionen." ("The functions ... are, at the same time simple examples of everywhere continuous and nowhere differentiable functions.") On first glance one might have reservations about calling these examples "simple" when they are not even presented explicitly. It is, however, indeed easy to see that these functions are nowhere differentiable: For $n \geqslant 3$, pick for any $t\in[0,1]$ a $t_n\in[0,1]$ such that $|t-t_n| \leqslant 10/2^{2n}$ and the coordinates of the image $\Phi_h(t)$ are separated from the coordinates of the image $\Phi_h(t_n)$ by at least a square of sidelength $1/2^n$ (see Fig. 2). Then,

$$|(f_h(t)-f_h(t_n))/(t-t_n)| \geqslant 2^n/10$$

and

$$|(g_h(t)-g_h(t_n))/(t-t_n)| \geqslant 2^n/10.$$

IV. Concluding Remarks

Once the principles underlying the construction of space-filling curves were understood, such curves could be churned out at will. The following three generating procedures emerged from the preceding discussion:

First, one may start with the continuous surjective map from Γ onto $[0,1]^2$ in (4) and extend the map continuously into $[0,1]$, though it is hard to imagine how anyone could surpass the constructions of Lebesgue and Schoenberg in ingenuity.

Secondly, one may follow Peano's idea of choosing the image $(0\underset{b}{.}\beta_1\beta_2\beta_3\ldots,\ 0\underset{b}{.}\gamma_1\gamma_2\gamma_3\ldots)$ of $t = 0\underset{b}{.}a_1a_2a_3\ldots$ (for some base b) in such a manner that a change in $a_{n+1}, a_{n+2}, \ldots$ won't affect $\beta_1, \beta_2, \ldots, \beta_{k(n)}$ and $\gamma_1, \gamma_2, \ldots, \gamma_{k(n)}$, where $k(n) \nearrow \infty$ is a suitable integer valued function, to ensure continuity and so that for any point $(0\underset{b}{.}b_1b_2b_3, \ldots, 0\underset{b}{.}c_1c_2c_3, \ldots) \in [0,1]^2$, the infinitely many equations $\beta_i = b_i, \gamma_i = c_i, i = 1, 2, 3, \ldots$, can be solved successively for $a_1, a_2, a_3, \ldots$, to ensure that the map is surjective.

