Chapter 2

Computing Connectedness

2.1 Introduction

This chapter examines some elementary concepts from point-set topology — specifically connectedness, total disconnectedness and perfectness. The goal is to obtain computational tools that allow us to determine the number and size of connected components of a data set at any given resolution.

The first part of the chapter reformulates the classical definitions in terms of a resolution parameter, ϵ. We then show that for a compact space, X, the topological properties of connectedness, total disconnectedness, and perfectness can be deduced by examining the limiting behavior of the number, $C(\epsilon)$, and size, $D(\epsilon)$, of ϵ-connected components as $\epsilon \to 0$. We characterize the limiting behavior of these two quantities by a power law, and compute the corresponding disconnectedness and discreteness indices.

In Section 2.3, we give a new algorithm based on the minimal spanning tree (MST) that implements these ideas numerically for arbitrary finite point-set data. We show that the MST is an ideal data structure for representing ϵ-components of a finite point-set. Essentially, this is because the MST always joins two subsets by the smallest possible edge. An important step is to determine a cutoff resolution, ρ, so that the computed results are a good representation of the true space for $\epsilon > \rho$. When the underlying space is perfect, ρ is well approximated by the resolution at which the data first appears to have an isolated point.

Finally, in Section 2.4, we demonstrate the effectiveness of our techniques by applying them to a variety of examples. We present data that exhibit different types of scaling in the number and size of their ϵ-components. We also investigate the dependence of the cutoff resolution ρ on the number of data points and on the uniformity of their distribution over the attractor. The first set of examples are fractals generated by closely related iterated function systems. Each has a distinct topology but all have the same Hausdorff dimension. We show that the disconnectedness and discreteness indices classify the sets according to their topology. We next analyze five Cantor sets to demonstrate different types of scaling that can occur in the functions $C(\epsilon)$ and $D(\epsilon)$ as $\epsilon \to 0$. These examples lead us to conjecture that Cantor sets with zero Lebesgue measure have disconnectedness index equal to their box-counting dimension. Results that prove some special cases of this conjecture are given in Chapter 5.

The material in this Chapter has been published in [71, 72].

2.2 Foundations for computing connectedness

2.2.1 Concepts from point-set topology

Connectedness is a very intuitive concept that captures the notion of continuity of a set of points in a topological way. The mathematical formulation goes back a little over one hundred years and the following definition can be traced to Jordan [42]. A topological space X is *connected* if it cannot be decomposed into the union of two non-empty, disjoint, closed sets. If such a decomposition exists then X is said to be *disconnected* — that is, if there are two closed sets U and V such that $U \cap V = \emptyset$ and $U \cup V = X$.

This definition does not lend itself to computation and in the following section we will give a definition that is equivalent to an early formulation of connectedness in metric spaces due to Cantor [35]. Cantor's definition uses ϵ*-chains*: a finite sequence of points $x_0, \ldots, x_n$ that satisfy $d(x_i, x_{i+1}) < \epsilon$ for $i = 1, \ldots, n$. A set, X, is *Cantor-connected* when every two points in X can be linked by an ϵ-chain for arbitrarily small ϵ. This definition agrees with the one above only in the special case of compact metric spaces. For example, the rational numbers are Cantor-connected but disconnected in the regular sense.

One object we are particularly interested in is the *connected component* of a point. Given $x \in X$, this is the largest connected subset of X containing x. For example, if $X = [0,1] \cup [2,3]$ then the connected component of $\frac{1}{2}$ is $[0,1]$. If the connected component of every point is only the point itself then the set is said to be *totally disconnected.* The rationals are totally disconnected, as is the middle-thirds Cantor set.

Another concept from point-set topology is the property of perfectness. This means every point has arbitrarily small neighborhoods containing infinitely many other points, so that no point is isolated. Formally a set is *perfect* if it is equal to the set of its accumulation points. Notice that this definition implies that only closed sets can be perfect.

Any compact metric space that is totally disconnected and perfect is homeomorphic to the middle-thirds Cantor set (see [35] for a proof). The combination of these properties may seem somewhat paradoxical, since they tell us that each point is isolated in one sense — its connected component is that single point, and yet no point is isolated, since each must be the limit of some sequence of points in the Cantor set.

In the following section we reformulate these definitions in a way that relies on extrapolation, making it possible to implement the ideas numerically. The basic approach is to look at the set with a finite resolution ϵ, see how certain properties change as $\epsilon \to 0$, and infer information about the topology.

2.2.2 ϵ-Resolution definitions

Given a compact subset X of a metric space, we say it is ϵ*-disconnected* if it can be written as the union of two sets that are separated by a distance of at least ϵ — i.e., there are two closed subsets, U and V with $U \cup V = X$ and $d(U,V) \equiv \inf_{x \in U, y \in V} d(x,y) \geq \epsilon$. Otherwise, X is ϵ*-connected.* As with Cantor's definition, a set that is ϵ-connected for any $\epsilon > 0$ is connected if and only if it is compact. This follows from the simple lemma below.

