

What is a Curve?

Nicholas J. Rose

1 Introduction

What is a curve? All of us have used the word curve since early childhood; it evokes certain intuitive ideas in each of us such as ‘length without width’ and ‘one dimensional’. We can all agree that certain simple geometric objects such as circles ellipses, polygons and helices are curves while a sphere and the interior of a square are not curves.. To deal with more complicated geometric objects, it is necessary to have a satisfactory definition of a curve. Such a definition should include all objects that everyone agrees are curves and exclude all objects everyone agrees are not curves. In addition the definition should apply to more complicated geometric objects where intuition is not a guide. Finally, a satisfactory definition should yield significant consequences and connections with related ideas. It took mathematicians until the 1920’s to come up with a satisfactory definition of a curve. In this short essay we will sketch some of the highlights in the struggle to answer the question “what is a curve?” For a more complete mathematical treatment see Blumenthal and Menger, Chapter 12.

2 Curves in the Cantor Sense

Before Cantor’s pioneering work in the 1870’s it was thought that *dispersed sets* such as a finite set of points or the Cantor set differed from curves, surfaces or solids. (See Section (5) for a technical definition of a dispersed set) Solids seem to contain more points than surfaces, surfaces more than curves and curves more than dispersed sets. However Cantor showed that there is a one-to-one correspondence between all the points on a line segment and all the points in the interior of square or the interior of a cube. Even some dispersed sets, such as the Cantor set, can be put into one-to-one correspondence with a line segment. In this sense, all of the objects contain the same ‘quantity’ of points. Thus the concept of quantity of points cannot be used to distinguish curves from non-curves.

Although Cantor is often credited with giving a satisfactory definition of a plane curve, there seems to be little evidence to support this. However, Cantor did provide one of the essential ingredients for such a definition, namely the notion of a *continuum*. A continuum is a closed, bounded connected set that contains at least two points. Intuitively, a connected set is “all of one piece.” A line segment, a circle, a square (interior plus boundary) and

the solid cube are all examples of continua. To obtain a satisfactory definition of a plane curve, all we need require is that a plane continuum contain no square. Therefore a plane continuum is a curve if every square (no matter how small) contains points that are not on the curve, or in modern language, a curve is a continuum that is *nowhere dense* in the plane. All of the usual curves such as circles, ellipses, cycloids, etc., satisfy this definition. Finite sections of hyperbolas or parabolas are also included, but this definition excludes entire parabolas or hyperbolas since they are not bounded. However there are more unusual examples. The *sinusoid* is defined by the set of points $(x, \sin(1/x))$, for $0 < x \leq 1$ together with the portion of the y -axis with $0 \leq y \leq 1$ shown in Figure (1). It is not intuitively clear that this set is connected; this requires use of the technical definition. Another example of a Cantorian Line is the Cantor Brush defined by connecting the point $(1/2, 1/2)$ with straight lines to the points in the Cantor set S lying in the interval in $[0, 1]$ on the x -axis shown in Figure (3).

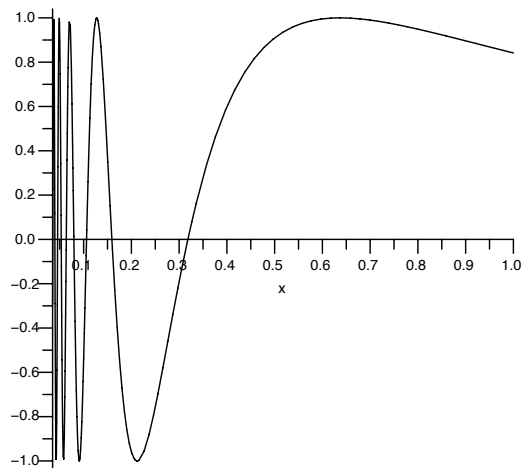


Figure 1: The Sinusoid

Let us now look at the Sierpinski Carpet. This is defined in stages, similar to the definition of the Cantor set. Start with the closed unit square. Divide it into 9 equal subsquares and remove the open middle square. Divide each of the remaining 8 squares into 9 equal subsquares and remove the open middle square in each one. Continue this process ad infinitum. The set of points that remain is called the Sierpinski Carpet (See Figure (4)).