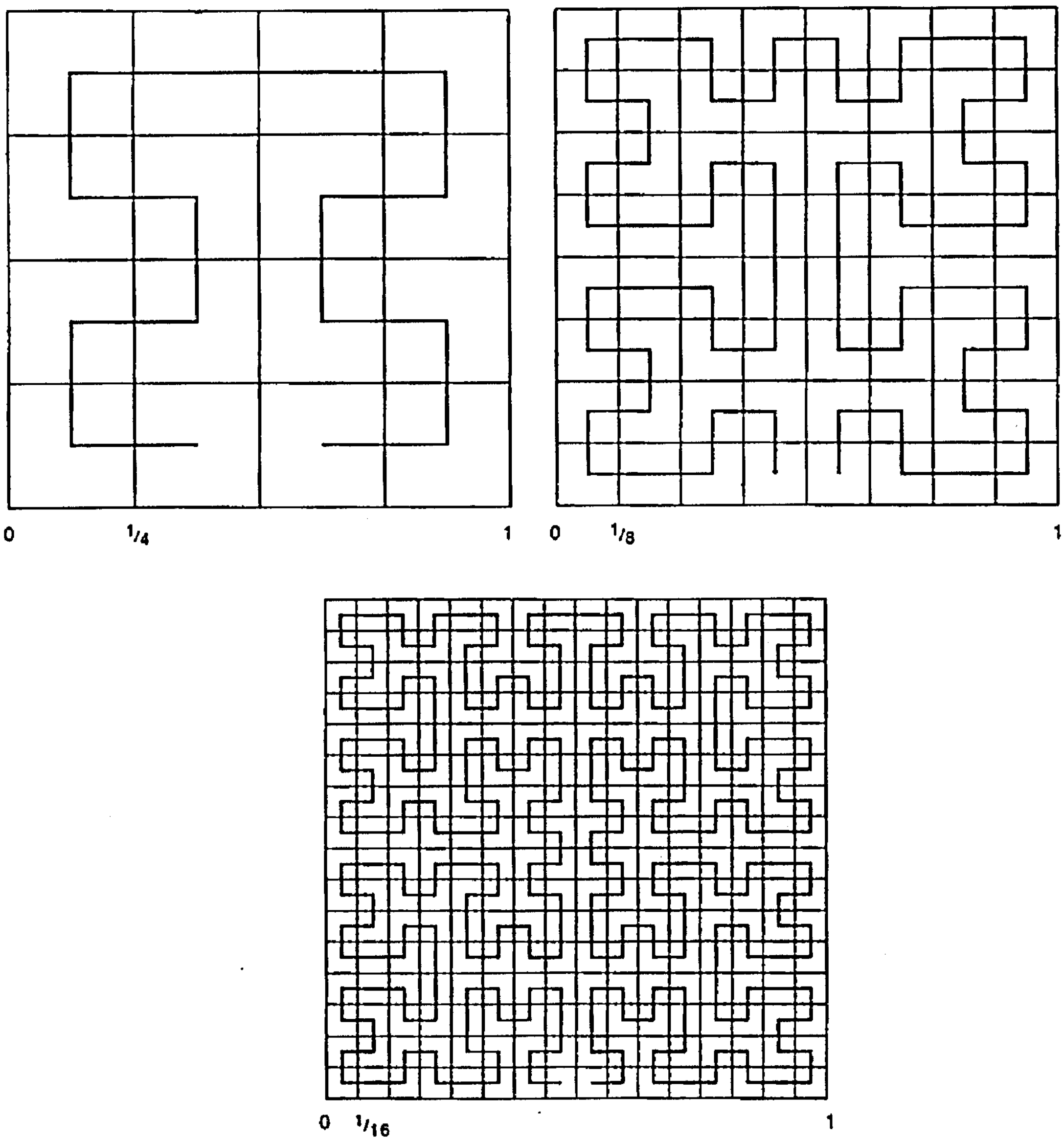

FIG. 3. Generating a Hilbert-type space-filling curve.

Thirdly, Hilbert's construction may be modified by changing the order of the subsquares as, for example in Fig. 3, or, by changing the number of subintervals and subsquares each interval and square are partitioned into at every step of the process, such as in Fig. 4 (which leads to Peano's curve), or, by changing the shape of the target set as we have illustrated in Fig. 5. [See also (**16–19**).]

With the square and, hence, all its continuous images revealed as continuous images of a line segment, the question arose as to the characterization of such sets. In 1913, this question was pondered by H. Hahn (**20**) and St. Mazurkiewicz (**21**). [This paper, consisting of two parts, is not easily accessible and is written in Polish. Those who read French are referred to (**22**) instead.] They arrived, independently of each other, at the following complete characterization of such sets: A set is the continuous image of a line segment if and only if it is compact, connected, and locally connected. This theorem is now known as the Hahn–Mazurkiewicz theorem. [For a more general version, see (**23**).]

Since compactness and connectedness are preserved under continuous mappings, and since a continuous mapping on a line segment is a closed continuous mapping