Lemma 2. *If a compact metric space X is disconnected, then it is ϵ-disconnected for some $\epsilon > 0$.*

Proof. Since X is disconnected, there are closed disjoint sets U and V such that $U \cup V = X$. Now suppose $d(U,V) = 0$. This implies that there are sequences $x_n \in U$, $y_n \in V$ with

$d(x_n, y_n) \to 0$. Since U and V are compact, there must be convergent subsequences $x_{n_i} \to x^*$ and $y_{n_j} \to y^*$ with $d(x^*, y^*) = 0$. But this implies $x^* = y^*$, which is impossible since U and V are disjoint, so there must be an $\epsilon > 0$ such that $d(U, V) = \epsilon$. □

Restricting our attention to compact sets is not unreasonable, since we are primarily interested in sets that are well approximated on a computer as finite point-sets. In the context of dynamical systems, we are interested in attractors and other invariant sets. These are closed subsets of metric spaces, and are therefore compact when they are bounded.

We now make an ϵ-resolution definition of connected component. A subset $A \subset X$ is an *ϵ-component* if A is ϵ-connected and $d(A, X \setminus A) \geq \epsilon$. Given a resolution, ϵ, X has a natural decomposition as the disjoint union of its ϵ-components. We can exploit this decomposition to deduce topological properties of the set. For example, if the only ϵ-component is X itself for all ϵ, then we can conclude that X is connected. We also know that at any fixed resolution $\epsilon > 0$, a compact set has a finite number of ϵ-components. To see this, take a covering of X by ϵ-balls. Since X is compact, there is a finite sub-cover. By their definition, every ϵ-component must contain at least one ϵ-ball$\cap X$, so there is a finite number of ϵ-components. This motivates the introduction of a function $C(\epsilon)$ that counts the number of ϵ-components at resolution ϵ. This function is monotonic: if $\epsilon_1 < \epsilon_2$ then $C(\epsilon_1) \geq C(\epsilon_2)$.

By looking at the size of the ϵ-components we can deduce the properties of total disconnectedness and perfectness. There are a number of ways to measure the size of a set; the quantity used depends on the context. We use the diameter, since this is defined in any metric space. Recall that the diameter of a set A is the largest distance between any two points in the set: $\mathrm{diam}(A) = \sup_{x,y \in A} d(x, y)$. At a resolution ϵ, let $\mathcal{D}(\epsilon)$ represent the set of diameter measurements of the ϵ-components. For notational convenience, we write $D(\epsilon) = \max \mathcal{D}(\epsilon)$ for the function that describes how the largest diameter changes with resolution. $D(\epsilon)$ is a monotonic non-increasing non-negative function, so the limit as $\epsilon \to 0$ must always exist.

From the definition of total disconnectedness, we have the following result.

Lemma 3. *A compact set X is totally disconnected if and only if* $\lim_{\epsilon \to 0} D(\epsilon) = 0$.

Proof. By way of obtaining a contradiction to the forward direction, suppose that

$$\lim_{\epsilon \to 0} D(\epsilon) = \delta > 0.$$

Take any sequence, $\epsilon_n \to 0$ and construct a tree as follows. At level n list all the ϵ_n-components with diameter $\geq \delta$. There are a finite number of these. Order the tree by set inclusion, i.e. an edge connects A_{ϵ_j} and $A_{\epsilon_{j+1}}$ if and only if $A_{\epsilon_{j+1}} \subset A_{\epsilon_j}$. We know that there must be ϵ-components with diameters $\geq \delta$ for all ϵ, so this tree must have an infinite branch. This gives a sequence of nested components A_{ϵ_n}, with $\mathrm{diam}(A_{\epsilon_n}) \searrow \delta$. Since the sets are nested, they have a limit, $A_* = \lim_{n \to \infty} A_{\epsilon_n} = \bigcap_n A_{\epsilon_n}$. It then follows that A_* is ϵ-connected for all ϵ and $\mathrm{diam}(A_*) = \delta$. This implies A_* has at least two points in it and so X could not be totally disconnected.

The converse follows directly from the definition. If the diameters of the ϵ-components go to zero, then the connected component of any $x \in X$ is just $\{x\}$, so X is totally disconnected. □

In ϵ-resolution terms, a compact set X is perfect if and only if $\min \mathcal{D}(\epsilon) > 0$ for all $\epsilon > 0$. Alternatively, we can look for isolated points, since these are easy to detect numerically. We say a point $x \in X$ is isolated at resolution ϵ if $d(x, X - x) \geq \epsilon$. For a set to be perfect, the number of ϵ-isolated points, $I(\epsilon)$, must be zero for any $\epsilon > 0$.

The three quantities $C(\epsilon)$, $D(\epsilon)$, and $I(\epsilon)$ form the basis of our computational approach to determining the connectedness properties of data.

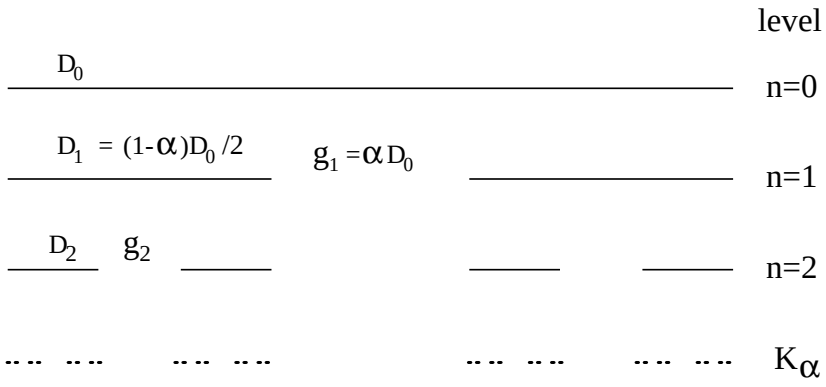


Figure 2.1: The construction of a K_α Cantor set.