The Sierpinski Carpet is a curve in the Cantor sense. It is defined by a decreasing

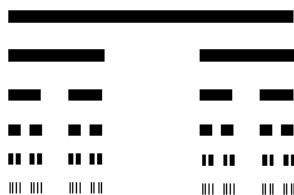


Figure 2: Cantor Set



Figure 3: Cantor Brush

sequence of closed bounded connected sets. The intersection of such is always closed, bounded and connected; namely, it is a continuum. It is also clear that every square will contain points not in the the carpet, so the Sierpinski Carpet is a curve! Furthermore, it can be shown that in a certain topological sense, the Sierpinski Carpet is universal. This means that every curve corresponds in a one-one and continuous way to a subset of the Sierpinski Carpet.

One of the difficulties with the definition of curve given here is that it cannot be extended to curves in three dimensions. The notion of continuum causes no difficulty, however the notion of being nowhere dense does not distinguish between a curve and a surface in three dimension. For instance a straight line segment and the surface of a sphere are both nowhere dense; every cube contain points that are not on the line or the surface. The proper notion that is needed is that of being ‘one-dimensional’; this will be discussed later.

3 Can A Plane Cantor Curve Have a Non-Zero Area?

What is the area of a straight line segment? Surely, since it has zero ‘width’, it’s area must be zero. More precisely the area of a line segment is zero since we may cover it with a number of small squares, the total area of which can be made as small as desired. similarly, the area of any of the usual curves, such as a circle or lemniscate, is also zero. Note that we are not talking about the area enclosed by the curve but the area *of* the curve itself. Is the area of any plane Cantor curve necessarily equal to zero? It would appear that since a

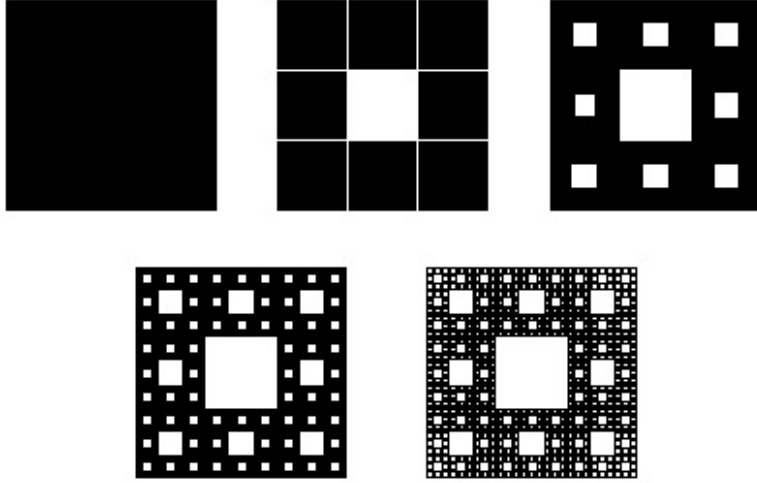


Figure 4: Sierpinski Carpet

Cantor curve cannot contain a square, it must necessarily have zero ‘width’ and therefore zero area. It is rather surprising that this is not the case.

We can obtain a curve in the Cantor sense with a non-zero area by modifying the construction of the Sierpinski Carpet. The Sierpinski Carpet was constructed by starting with a unit square and, at the first stage, removing a square of area $1/9$, at the second stage removing 8 squares, each of area $1/9^2$, at the n^{th} stage removing 8^{n-1} squares each of area $1/9^n$. The total area removed is obtained by summing infinite geometric series:

$$\frac{1}{9} + \frac{1}{9} \cdot \left(\frac{8}{9}\right) + \dots + \frac{1}{9} \cdot \left(\frac{8}{9}\right)^{n-1} + \dots = \frac{1}{9} \frac{1}{1 - \frac{8}{9}} = 1$$

