A Note on the History of the Cantor Set and Cantor Function

JULIAN F. FLERON
SUNY at Albany
Albany, New York 12222

A search through the primary and secondary literature on Cantor yields little about the history of the Cantor set and Cantor function. In this note, we would like to give some of that history, a sketch of the ideas under consideration at the time of their discovery, and a hypothesis regarding how Cantor came upon them. In particular, Cantor was not the first to discover "Cantor sets." Moreover, although the original discovery of Cantor sets had a decidedly geometric flavor, Cantor's discovery of the Cantor set and Cantor function was neither motivated by geometry nor did it involve geometry, even though this is how these objects are often introduced (see e.g. [1]). In fact, Cantor may have come upon them through a purely arithmetic program.

The systematic study of point set topology on the real line arose during the period 1870–1885 as mathematicians investigated two problems:

1) conditions under which a function could be integrated, and
2) uniqueness of trigonometric series.

It was within the framework of these investigations that the two apparently independent discoveries of the Cantor set were made; each discovery linked to one of these problems.

Bernhard Riemann (1826–1866) spent considerable time on the first question, and suggested conditions he thought might provide an answer. Although we will not discuss the two forms his conditions took (see [2, pp. 17–18]), we note that one of these conditions is important as it eventually led to the development of measure theoretic integration [2, p. 28]. An important step in this direction was the work of Hermann Hankel (1839–1873) during the early 1870s. Hankel showed, within the framework of Riemann, that the integrability of a function depends on the nature of certain sets of points related to the function. In particular, "a function, he [Hankel] thought, would be Riemann-integrable if, and only if, it were *pointwise discontinuous* [2, p. 30]," meaning, in modern terminology, that for every $\sigma > 0$ the set of points x at which the function oscillated by more than σ in every neighborhood of x was nowhere dense. Basic to Hankel's reasoning was his belief that sets of the form $\{1/2^n\}$ were prototypes for all nowhere dense subsets of the real line. Working under this assumption Hankel claimed that all nowhere dense subsets of the real line could be enclosed in intervals of arbitrarily small total length (i.e. had zero outer content) [2, p. 30]. As we shall see, this is not the case. (See also [3].)

Although Hankel's investigation into the nature of certain point sets would become extremely important, "as was the case with Dirichlet and Lipschitz, it was the inadequacy of his understanding of the possibilities of infinite sets—in particular, nowhere dense sets—that led him astray. It was not until it was discovered that nowhere dense sets can have positive outer content that the importance of negligible sets in the measure-theoretic sense was recognized [2, p. 32]." The discovery of such sets, nowhere dense sets with positive outer content, was made by H. J. S. Smith (1826–1883), Savilian Professor of Geometry at Oxford, in a paper [4] of 1875. After an exposition of the integration of discontinuous functions, Smith presented a method for constructing nowhere dense sets that were much more "substantial" than the set

$\{1/2^n\}$. Specifically, he observed the following:

> Let m be any given integral number greater than 2. Divide the interval from 0 to 1 into m equal parts; and exempt the last segment from any subsequent division. Divide each of the remaining $m-1$ segments into m equal parts; and exempt the last segments from any subsequent subdivision. If this operation be continued *ad infinitum*, we shall obtain an infinite number of points of division P upon the line from 0 to 1. These points lie in loose order... [4, p. 147].

In modern terminology Smith's 'loose order' is what we refer to as nowhere dense. Implicit in Smith's further discussion is the assumption that the exempted intervals are open, so the resulting set is closed. Today this set would be known as a general Cantor set, and this seems to be the first published record of such a set.

Later in the same paper, Smith shows that by dividing the intervals remaining before the nth step into m^n equal parts and exempting the last segment from each subdivision we obtain a nowhere dense set of positive outer content. Smith was well aware of the importance of this discovery, as he states, "the result obtained in the last example deserves attention, because it is opposed to a theory of discontinuous functions, which has received the sanction of an eminent geometer, Dr. Hermann Hankel [4, p. 149]." He continues by explaining the difficulties in the contemporary theories of integration that his examples illuminate.

It is interesting to note that an editor's remark at the conclusion of Smith's paper states "this paper, *though it was not read*, was offered to the society and accepted in the usual manner." (Emphasis added.)[1] In fact, this paper went largely unnoticed among mathematicians on the European continent and unfortunately Smith's crucial discoveries lay unknown. It took the rediscovery, almost a decade later, of similar ideas by Cantor to illuminate the difficulties of contemporary theories of integration and to begin the evolution of measure-theoretic integration.