2.2.3 Disconnectedness and discreteness growth rates

In some situations of interest — Cantor sets, for example — it is expected that $C(\epsilon) \to \infty$ and $D(\epsilon) \to 0$ as $\epsilon \to 0$. An effective way to evaluate this behavior is to assume an asymptotic form and compute appropriate indices. We use a simple polynomial power law, although finer scales of functions exist [81]. Thus, near $\epsilon = 0$, we look for real numbers γ and δ such that $C(\epsilon) \sim \epsilon^{-\gamma}$ and $D(\epsilon) \sim \epsilon^{\delta}$. The exponents may be found as the following limits:

$$\gamma = \lim_{\epsilon \to 0} \frac{\log C(\epsilon)}{\log(1/\epsilon)} \tag{2.1}$$

$$\delta = \lim_{\epsilon \to 0} \frac{\log D(\epsilon)}{\log \epsilon}. \tag{2.2}$$

If the limits do not exist we can compute the $\liminf$ and the $\limsup$ of each quantity.

The component growth rate, γ, is called the *disconnectedness* index. A positive value of γ implies that the set has infinitely many components. We call δ the *discreteness* index. If δ is positive, the set must be totally disconnected. The function $D(\epsilon)$ relates the size of the ϵ-components to the distance between them, so δ measures the relative rate of decrease of component and gap sizes. If the set is connected or has a finite number of components, then γ and δ are both zero. Three simple examples of Cantor subsets of $\mathbb{R}$ are given below.

Middle-α Cantor sets

These Cantor sets arise in piecewise-linear one-dimensional maps. Let $0 < \alpha < 1$ and consider the Cantor set $K_\alpha \subset [0,1]$ constructed by successively removing the middle α-proportion of each remaining interval. This construction has a natural correspondence with the ϵ-components. At a given level n, there are $C_n = 2^n$ intervals of equal length $D_n = \frac{1}{2}(1-\alpha)D_{n-1}$, separated by gaps of at least $g_n = \alpha D_{n-1}$; see Figure 2.1. With $D_0 = 1$, the recursion relations may be solved to find $D_n = [\frac{1}{2}(1-\alpha)]^n$ and $g_n = \alpha[\frac{1}{2}(1-\alpha)]^{n-1}$, so that

$$\delta = \lim_{n\to\infty} \frac{\log D_n}{\log g_n} = \lim_{n\to\infty} \frac{n \log[\frac{1}{2}(1-\alpha)]}{(n-1)\log[\frac{1}{2}(1-\alpha)] + \log\alpha} = 1 \ ,$$

and

$$\gamma = \lim_{n\to\infty} -\frac{\log C_n}{\log g_n} = \lim_{n\to\infty} \frac{-n \log 2}{(n-1)\log[\frac{1}{2}(1-\alpha)] + \log\alpha} = \frac{\log 2}{\log 2 - \log(1-\alpha)} \ .$$

The discreteness index, δ, is independent of α because the Cantor set is constructed in such a way that the ϵ-components and gaps decrease at the same rate. The disconnectedness index, γ, has the same value as the Hausdorff dimension $\dim_H$. Since $\dim_H < 1$, it follows that these Cantor sets have zero Lebesgue measure. This also follows from the fact that the sum of the gap-lengths is equal to one.

A Cantor set with positive measure

Now consider a Cantor set with gaps that decrease more rapidly. Let K be the subset of $[0,1]$ obtained by successively removing gaps from the center of remaining intervals, with widths $g_n = (\frac{1}{2})^{2n-1}(\frac{1}{10})$ for $n = 1, 2, \ldots$. The total Lebesgue measure of the gaps is just $\frac{1}{10}$, so the measure of K is $\frac{9}{10}$. After a bit of algebra, we find that for $(\frac{1}{2})^{2n+1}(\frac{1}{10}) \le \epsilon < (\frac{1}{2})^{2n-1}(\frac{1}{10})$, there are $C_n = 2^n$ components with diameters $D_n = (\frac{1}{2})^n[\frac{9}{10} + (\frac{1}{2})^n(\frac{1}{10})]$. So,

$$\delta = \lim_{n\to\infty} \frac{n\log(\frac{1}{2}) + \log[\frac{9}{10} + (\frac{1}{2})^n(\frac{1}{10})]}{(2n-1)\log(\frac{1}{2}) + \log(\frac{1}{10})} = \frac{1}{2} \ ,$$

and

$$\gamma = \lim_{n\to\infty} \frac{-n\log 2}{(2n-1)\log(\frac{1}{2}) + \log(\frac{1}{10})} = \frac{1}{2} \ .$$

Since K has positive Lebesgue measure, its Hausdorff dimension is 1. Therefore, the disconnectedness index and the dimension are distinct. In fact, in Chapter 5 we show that for subsets of $\mathbb{R}$, the disconnectedness index is equivalent to the fat fractal exponent.