Since the total area removed is 1, the area that remains, namely the area of the Sierpinski Carpet, must be zero. We now modify the construction of the Sierpinski Carpet to obtain a curve with a non-zero area which we shall call the Fat Sierpinski Carpet. Start again with the unit square. At the first stage we remove a cross-like figure of area $9/25$, as shown in Figure (5). However in order to keep the remaining figure connected, we retain the lines AB , CD , EF . At the second stage we remove a cross like figure from each of the remaining 4 squares except for certain lines to keep the figure connected. We can certainly arrange it so that the total area removed at this stage is $(9/25)^2$. We continue this process indefinitely. The points that are never removed form the Fat Sierpinski Carpet. This is a curve in the Cantor sense since it is the intersection of a decreasing sequence of continua, which is a continuum; furthermore the Fat Carpet contains no square. Now the total area

removed is given by the infinite geometric series

$$\frac{9}{25} + \left(\frac{9}{25}\right)^2 + \cdots + \left(\frac{9}{25}\right)^n + \cdots = \frac{9}{25} \cdot \frac{1}{1 - \frac{4}{25}} = \frac{9}{16}.$$

Thus the area of the Fat Sierpinski Carpet must be $7/16$.

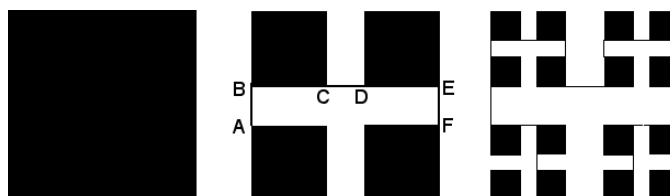


Figure 5: Fat Sierpinski Carpet

Nothing is special about the factor $9/25$ in the above construction. We can just as easily remove areas equal to $a, a^2, a^3 \dots$ at each state where $0 \leq a \leq 1/2$. In this manner we can construct a Fat Sierpinski Carpet of area as close to 1 as we please.

The Fat Sierpinski Carpet can also be defined as the limit of a sequence of simple curves as shown in Figure (6). It can be shown that the limiting curve is the same as the Fat Sierpinski Carpet described in the previous paragraph. Furthermore, the limiting curve is continuous and non-intersecting. We can construct a simple Jordan Curve by connecting the points P and Q in Figure (6) by say a semi-circle. What is the area of the interior region of the Jordan Curve? Notice that the Fat Sierpinski Carpet which is part of the boundary of the Jordan Curve and it has an area of $12/21$. Should this be included in the enclosed area or not? Something is peculiar! One cannot settle this question without giving a precise definition of area. It turns out, according to the accepted definition of area, that the region we have just described *does not possess an area*. It is a mathematical fact of life that we cannot associate an area with every set of points in the plane.

4 Jordan Curves

In the 1880's Camille Jordan came up with a different approach to the definition of a curve. Jordan thought of a curve as the set of points generated by a moving particle. Thus, he defined a curve in space, as the set of points obtained from the parametric equations $x = f(t), y = g(t), z = h(t)$, where f, g , and h are continuous functions on, say, $[0, 1]$ (for a plane curve, the z -equation is omitted). Since only one real parameter is needed to define a curve, it seems reasonable that this should force the curve to be 'one-dimensional'. The properties of continuous functions force a Jordan Curve to be a continuum (if the curve contains at least two points).

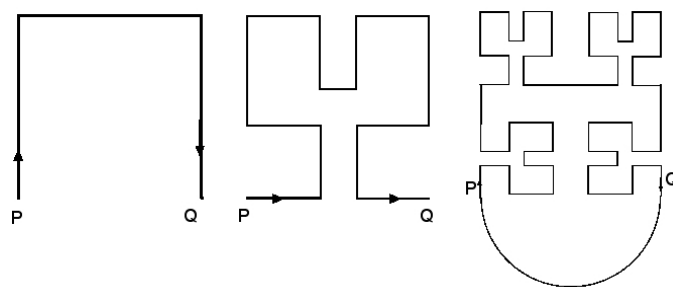


Figure 6: Fat Sierpinski Carpet 2

In 1890, Peano startled the mathematical world by discovering a Jordan curve that passes through each point of the square at least once. The illustrations in Figure (7) show some of the approximations to such a ‘space filling curve’. In the first approximation the point starts out at the lower left hand corner of the square at $t = 0$ and moves along the diagonal, ending up at the right hand corner at $t = 1$.