Georg Cantor (1845–1918) came to the study of point set topology after completing a thesis on number theory in Berlin in 1867. He began working with Eduard Heine (1821–1881) at the University of Halle on the question of the uniqueness of trigonometric series. This question can be posed as follows:

If for all x except those in some set P we have

$$\frac{1}{2}a_0 + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx)) = 0$$

must all the coefficients a_n and b_n be zero?

Heine answered the question in the affirmative "when the convergence was *uniform in general* with respect to the set P, which is thus taken to be finite [2, p. 23]," meaning, by definition, that the convergence was uniform on any subinterval that did not contain any points of the finite set P.

Cantor proceeded much further with this problem. In papers [5, 6] of 1870 and 1871, he removed the assumption that the convergence was "uniform in general" and began to consider the case when P was an infinite set. In doing so he began to look at what we now consider the fundamental point set topology of the real line. In a paper [8] of 1872, Cantor introduced the notion of a *limit point* of a set that he defined as we do today, calling the limit points of a set P the *derived set*, which he denoted by P'. Then P'' was the derived set of P', and so on. Cantor showed that if the set P was

[1] It is possible that "not read" simply meant that the paper was not presented at a meeting of the London Mathematical Society. However, in weighing the significance of this note, one must consider that in vols. 3–10 of the *Proceedings of the London Mathematical Society* (1871–1879), and perhaps even further, no other paper was similarly noted.

such that $P^{(n)} = \varnothing$ for some integer n and the trigonometric series $\frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n \cos(nx) + b_n \sin(nx)) = 0$, except possibly on P, then all of the coefficients had to be zero. Cantor's work on this problem was "decisive" [9, p. 49], and doubly important as his derived sets would play an important role in much of his upcoming work.

In the years 1879–1884 Cantor wrote a series of papers entitled "Über unendliche, lineare Punktmannichfaltigkeiten [10–15]," that contained the first systematic treatment of the point set topology of the real line.[2] It is the introduction of three terms in this series that concerns us most here. In the first installment of this series Cantor defines what it meant for a set to be *everywhere dense* (literally "überall dicht"), a term whose usage is still current. He gives a few examples, including the set of numbers of the form $2^{2n+1}/2^m$ where n and m are integers, and continues by noting the relationship between everywhere dense sets and their derived sets. Namely, $P \subseteq (\alpha, \beta)$ is everywhere dense in (α, β) if [and only if] $P' = (\alpha, \beta)$ [10, pp. 2–3]. In the fifth installment of this series Cantor discusses the partition of a set into two components that he terms *reducible* and *perfect* [14, p. 575]. His definition of a perfect set is also still current: A set P is perfect provided that $P = P'$.

After introducing the term *perfect* in the fifth installment, Cantor states that perfect sets need not be everywhere dense [14, p. 575]. In the footnote to this statement Cantor introduces the set that has become known as the *Cantor (ternary) set*: The set of real numbers of the form

$$x = \frac{c_1}{3} + \cdots + \frac{c_\nu}{3^\nu} + \cdots$$

where c_ν is 0 or 2 for each integer ν. Cantor notes that this is an infinite, perfect set with the property that it is not everywhere dense in any interval, regardless of how small the interval is taken to be. We are given no indication of how Cantor came upon this set.

During the time Cantor was working on the 'Punktmannichfaltigkeiten' papers, others were working on extensions of the Fundamental Theorem of Calculus to discontinuous functions. Cantor addressed this issue in a letter [18] dated November 1883, in which he defines the Cantor set, just as it was defined in the paper [14] of 1883 (which had actually been written in October of 1882). However, in the letter he goes on to define the Cantor function, the first known appearance of this function. It is first defined on the complement of the Cantor set to be the function whose values are

$$\frac{1}{2}\left(\frac{c_1}{2} + \cdots + \frac{c_{\mu-1}}{2^{\mu-1}} + \frac{2}{2^\mu}\right)$$

for any number between

$$a = \frac{c_1}{3} + \cdots + \frac{c_{\mu-1}}{3^{\mu-1}} + \frac{1}{3^\mu} \quad \text{and} \quad b = \frac{c_1}{3} + \cdots + \frac{c_{\mu-1}}{3^{\mu-1}} + \frac{2}{3^\mu},$$

where each c_ν is 0 or 2. Cantor then concludes this section of the letter by noting that this function can be extended naturally to a continuous increasing function on $[0, 1]$. That serves as a counterexample to Harnack's extension of the Fundamental Theorem of Calculus to discontinuous functions, which was in vogue at the time (see e.g. [2, p. 60]). We are given no indication of how Cantor came upon this function.