Cantor sets with zero Hausdorff dimension

It is possible for a Cantor set to have $\dim_H = 0$. Such Cantor sets are observed as invariant sets for the symplectic twist maps we study in Chapter 4. As an example, let $0 < \lambda < 1$ and $c = (1-\lambda)/\lambda$, and suppose the Cantor set has a single gap $g_n = c\lambda^n$ for each $n = 1, 2, \ldots$. We construct the Cantor set as a subset of $[0,1]$, so the constant c is chosen the make the sum

$$\sum_{i=1}^{\infty} g_i = \sum_{i=1}^{\infty} c\lambda^i = 1.$$

This means the Lebesgue measure of the Cantor set is zero. By setting $\epsilon = g_n$, we have that $C(g_n) = n+1$ and therefore

$$\gamma = \lim_{n\to\infty} -\frac{\log(n+1)}{\log g_n} = \lim_{n\to\infty} -\frac{\log(n+1)}{n\log\lambda + \log c} = 0.$$

The diameters of the ϵ-components depend on the exact positioning of the gaps in $[0,1]$. However, if $\epsilon = g_n$, then the largest remaining ϵ-component must be longer than the next gap to be resolved, and its length cannot exceed the sum of all the remaining gaps. That is,

$$g_{n+1} \le D(g_n) \le 1 - \sum_{i=1}^{n} g_i.$$

It follows that

$$\frac{\log g_{n+1}}{\log g_n} \ge \frac{\log D(g_n)}{\log g_n} \ge \frac{\log(1 - \sum_{i=1}^{n} g_i)}{\log g_n}.$$

Now,

$$1 - \sum_{i=1}^{n} g_i = 1 - \sum_{i=1}^{n} c\lambda^i = 1 - \frac{c\lambda}{1-\lambda}(1-\lambda^{n+1}) = \lambda^{n+1}.$$

Thus,

$$\frac{\log c\lambda^{n+1}}{\log c\lambda^n} \geq \frac{\log D(g_n)}{\log g_n} \geq \frac{\log \lambda^{n+1}}{\log c\lambda^n}.$$

Taking the limit as $n \to \infty$ in each expression, we find that $\delta = 1$.

In Section 2.4, we examine some Cantor subsets of $\mathbb{R}^2$. For many of these examples we find that γ is approximately equal to the box-counting dimension, $\dim_B$, and that δ is close to one. We give formal results relating fractal dimensions and the disconnectedness and discreteness indices in Chapter 5.

2.3 Implementation

The theory outlined in the previous section applies to arbitrary compact sets. Recall that compactness is the property that any covering of the set by open subsets can also be accomplished by a finite number of those subsets. This is why it is possible to represent a compact set by a finite number of points. One way to do this is to take a covering of the set by open balls, choose a finite sub-cover, and let the centers of the balls be the finite point-set approximation. Suppose each point in the compact set, X, is within a maximum distance $\rho/2$ of some point in the approximating set, S. The original set and its finite point-set approximation therefore exhibit essentially the same $C(\epsilon)$, $D(\epsilon)$, and $I(\epsilon)$ behavior when $\epsilon > \rho$. For example, if the original set, X, is connected and perfect, then examining S with resolution $\epsilon > \rho$ we see $C_S(\epsilon) = 1$, $\mathrm{diam}(X) - \rho \leq D_S(\epsilon) \leq \mathrm{diam}(X) + \rho$, and $I_S(\epsilon) = 0$. For $\epsilon < \rho$, S has many components and the majority of these are isolated points. For general compact sets we have the bound $C_S(\epsilon - \rho) \geq C_X(\epsilon) \geq C_S(\epsilon + \rho)$ for $\epsilon > \rho$. These ideas are developed further in Chapter 3.

In practice, we are given only the finite point-set approximation and wish to determine the connectedness properties of the underlying set. To do this, we need to estimate the minimum resolution, ρ, from the point-set itself. We can then attempt to extrapolate the limiting behavior of $C(\epsilon)$ and $D(\epsilon)$ from the data with $\epsilon > \rho$. The question that naturally arises is: how confident can we be that the limiting topology is that implied by the data for $\epsilon > \rho$? It is possible to construct examples that appear to be connected down to a given resolution, but are in fact Cantor sets. Conversely, there are connected sets whose finite point-set approximations may appear to be totally disconnected. Both these problems can be addressed to some extent by checking the effect on $C(\epsilon), D(\epsilon)$ and $I(\epsilon)$ of increasing the number of points approximating the underlying set. Ultimately, though, we are restricted by machine precision.

In order to compute $C(\epsilon)$, $D(\epsilon)$ and $I(\epsilon)$ numerically, we need an appropriate way to organize the finite point-set data. The structure we use is a graph — the Minimal Spanning Tree (MST) [67]. This choice was inspired by K. Yip's work on computer recognition of orbit structures in two dimensional area-preserving maps [89]. The following section describes why the edge lengths of the MST naturally define the resolutions at which one should see a change in the number of components. In fact, the MST of a data set contains all the information we need about the ϵ-components.

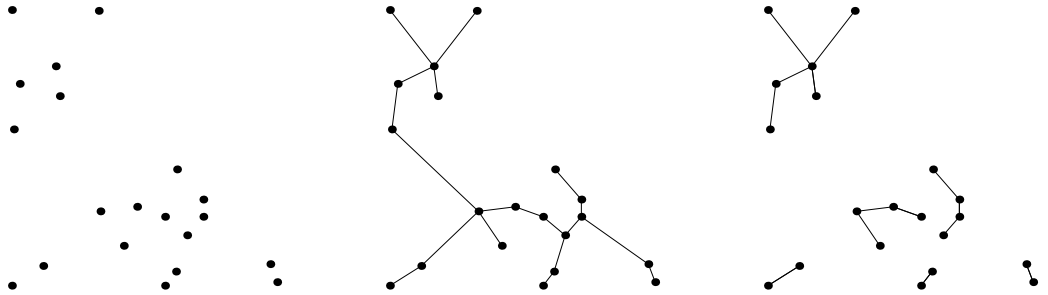


Figure 2.2: A finite set of points (a), its minimal spanning tree (b), and the nearest neighbor graph (c). The weight of an edge is the Euclidean distance between the points it joins.