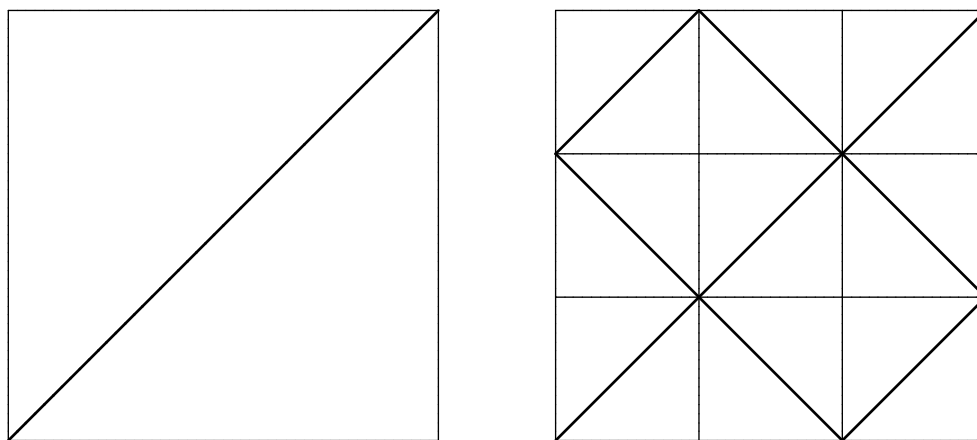


Figure 7: Peano Space-Filling Curve

In the second approximation the square is divided into 9 subsquares and the point moves as shown in along the diagonals of the subsquares tracing out a *bow-tie-like shape* as shown in the second graphic of Figure (7). Each of the 9 diagonals is traversed in $1/9$ of a second. This procedure is repeated indefinitely. The resulting curve is continuous

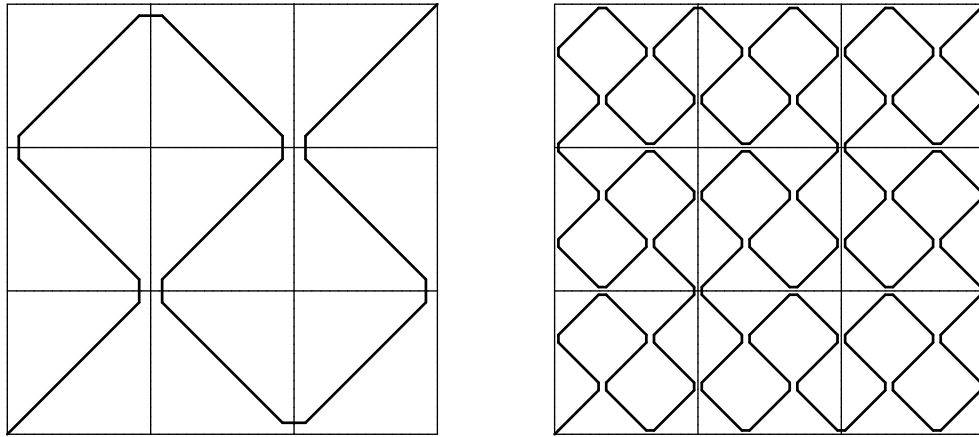


Figure 8: Peano Space-Filling Curve With Cut Corners

and passes through each point of the square at least once; some points are passed through many times. Thus the correspondence between the interval and the square is continuous but not one-to-one. The correspondence that Cantor had obtained between the interval and the square was one-to-one but not continuous. It was later proved by Netto that it is impossible to have a correspondence between the interval and the square that is both one-to-one and continuous. One additional feature of the Peano curve deserves mention: at no point does it possess a tangent line; it is ‘infinitely crinkly’. In terms of motion we have the seemingly paradoxical situation of a particle moving continuously from one corner of the square to another, yet at no point does the particle have a well defined velocity! The x and y components of the Peano-curve are continuous functions which do not possess a derivative at any point. Figures (9) and (10) show the first two approximations to these curves.

In 1891 Hilbert produced another example of a space-filling curve (see Figure(11)). Many other examples of space filling curves have been discovered. See the book by Sagan for a thorough treatment of this topic.