There are two other topics that interested Cantor that we would like to mention because they are indicative of Cantor's facility with arithmetic constructions and it is

[2] In addition, these papers contained many other topics that had far reaching implications (see [16, 17]), including Cantor's investigation of higher order derived sets that marked the "beginnings of Cantor's theory of transfinite numbers [2, p. 72]."

possibly within this setting that Cantor came upon the Cantor set and Cantor function. First, Cantor spent some time in the mid 1870s considering the possible existence of a bijective correspondence between a line and a plane, a question most of his contemporaries had dismissed as absurd. In 1877, in a letter to Richard Dedekind (1831–1916), Cantor explained that he had found such a correspondence. This "correspondence" can be expressed as follows:

> Let (x_1, x_2) be a point in the unit square, and let $0.x_{1,1}x_{1,2}x_{1,3}\cdots$ and $0.x_{2,1}x_{2,2}x_{2,3}\cdots$ be decimal expansions of x_1 and x_2 respectively. Map the point (x_1, x_2) to the point on the real line whose decimal expansion is $0.x_{1,1}x_{2,1}x_{1,2}x_{2,2}\cdots$ (See e.g. [19, p. 187].)

Dedekind pointed out that there was a problem with this approach. The decimal expansions of rationals are not unique, so to avoid duplication we must not allow expansions of some type, say expansions that contain infinite strings of zeros. However, by disallowing expansions with infinite strings of zeros, the irrational number 0.11010201010201010102... could never be obtained under Cantor's correspondence.

This reasoning does however give us an injection of $[0,1]\times[0,1]$ into $[0,1]$. It is trivial to find an injection of $[0,1]$ into $[0,1]\times[0,1]$. These two facts, together with the Schroeder-Bernstein Theorem (if there are injections of the set A into the set B and B into A respectively, then there is a bijective correspondence between A and B; see e.g. [20]), allow us to conclude that there is a bijective correspondence between $[0,1]$ and $[0,1]\times[0,1]$. However, set theory was in its infancy during the period in question and it would be 20 years before E. Schroeder and Felix Bernstein independently proved the theorem that bears their names [16, p. 172–173] and occasionally Cantor's name as well (e.g. [21, 22]). So this was not an option for Cantor.

Instead, Cantor needed to explicitly exhibit a bijection. To do this he modified his previous approach to use continued fractions [23]. Denote the continued fraction

$$\cfrac{1}{a_1+\cfrac{1}{a_2+\cfrac{1}{a_3+\cdots}}} \quad \text{by } [a_1, a_2, a_3, \ldots] \quad \text{where} \quad a_1, a_2, a_3, \ldots > 0 \text{ are integers.}$$

Since a continued fraction is infinite if, and only if, it represents an irrational number, in which case the representation is unique [see e.g. 24], Cantor could set up the correspondence

$$([a_{1,1}, a_{1,2}, \ldots], [a_{2,1}, a_{2,2}, \ldots], \ldots, [a_{n,1}, a_{n,2}, \ldots]) \leftrightarrow [a_{1,1}, a_{2,1}, \ldots a_{n,1}, a_{1,2}, a_{2,2}, \ldots, a_{n,2}, \ldots]$$

between n-tuples of irrationals in $(0,1)^n = (0,1)\times(0,1)\times\cdots\times(0,1)$ and irrationals in $(0,1)$. This avoids the difficulties of the previous approach and gives a bijective correspondence between $([0,1]-\mathbf{Q})^n$ and $[0,1]-\mathbf{Q}$. Cantor then took great lengths to prove there was a bijective correspondence between $[0,1]$ and $[0,1]-\mathbf{Q}$. Repeated application of this fact combined with the previous correspondence gives a bijective correspondence between $[0,1]^n$ and $[0,1]$.

Secondly, it is known that Cantor studied binary expansions. In fact:

> Cantor recognised that the power of the linear continuum, denoted by **o**, could be represented as well by [the power of] the set of all representations:
>
> $$x = \frac{f(1)}{2} + \cdots + \frac{f(\nu)}{2^\nu} + \cdots,$$
>
> where $f(\nu) = 0$ or 1 [for each integer ν] [19, p. 209].

There is, so it seems, no substantive evidence about how Cantor came upon the Cantor set and Cantor function. However, given Cantor's route into point set topology, his arithmetic introduction of the Cantor set and Cantor function, and his facility with arithmetic methods, as we have just illustrated, it is feasible that it is within the arithmetic framework of binary and ternary expansions that Cantor came upon the Cantor set and Cantor function.