2.3.1 Minimal spanning trees

A *graph* is a structure consisting of a finite set of points called *vertices* and a list of pairs of these points called *edges*. A *weighted graph* has a weight or cost assigned to each edge. A graph is *connected* if there is a sequence of edges (a *path*) joining any point to any other point. When it contains no closed paths, a connected graph is called a *tree*. Given a graph, a *spanning tree* is a subgraph that is a tree and contains all the vertices from the original graph. For our purposes, the vertices are data points in a metric space (usually $\mathbb{R}^n$), the edges are lines joining two points, and the weight of an edge is just the metric distance between the two points joined by that edge.

The *minimal spanning tree* (MST) of a graph is a spanning tree of that graph that has minimal total weight. It is unique when all the edge weights are distinct. For our application, we build the MST from the complete graph — i.e., one that has an edge between every pair of points. The intuitive way to do this is to delete edges from the complete graph as follows. Start with the longest edge (largest weight), and then remove successively shorter ones, provided that doing so leaves the sub-graph connected. The algorithm ceases when removing any edge would leave a disconnected graph. An alternative constructive algorithm, due to R.S. Prim [68], is more readily implemented on a computer. The initial subtree consists of any point, its nearest neighbor, and the edge between them. The subtree grows by adding the point that has the shortest distance to some point (called its parent) in the subtree. This step is repeated until all points are in the tree. The cost of this algorithm is $O(n^2)$ where n is the number of points in the set. It is also possible to construct the MST from a special graph called the Delaunay triangulation (c.f., Section 3.4). For sets in the plane, this results in an algorithm with cost $O(n \log n)$. See [67] for details.

The property of connecting closest points makes the MST a natural structure for our analysis. This is illustrated by the data set and MST presented in Figure 2.2. Notice that the MST bridges the "gap" between two subsets with the shortest edge possible, automatically organizing the point-set in a way that corresponds closely to ϵ-connected components of the set. The following results formalize this intuition. More results about minimal spanning trees and their relationship to clustering can be found in [90].

First, we show how the edge lengths in the MST relate to ϵ-connectedness and the ϵ-components.

Lemma 4. *If S is a finite set of points that is ϵ-connected for all $\epsilon \geq \epsilon_0 > 0$, then the longest edge in the MST has length $l < \epsilon_0$.*

Conversely, if the MST of a finite set of points, S, has largest edge length $l < \epsilon_0$, then S is ϵ-connected for all $\epsilon \geq \epsilon_0$.

Proof. Recall from Section 2.2 that S is ϵ-connected if and only if there is an ϵ-chain joining every pair of points in S. This implies that in the complete graph over the points of S, it is possible to move from one vertex to any other along edges of length less than ϵ. Therefore, in the intuitive construction of the MST, all edges of length greater than ϵ_0 are removed, since they could not disconnect the graph. It follows that the MST of S can have no edge longer than ϵ_0.

Conversely, the MST is a connected graph, so given any points $x, y \in S$, we can find an ϵ_0-chain: $x_0 = x, \ldots, x_n = y$, with $d(x_i, x_{i+1}) < \epsilon_0$. This implies that S is ϵ-connected for all $\epsilon \geq \epsilon_0$. □

The next result shows that the MST is a suitable structure for representing ϵ-components. First, we note that every edge in a MST defines a partition of S (the point set) into two subsets P and Q containing the points from the two subtrees generated by removing the edge. It follows that:

Lemma 5. *Removing an edge from a MST generates two subgraphs, each of which is a MST of its points.*

Suppose that the longest edge has length l and assume that this edge length is unique. If we remove this edge, we are left with two subsets P and Q that are ϵ-connected for some $\epsilon < l$. These subsets genuinely are ϵ-components since $d(P, Q) = l > \epsilon$. If there are $n > 1$ edges of length l then removing them leaves $n + 1$ connected components.

The minimal spanning tree, once constructed, holds all the information we need to deduce connectedness properties from ϵ-components. The ϵ-components of the set S correspond directly to the connected components of the sub-graph that is generated by removing edges from the MST of lengths, $l \geq \epsilon$. Also, the edge lengths of the MST automatically give the resolutions at which one sees a change in the number of components. We now discuss ways to extract this information efficiently from the MST.

2.3.2 Practical issues

This section addresses some of the details of the numerical implementation. In particular, it describes how we compute $C(\epsilon)$, $D(\epsilon)$, and $I(\epsilon)$. We give Matlab code for the following algorithms in an appendix to this thesis.

Finding the number of ϵ-components is straightforward: $C(\epsilon)$ is just one more than the number of edges with length greater than ϵ. We build the MST using Prim's algorithm with the Euclidean distance between points as the edge weight, and store the MST as three arrays. The first lists the coordinates of the data points in arbitrary order, the second contains the index of the parent of each point, and the third, the length of the edge between the point and its parent. This allows us to compute $C(\epsilon)$ from the cumulative distribution of edge-lengths, obtained by simply sorting the array.