Peano’s example shows that Jordan’s definition is too broad; it is also too narrow in that the sinusoid discussed in Section (2) is not a Jordan curve. Nevertheless, Jordan’s definition did lead to one notable achievement—the famous Jordan Curve Theorem. This theorem states that a simple, closed continuous curve divides the plane into two open connected regions, an interior and an exterior, with the curve being the boundary of both. Here by simple curve, we mean one that does not intersect itself. Although this theorem

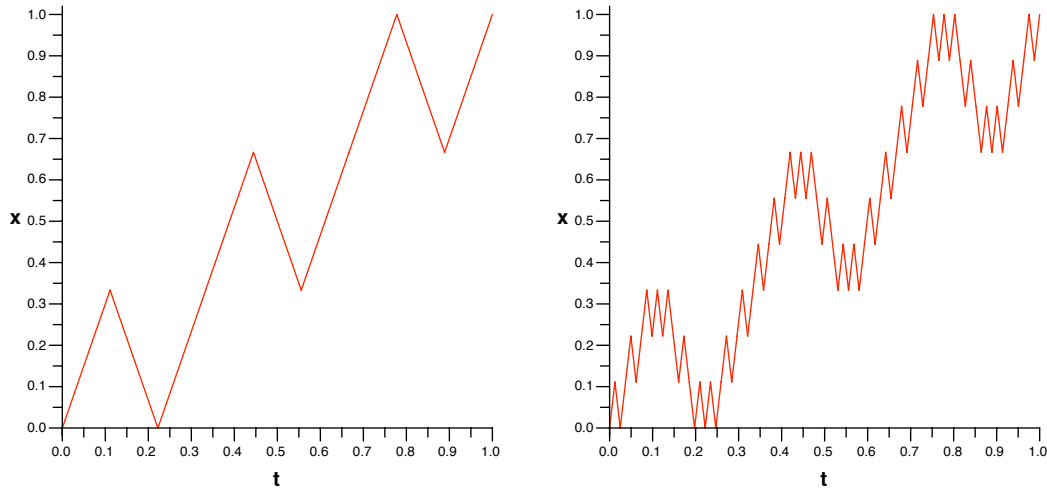


Figure 9: x-Coordinate of 2nd and 3rd Approximations of the Peano Curve

may seem obvious, it deals with the whole class of continuous curves which is exceedingly complex. The theorem is really quite difficult to prove. Even Jordan’s proof had certain flaws in it. The first rigorous proof is due to Veblen, in 1903.

The failure of Jordan’s definition to provide a satisfactory definition of a curve did not end the interest in Jordan curves. Instead interest was stimulated in the question “what kind of objects are the continuous images of a segment?”. In 1908, Schoenflies discovered a geometric characterization of plane Jordan curves. In 1924, working independently, Hahn and Mazurkiewicz characterized Jordan curves in three dimensions (in fact, in any topological space). They showed that a Jordan curve was a continuum with the additional property of being locally connected. For example, the sinusoid, discussed earlier, is a continuum that is not locally connected—a point of the sinusoid on the y -axis has a small neighborhood that does not contain a connected part of the sinusoid. Hence, the sinusoid is not a Jordan curve.

5 Dimension—Pincers, Scissors and Saws

We have seen that the definition of curve due to Cantor is suitable for plane curves but not for space curves. What then is a curve in space? To answer this question we need the concept of *dimension*.

The basic idea is due to Poincare (1912). Poincare observed that to separate a piece of curve from the remainder of the curve, it is necessary to remove only one point. To separate