REFERENCES

1. R. L. Wheeden and A. Zygmund, *Measure and Integral: An Introduction to Real Analysis*, Marcel Decker, Inc., New York, 1977, p. 35.
2. T. Hawkins, *Lebesgue's Theory of Integration: Its Origins and Development*, Chelsea Publishing Co., Madison, WI, 1975.
3. H. L. Royden, *Real Analysis*, Macmillan Publishing Co., New York, 1988, p. 64.
4. H. J. S. Smith, On the integration of discontinuous functions, *Proc. London Math. Soc.* (*1*) 6 (1875), 140–153.
5. G. Cantor, Beweis, daß eine für jeden reellen Werth von x durch eine trigonometrische Reihe gegebene Function $f(x)$ sich nur auf eine einzige Weise in dieser Form darstellen läßt, Part 1, *Crelle Jl. Math.* 72 (1870), 139–142. Reprinted [7, pp. 80–83].
6. G. Cantor, Beweis, daß eine für jeden reellen Werth von x durch eine trigonometrische Reihe gegebene Function $f(x)$ sich nur auf eine einzige Weise in dieser Form darstellen läßt, Part 2, *Crelle Jl. Math.* 73 (1871), 294–296. Reprinted [7, pp. 84–86].
7. G. Cantor, *Gesammelte Abhandlungen mathematischen und philosophischen Inhalts*, E. Zermelo (ed.), Springer-Verlag, New York, 1980.
8. G. Cantor, Über die Ausdehnung eines Satzes aus der Theorie der trigonometrischen Reihen, *Math. Ann.* 5 (1872), 123–132. Reprinted [7, pp. 92–102].
9. W. Purkert, Cantor's Philosophical Views, in *The History of Modern Mathematics*, Vol. 1: *Ideas and Their Reception*, D. E. Rowe and J. McCleary (eds.), Academic Press, Boston, 1989.
10. G. Cantor, Über unendliche, lineare Punktmannichfaltigkeiten, Part 1, *Math. Ann.* 15 (1879), 1–7. Reprinted [7, pp. 139–145].
11. G. Cantor, Über unendliche, lineare Punktmannichfaltigkeiten, Part 2, *Math. Ann.* 17 (1880), 355–358. Reprinted [7, pp. 145–148].
12. G. Cantor, Über unendliche, lineare Punktmannichfaltigkeiten, Part 3, *Math. Ann.* 20 (1882), 113–121. Reprinted [7, pp. 149–157].
13. G. Cantor, Über unendliche, lineare Punktmannichfaltigkeiten, Part 4, *Math. Ann.* 21 (1883), 51–58. Reprinted [7, pp. 157–164].
14. G. Cantor, Über unendliche, lineare Punktmannichfaltigkeiten, Part 5, *Math. Ann.* 21 (1883), 545–591. Reprinted [7, pp. 165–209].
15. G. Cantor, Über unendliche, lineare Punktmannichfaltigkeiten, Part 6, *Math. Ann.* 23 (1884), 453–488. Reprinted [7, pp. 210–246].
16. J. W. Dauben, *Georg Cantor: His Mathematics and Philosophy of the Infinite*, Princeton University Press, Princeton, NJ, 1990.
17. W. Purkert and H. J. Ilgauds, *Georg Cantor: 1845–1918*, Birkhäuser Verlag, Basel, 1987.
18. G. Cantor, De la puissance des ensembles parfaits de points, *Acta Math.* 4 (1884), 381–392. Reprinted [7, pp. 252–260].
19. J. W. Dauben, The development of Cantorian set theory, in *From the Calculus to Set Theory: 1630–1910*, I. Grattan-Guinness (ed.), Gerald Duckworth & Co., London, 1980.
20. G. F. Simmons, *Introduction to Topology and Modern Analysis*, R. E. Krieger Publishing Co., Malabar, FL, 1983, pp. 28–30.
21. K. Hrbacek and T. Jech, *Introduction to Set Theory*, Marcel Decker, Inc., New York, 1984, p. 72.
22. G. Takeuti and W. M. Zaring, *Introduction to Axiomatic Set Theory*, Springer-Verlag, New York, 1982, p. 86.
23. G. Cantor, Ein Beitrag zur Mannigfaltigkeitslehre, *Crelle Jl. Math.* 84, (1878), 242–258. Reprinted [7, pp. 119–133].
24. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Clarendon Press, Oxford, 1989, pp. 129–140.