We know that a point will be isolated at resolution ϵ if all the MST edges incident on that point are longer than ϵ. One way to compute $I(\epsilon)$ is to delete all edges longer than ϵ from the MST and then count the number of isolated points. Another method is to use the *nearest neighbor graph* (NNG); see Figure 2.2 for an example. The nodes of this directed graph are again the data points; an edge points from one node, x_i, to another, x_j, if

$$d(x_i, x_j) \leq d(x_i, x) \quad \text{for all} \quad x \in S \setminus \{x_i, x_j\}. \tag{2.3}$$

This relationship is not reflexive, i.e., $x_i \to x_j$ does not imply $x_j \to x_i$. Ignoring their directions, the edges of the NNG are a subset of those in the MST. It is therefore very simple to build the NNG from the MST, again by sorting the MST edges by length and then recording the shortest edge incident on each point. The number of ϵ-isolated points, $I(\epsilon)$, is just the number of edges in the NNG that are longer than ϵ (counting an edge that points both ways twice). $I(\epsilon)$ is then a cumulative distribution of edge-lengths in the NNG.

To find the diameter of an ϵ-component we first need to list the points in the ϵ-component. This involves a tree walk algorithm, which we will discuss in the following three paragraphs. Finding the Euclidean metric diameter of a set of n points in the plane is an order $O(n \log n)$ algorithm, since the computations can be restricted to points lying on the boundary of the convex hull. For subsets of higher-dimensional spaces, this restriction does not necessarily help [67]. Instead, the algorithm we use is the brute-force comparison of distances between all pairs of points, which is $O(n^2)$. Finding the diameter using the supremum metric ($d_{\sup}(x, y) = \sup_i |x_i - y_i|$) is faster, since all that is required is to find the maximum and minimum value in each coordinate over the set of points, and this is linear in the number of points.

A tree walk on the MST is not particularly efficient since its root is arbitrary and little can be said in general about its branching structure. There is, however, a natural binary structure to the MST, since each edge defines a partition of the MST into two components. This fact can be used to construct a binary tree from the MST, which is then faster to search.

The binary tree represents information about the MST in the following way. Nodes (vertices) of the binary tree represent edges from the MST, ordered by length; the root is the longest edge. The two children of an edge node are the longest edges of the two sub-MSTs generated by removing that edge. The leaves (nodes with no children) of the binary tree represent the data points; the parent node of a leaf is the shortest edge incident to that point in the MST. In essence, each edge node in the binary tree represents a connected component of the MST. Given a value of ϵ, each ϵ-connected component is represented by an edge-node with length less than ϵ, but whose parent has length greater than ϵ. The points in an ϵ-component can be found by listing the leaves "under" its representative node. This binary tree can be built so that each edge-node has information about the component it represents — for example, the number of points in the component.

The time to list the points in a component is proportional to the depth of the tree, which in turn depends on the number of points in the data set and how well the tree is balanced. For fairly uniformly distributed data, the tree is well balanced, but for non-uniformly distributed data this is not the case and the time to list the points in an ϵ-component can become very long.

Our algorithms to build the MST and the binary component tree are not the most efficient implementations. There is a large literature on special data structures for graph algorithms [9] that could be exploited to give faster implementations.

Finally, we must address the problem of how to determine the finest appropriate resolution, ρ, as discussed at the beginning of Section 2.3. To do this, we examine how the number of isolated points in the ϵ-decomposition of the set varies with resolution —i.e., the function $I(\epsilon)$. In all of the examples below, the underlying sets are perfect, so the finite point-set approximation is "bad" at any resolution for which there are isolated points. It follows that the resolution at which we start to see isolated points is an estimate of ρ, i.e. $\rho = \inf\{\epsilon : I(\epsilon) = 0\}$. The validity of this approach is supported by the numerical evidence in the next section; the data for $C(\epsilon)$ and $D(\epsilon)$ diverge from their true values at the resolution when isolated points are first detected.

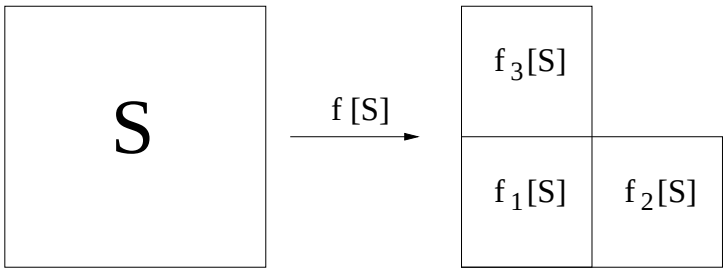


Figure 2.3: Template for the iterated function system that generates the Sierpinski triangle relatives

2.4 Examples

In this section we present some examples that illustrate the behavior of the number of ϵ-components, $C(\epsilon)$, the largest diameter, $D(\epsilon)$, and the number of isolated points, $I(\epsilon)$, for fractals with different topology. The goal is to show that these quantities give useful information about the topology.

The first examples are relatives of the Sierpinski triangle. These sets are generated from a family of iterated function systems (IFS). The Hausdorff dimension of each set is identical, even though they have different topological structure, as the disconnectedness and discreteness indices highlight. For these fractals, we show that the cutoff resolution decreases when the number of data points is increased, which is not surprising, since more points sampled from an attractor constitute a better approximation of the underlying set. We also vary the way in which the data cover the set and find that for a fixed number of points the cutoff resolution is a minimum when the data are uniformly distributed. Again, this is exactly what we expect, since isolated points appear at larger values of ϵ when points are not evenly spaced.

We next present a number of Cantor-set examples to illustrate different types of scaling behavior in $C(\epsilon)$ and $D(\epsilon)$. We observe that, for Cantor sets with zero Lebesgue measure, the computed value of γ is approximately equal to the dimension of the set. For a Cantor set with positive measure, though, γ and the dimension are significantly different.