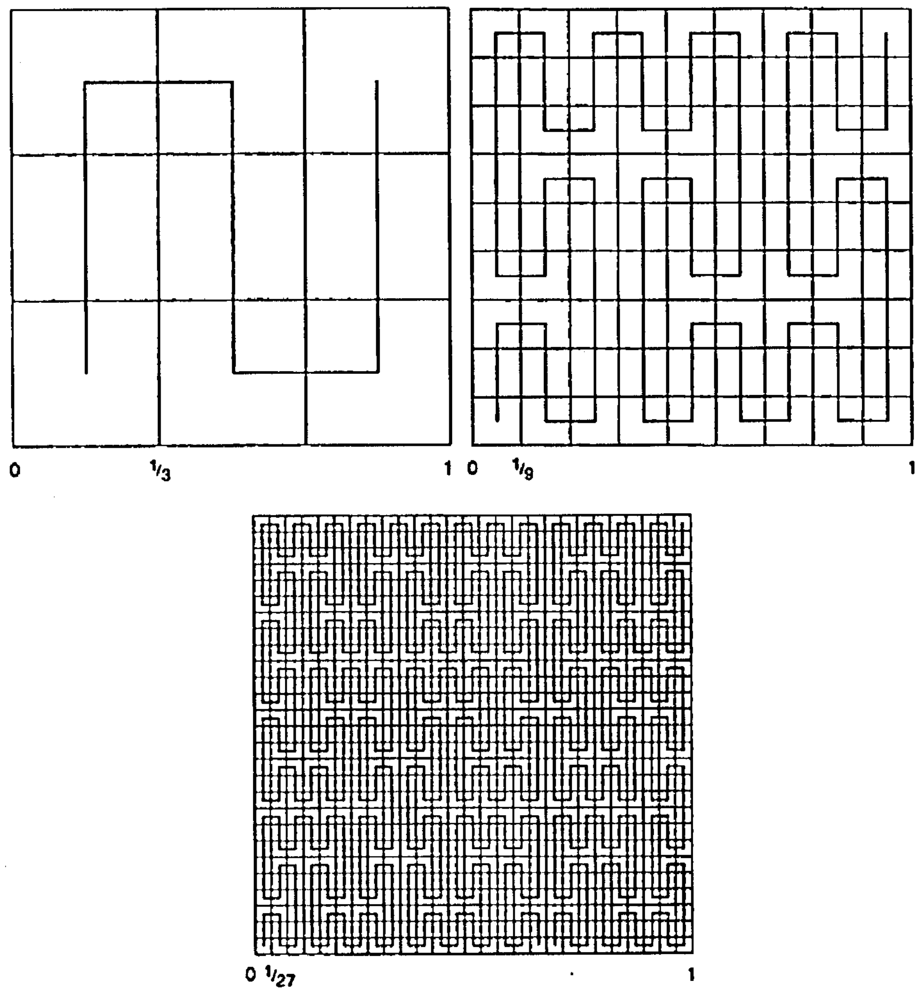

FIG. 4. Generating Peano's space-filling curve.

Vol. 328, No. 4, pp. 419–430, 1991
Printed in Great Britain

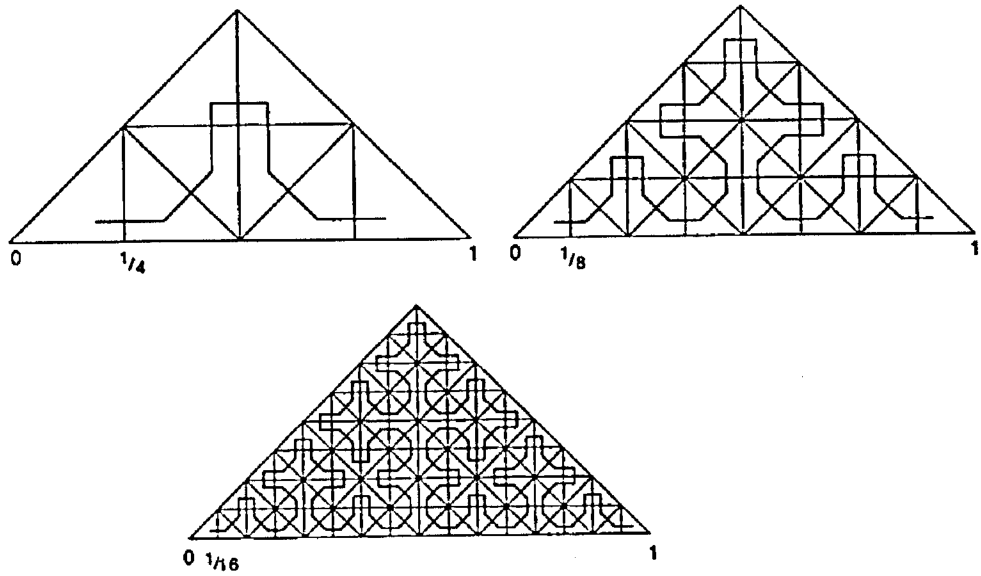

FIG. 5. Generating Sierpiński's space-filling curve.