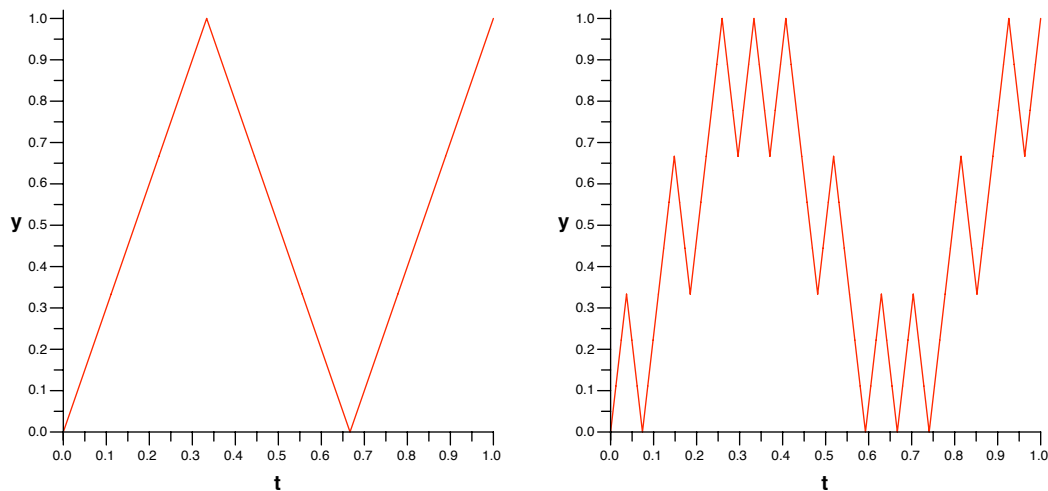


Figure 10: y -Coordinate of 2nd and 3rd Approximations of the Peano Curve

piece of a surface from the remainder of the surface it is necessary to remove a curve. To separate a solid it is necessary to remove a surface. These three different situations are reflected practically in the types of tools one needs to cut wires (curves), cloth (surfaces) and blocks of wood (solids). To cut a wire one needs a *pincer* to cut it at points; to cut cloth one needs *scissors* to cut along curves and to cut a block of wood one needs a *saw* to cut it along surfaces.

Poincaré's observations suggest an inductive definition of the dimension of a point set. A finite number of points is defined to have dimension zero; if a point set can be separated by a zero dimensional set, it is called one-dimensional; if a one-dimensional set is required to separate it, it is called two-dimensional; if a two dimensional set is required to separate it, it is called three-dimensional. Using these ideas, Brouwer came up with a satisfactory definition of dimension. Later, in the 1920's, Menger, and working independently, Urysohn, constructed an improved version which is now commonly accepted as the definition of dimension.

We present a brief outline of the Menger-Urysohn approach. A point set in space is called 0-dimensional if it is non-empty, and if each point has an arbitrarily small neighborhood whose boundary has no points in common with the set (see Figure (12)). A finite set of points has dimension zero. The set of rational points in $[0, 1]$ and has dimension 0 since each rational point can be made the center of of an arbitrarily small interval with irrational end points. The Cantor set also has dimension 0. Sets of dimension zero are also called *dispersed sets*.

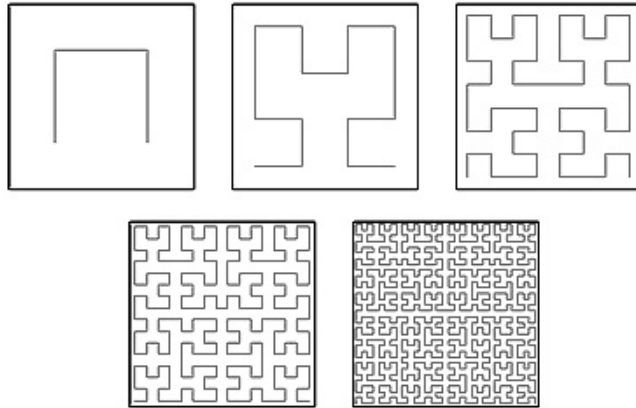


Figure 11: Hilbert Space-Filling Curve

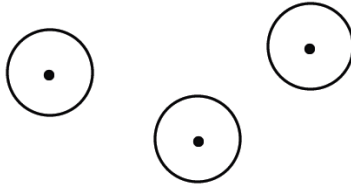


Figure 12: 0-Dimensional Set

A point set S is called 1-dimensional if it is not 0-dimensional, and if each point of S has an arbitrarily small neighborhood which intersects S in a 0-dimensional set (see Figure (13)). The usual curves—circles, lemniscates and the like—have dimension 1. The Sierpinski Carpet, the sinusoid, and the Cantor Brush also have dimension 1. With this definition, we can finally give the modern definition of a curve: *a curve is a one-dimensional continuum*. Point sets of dimension 2 or 3 may be defined in a similar manner. A *surface* is then defined as a continuum of dimension 2, and a *solid* as a continuum of dimension 3.