2.4.1 Relatives of the Sierpinski triangle

Sierpinski triangle relatives are attracting fixed sets of a family of iterated function systems:

$$S = f[S] = f_1[S] \cup f_2[S] \cup f_3[S].$$

In each case, the functions f_i are similarity transformations[1] of the unit square with contraction ratio $\frac{1}{2}$, as illustrated by the template in Figure 2.3. The functions that generate the Sierpinski triangle, Figure 2.4, are simple contractions composed with a translation; the generators of the examples in Figure 2.11 and Figure 2.13 involve additional rotation or reflection symmetries of the square. There are 232 different fractals in this family [66] and their Hausdorff dimensions are identical: $\dim_H = \log 3 / \log 2$. Their topology, however, ranges from simply connected to connected to totally disconnected to a class of examples with infinitely many connected components of non-zero diameter [71]. This range of structure makes them ideal test cases for our techniques.

[1]Recall from Section 1.2.3 that similarity transformation, S, is an affine transformation that contracts or dilates distance uniformly; i.e., for all x and y, there is a positive number r such that $|S(x) - S(y)| = r|x - y|$.

It is easy to generate a finite number of points on the attractor of an iterated function system. One way (Barnsley's chaos game [4]) is to chose an initial point x_0 in the domain of the IFS and then record its trajectory under the iteration $x_{n+1} = f_{i_n}(x_n)$, setting $i_n = 1, 2$ or 3 with probability p_1, p_2 and p_3 respectively. If x_0 is in the attractor then its entire orbit is in the attractor; if not, the iterates converge to it. In the examples below, we choose $x_0 = (0, 0)$. Thus, the orbit can be viewed as a random sampling of the attractor by a finite number of points. When $p_1 = p_2 = p_3 = \frac{1}{3}$ the data cover the fractal uniformly; if the probabilities are not equal the distribution of points is nonuniform and their density approximates a multifractal measure. In the appendix, we give Matlab code for generating the Sierpinski triangle relatives in the above manner.

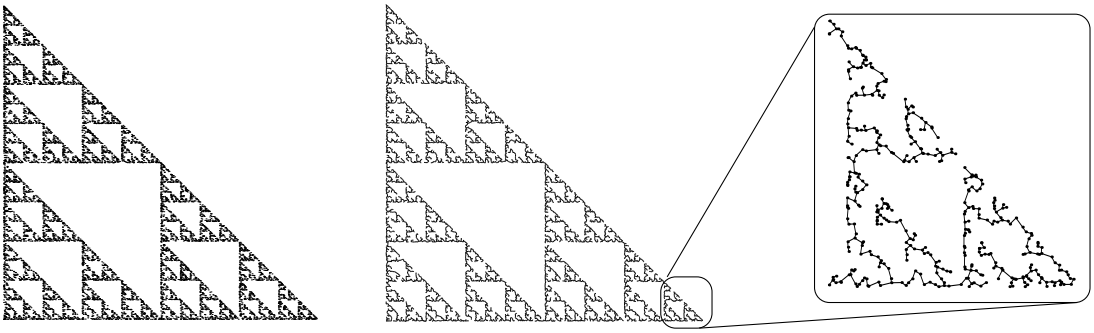
The Sierpinski triangle

The generating functions for the Sierpinski triangle are:

$$\begin{aligned} f_1(x, y) &= \tfrac{1}{2}(x, y) \\ f_2(x, y) &= \tfrac{1}{2}(x + 1, y) \\ f_3(x, y) &= \tfrac{1}{2}(x, y + 1). \end{aligned} \tag{2.4}$$

A finite point-set approximation to the triangle and the corresponding minimal spanning tree are shown in Figure 2.4. The underlying set is connected and perfect, so we expect to see $C(\epsilon) = 1$, $D(\epsilon) = \sqrt{2}$, and $I(\epsilon) = 0$ for $\epsilon > \rho$. This is corroborated by the calculations of $C(\epsilon)$ and $D(\epsilon)$ for 10^4 and 10^5 point approximations to the triangle, as shown in Figure 2.5. We see that for ϵ above a threshold value, the computed values of $C(\epsilon)$ and $D(\epsilon)$ are in exact agreement with our expectations. The point at which $C(\epsilon)$ and $D(\epsilon)$ deviate from the ideal values is the value of ϵ at which the number of isolated points, $I(\epsilon)$, becomes positive. This ϵ value is, of course, the cutoff resolution ρ discussed in Section 2.3.2. At finer resolutions — i.e., $\epsilon < \rho$ — we see a sharp transition in the number of connected components from one to the number of points in the set; the diameters show a correspondingly sharp decrease. Both these effects are due to the narrow distribution of edge-lengths of the MST. Clearly, the value of ρ depends on the number of points, N, covering the set. For the 10^4-point approximation, $\rho \approx 0.008$ and for $N = 10^5$, $\rho \approx 0.0022$; We expect the relationship to be $\rho \approx 1/\sqrt{N}$, since the data are homogeneously distributed on a subset of $\mathbb{R}^2$. This is supported by the data in Figure 2.6(a). Here, we plot cutoff resolution versus the number of points for $10^3 \leq N \leq 10^5$; the slope of the least-squares fit line is -0.58.