which preserves local connectedness, the necessity of these conditions is obvious. It is not at all obvious that they are also sufficient. Hahn established the sufficiency (20) by generalizing Peano's construction of a space-filling curve. Fourteen years later, he offered a simpler and more lucid proof (24) by generalizing Lebesgue's construction.

References

(1) G. Cantor, "Ein Beitrag zur Mannigfaltigkeitslehre", *Crelle J.*, Vol. 84, pp. 242–258, 1878.

(2) E. Netto, "Beitrag zur Mannigfaltigkeitslehre", *Crelle J.*, Vol. 86, pp. 263–268, 1879.

(3) M. Greenberg, "Lectures on Algebraic Topology", p. 82, W. A. Benjamin, New York, 1967.

(4) G. Cantor, "Ueber unendliche lineare Punktmannigfaltigkeiten", *Math. Ann.*, Vol. 21, pp. 545–591, 1883.

(5) H. Sagan, "Approximating polygons for Lebesgue's and Schoenberg's space-filling curves", *Am. Math. Monthly*, Vol. 93, pp. 361–368, 1986.

(6) F. Hausdorff, "Set Theory", pp. 154, 220, Chelsea, New York, 1962.

(7) H. Lebesgue, "Leçons sur l'Integration et la Recherche des Fonctions Primitives", pp. 44–45, Gauthier-Villars, Paris, 1904.

(8) A. Devinatz, "Advanced Calculus", p. 253, Holt, Rinehart, Winston, New York, 1968.

(9) I. J. Schoenberg, Personal letter to the author, 7 May 1986.

(10) I. J. Schoenberg, "The Peano curve of Lebesgue", *Bull. Am. Math. Soc.*, Vol. 44, p. 519, 1938.

(11) J. Alsina, "The Peano curve of Schoenberg is nowhere differentiable", *J. Approx. Theory*, Vol. 33, pp. 28–42, 1981.

(12) I. J. Schoenberg, "Mathematical time exposures", *Math. Assoc. Am.*, pp. 135–148, 1982.

(13) H. Sagan, "An elementary proof that Schoenberg's space-filling curve is nowhere differentiable", *Math. Mag.* (in press).

(14) G. Peano, "Sur une courbe qui remplit toute une aire plane", *Math. Ann.*, Vol. 36, pp. 157–160, 1890.

(15) D. Hilbert, "Ueber die stetige Abbildung einer Linie auf ein Flaechenstueck", *Math. Ann.*, Vol. 38, pp. 459–460, 1891.

(16) W. Wunderlich, "Ueber Peano Kurven", *Elem. Math.*, Vol. 28, pp. 1–10, 1973.

(17) E. H. Moore, "On certain crinkly curves", *Trans. Am. Math. Soc.*, Vol. 1, pp. 72–90, 1900.

(18) G. Polya, "Ueber eine Peanosche Kurve", *Bull. Acad. Sci. Cracovie* (Sci. math. et nat., Série A), pp. 1–9, 1913.

(19) W. Sierpiński, "Sur une nouvelle courve continue qui remplit toute une aire plane", *Bull. Acad. Sci. de Cracovie* (Sci. math. et nat., Série A), pp. 193–219, 1912.

(20) H. Hahn, "Ueber die allgemeinste ebene Punktmenge, die stetiges Bild einer Strecke ist", *J. Ber. Deutsch. Math.-Verein.*, Vol. 23, pp. 318–322, 1914; "Mengentheoretische Charakterisierung der stetigen Kurven", *S. B. Kaiserl. Akad. Wiss. Wien*, Abt. 2a (Math), Vol. 123, pp. 2433–2487, 1914.