The striking illustration shown in Figure (14) is called the Sierpinski Sponge. The Sierpinski Sponge is the three dimensional version of the Sierpinski Carpet. The Sponge is constructed in Stages starting with a solid unit cube. At the first stage the cube is divided into 27 equal subcubes, each of side $1/3$. The innermost cube and the 6 cubes with which it has a face in common form a three dimensional cross. This cross is removed leaving a figure composed of 20 closed cubes. At the second stage, a three dimensional cross is deleted from each of these 20 cubes, leaving 20^2 cubes, each of side $1/3^2$. This process is continued indefinitely; the points that are never removed form the Sierpinski Sponge. It is

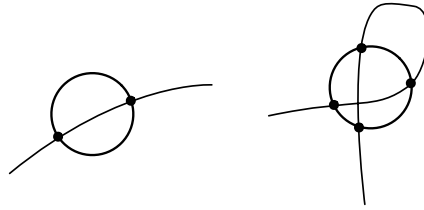


Figure 13: 1-Dimensional Sets

a one-dimensional continuum, and therefore, a curve.

The definition of dimension can be extended to point sets in Euclidean n -space, and even to more general topological spaces. The dimension of a point set is a *topological invariant*, that is, the dimension remains the same when the point set is transformed by a transformation that is one-to-one and continuous in both directions. Therefore, the image of a curve under such a transformation remains a curve. The Sierpinski Sponge is sometimes called the Universal Curve, since it can be shown that any curve in space corresponds in a one-to-one and continuous manner to a subset of the Sierpinski Sponge.

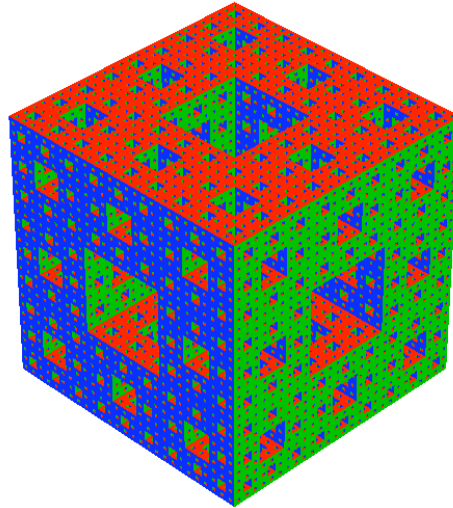


Figure 14: Sierpinski Sponge

6 The Koch Snowflake Curve

The Koch Snowflake Curve is defined as a limit of a sequence, K_n , of curves. K_0 is an equilateral triangle having 3 sides of length 1. K_1 is the curve obtained by replacing the middle third of each side with an equilateral triangle with sides of length $1/3$ (not including the base) as shown in Figure (15). The curve K_1 has $3 \cdot 4$ straight line segments, each of length $1/3$. The curve K_2 is formed by replacing the middle of each of the 12 sides of K_1 with equilateral triangles of length $1/3^2$. Thus K_2 is a curve with in $3 \cdot 4^2$ straight line segments each of length $1/3^2$. Continuing this process, we have, at the n -th stage a curve consisting of $3 \cdot 4^n$ straight line segments, each of length $1/3^n$. The limiting curve as n approaches ∞ is the Koch Snowflake Curve. It can be shown that the limit is a continuum with topological dimension 1 and therefore a curve as we have defined it.

The Koch Curve has many interesting properties. It is a continuous curve, yet at no point does it possess a derivative; it is “infinitely crinkly”. What is the length of the Koch curve? The length of the approximating curve K_n is $3 \cdot 4^n / 3^n$. This approaches infinity as with n , so that the Koch curve has infinite length. Let us now look at the enclosed area. Let A_n be the area enclosed by K_n . The area A_0 is the area of an equilateral triangle with side 1, namely $\sqrt{3}/4$. To obtain A_1 , 3 triangles of length $1/3$ are adjoined to K_0 , thus

$$A_1 = A_0 + 3 \cdot \frac{1}{9} \cdot A_0.$$

Similarly we find that

$$A_n = A_0 \left(1 + 3 \cdot \frac{1}{9} + 3 \cdot 4 \cdot \frac{1}{9^2} + \cdots + 3 \cdot 4^{n-1} \cdot \frac{1}{9^n} \right)$$