(21) St. Mazurkiewicz, "O arytmetyzacji kontinuów", *C. R. Soc. Sc. Varsovie*, Vol. VI, pp. 305–311, 1913; "O arytmetyzacji kontinuów II", *ibid.*, pp. 941–945.

(22) St. Mazurkiewicz, "Sur les lignes de Jordan", *Fund. Math.*, Vol. 1, pp. 166–209, 1920.

(23) St. Willard, "General Topology", pp. 219–222. Addison Wesley, Reading, Mass., 1970.

(24) H. Hahn, "Ueber stetige Streckenbilder", Atti del Congresso Internazionale dei Mathematici, Bologna, 3–10 Septembre 1928, VI, 217–220.

Appendix

(I) The Cantor set may be obtained from the closed interval [0, 1] by first removing the middle third $(\frac{1}{3}, \frac{2}{3})$, then the middle thirds of what remains, namely $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$, and again the middle thirds of what is left, $(\frac{1}{27}, \frac{2}{27})$, $(\frac{7}{27}, \frac{8}{27})$, $(\frac{19}{27}, \frac{20}{27})$, $(\frac{25}{27}, \frac{26}{27})$, and so on *ad infinitum*. (See also Fig. 6.)

If all numbers in [0, 1] are represented in ternary form

$$0_3 a_1 a_2 a_3 a_4 \ldots$$

where $a_i = 0$ or 1 or 2, then, by removing the middle third $(\frac{1}{3}, \frac{2}{3})$, all numbers $0_3 1 a_2 a_3 a_4 \ldots$ are removed except for $0_3 1$ which we rewrite as $0_3 0\bar{2}$. Next, we remove all numbers $0_3 01 a_3 a_4 a_5 \ldots, 0_3 21 a_3 a_4 a_5 \ldots$, except for $0_3 01$ and $0_3 21$ which we rewrite as $0_3 00\bar{2}$ and $0_3 20\bar{2}$, etc. Eventually, only those numbers are left that have a ternary representation with 0's and 2's only and every such number represents an element of the Cantor set.

FIG. 6. Generating the Cantor set.

Vol. 328, No. 4, pp. 419–430, 1991
Printed in Great Britain

(II) By construction, the Cantor set Γ can be covered by 2^n closed intervals of length $1/3^n$ each. These are the intervals that are left over after n consecutive removals of the middle thirds from what was left over after the preceding step. Since $(2/3)^n$ can be made arbitrarily small, Γ has Jordan content 0 and, consequently, Lebesgue measure 0.

(III) Let $t \in \Gamma$, i.e.

$$t = \frac{2a_1}{3} + \frac{2a_2}{3^2} + \frac{2a_3}{3^3} + \cdots, \quad a_i = 0 \text{ or } 1.$$

In order to evaluate f_s, as defined in (5), we observe that

$$3^{2k-2}t = 2a_1 3^{2k-3} + 2a_2 3^{2k-4} + \cdots + 2a_{2k-2} + \frac{2a_{2k-1}}{3} + \frac{2a_{2k}}{3^2} + \cdots$$

$$= \text{even integer} + \frac{2a_{2k-1}}{3} + \frac{2a_{2k}}{3^2} + \cdots$$

and we have from (6) that

$$p(3^{2k-2}t) = \begin{cases} 0 & \text{if } a_{2k-1} = 0 \\ 1 & \text{if } a_{2k-1} = 1 \end{cases}$$

because $(2a_{2k}/3^2) + (2a_{2k+1}/3^3) + \cdots \leqslant 1/3$, i.e.

$$p(e^{2k-2}t) = a_{2k-1}.$$

Consequently,

$$f_s(t) = \sum_{k=1}^{\infty} \frac{a_{2k-1}}{2^k} = 0_2 a_1 a_3 a_5 \ldots$$

Similarly, we obtain

$$g_s(t) = 0_2 a_2 a_4 a_6 \ldots$$

and we see that the mapping in (5) satisfies (4).