$$A = \lim_{n \rightarrow \infty} A_n \tag{1}$$

$$= A_0 \left(1 + 3 \cdot \frac{1}{9} + 3 \cdot 4 \cdot \frac{1}{9^2} + \cdots + 3 \cdot 4^{n-1} \cdot \frac{1}{9^n} + \cdots \right) \tag{2}$$

$$= A_0 \left(1 + \frac{1}{3} \sum_0^{\infty} \left(\frac{4}{9} \right)^n \right) \tag{3}$$

$$= \frac{8}{5} A_0 \tag{4}$$

Appendix — The Cantor Set

The Cantor set is constructed in stages. Start with a closed unit interval, $T_0 = [0, 1]$. T_1 is defined to be the set of points remaining after removing the open set $(1/3, 2/3)$ (the middle third) of T_0 . Thus T_1 is the union of the two closed intervals $[0, 1/3]$ and $[2/3, 1]$. T_2 is the

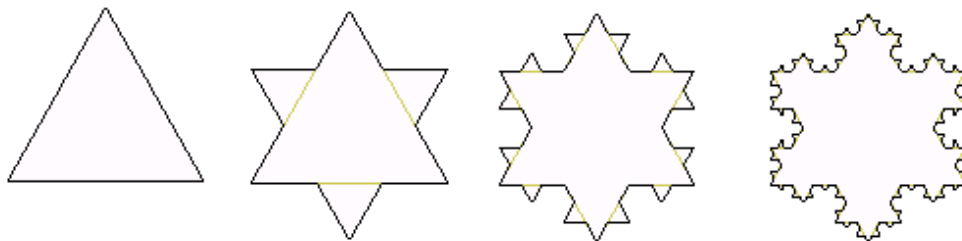


Figure 15: Koch Snowflake Curve

set remaining after removing the middle thirds of two closed intervals in T_1 . Continue in this manner; T_n is obtained by removing the open middle thirds of the 2^{n-1} sets in T_{n-1} . The Cantor set is defined to be the set of points that are never removed at any stage. T may also be defined as the intersection of the decreasing sequence of closed sets T_n .

What points are contained in the Cantor set? Clearly the end points of the the closed intervals that remained at each stage are in T ; these are the points $0, 1/3, 2/3, 1, 1/9, 2/9, 7/9, 8/9, \dots$. However there are other points in T . To describe these we use the ternary representation of number x in $[0, 1]$ namely

$$x = (.t_1t_2t_3 \dots t_j \dots)_3 = \sum_{j=1}^{\infty} t_j/3^j$$

Since $1/3 = (.1)_3$, $2/3 = (.2)_3$, any point removed in the first stage must have $t_1 = 1$ in its ternary representation. Likewise one can show that if a point is removed at the n -th stage, then $t_n=1$ in its ternary representation. Thus the Cantor set can be described as all numbers in $[0, 1]$ having a ternary representation that can be written without any 1's. For instance $1/4 = (.020202 \dots)_3$ and $3/4 = (.202020 \dots)_3$ are in T and these points are not the end points of any of the closed intervals that remain at any stage.

How 'big' is the set T ? From one point of view it is not very big since so many points are removed to obtain T . In fact the total length of all the middle thirds removed at any stage is given by the infinite geometric series $1/3 + 2/9 + 4/27 + \dots$ whose sum is 1; thus the set T has measure zero. From another point of view T has 'as many' points in it as $[0, 1]$ since the points of T can be put in one-to-one correspondence with the points in $[0, 1]$. The correspondence is very simple, namely

$$(.t_1t_2 \dots t_n \dots)_3 \longleftrightarrow (.b_1b_2 \dots b_n \dots)_2$$

where the right hand side is the binary representation and $2t_n = b_n$. Since $t_n = 0, 2$, we must have $b_n = 0, 1$. Thus every number in T corresponds to a unique number in $[0, 1]$, and every number in $[0, 1]$ corresponds to a unique number in T .

References

- Blumenthal, L. and Menger, K. *Studies in Geometry*, W. H. Freeman and Company, San Francisco (1970).
- Moore, E. H., On Certain Crinkly Curves, *Trans. Amer. Math Soc.*, 1, 72-90 (1900).
- Sagan, H. *Space Filling Curves*, Springer-Verlag, New York (1994).