The results discussed so far are for uniformly distributed data. In general dynamical systems, though, orbits cover attractors nonuniformly. Therefore, we want to understand the effects that nonuniformly distributed points have on the connectedness data. As described earlier, we can change the way an orbit covers the IFS attractor by choosing the functions f_1, f_2, and f_3 with different probabilities. To generate Figure 2.7(a), for example, we set $p_1 = 0.05$ and $p_2 = p_3 = 0.475$. This highly nonuniform distribution of points induces perceptible changes in the $C(\epsilon)$, $D(\epsilon)$ and $I(\epsilon)$ data, as shown in Figure 2.8, but the graphs remain qualitatively similar to those in Figure 2.5. The cutoff resolution is significantly larger: $\rho \approx 0.04$ compared with 0.008 for the uniform distribution with the same number of points. The growth in the number of ϵ-components for $\epsilon < \rho$ is also less rapid than that for the uniform data. Both of these changes are due to a greater spread in the edge-lengths of the MST. The geometry of the distribution is reflected in the graph of $D(\epsilon)$; the densely covered diagonal means that $D(\epsilon) = \sqrt{2}$ for ϵ values significantly less than ρ.



(a) (b) (c)

Figure 2.4: (a) 10^4 points uniformly distributed on the Sierpinski triangle. (b) The corresponding MST. (c) A close up of the bottom right corner of the MST.

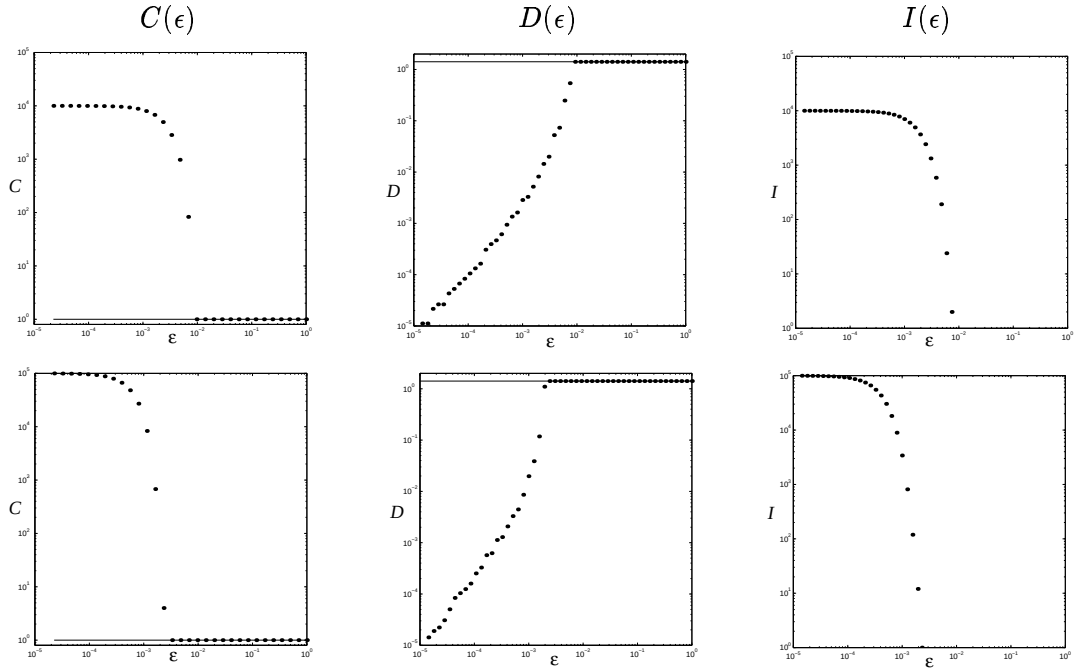


Figure 2.5: $C(\epsilon), D(\epsilon)$ and $I(\epsilon)$ for the Sierpinski triangle. The top row gives results for 10^4 uniformly distributed points on the fractal and the bottom row for 10^5 points. All axes are logarithmic. The horizontal axis range is $10^{-5} < \epsilon < 1$. The solid lines represent $C(\epsilon)$ and $D(\epsilon)$ for ideal data; the dots are the computed values.

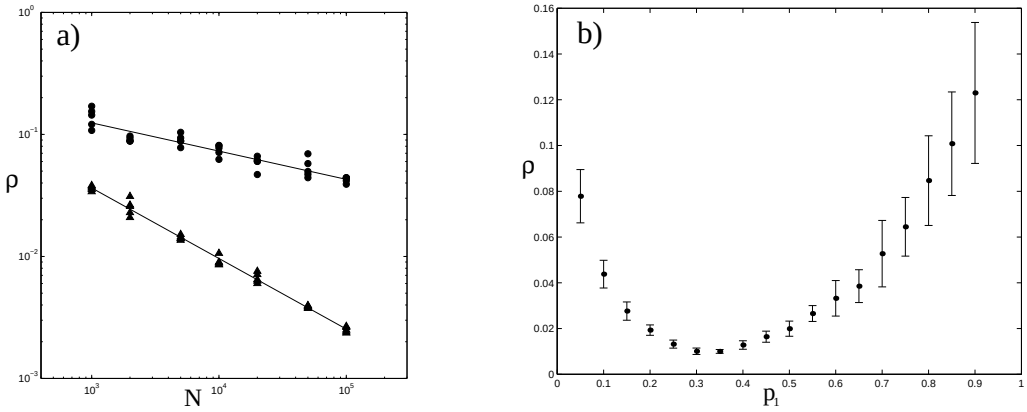
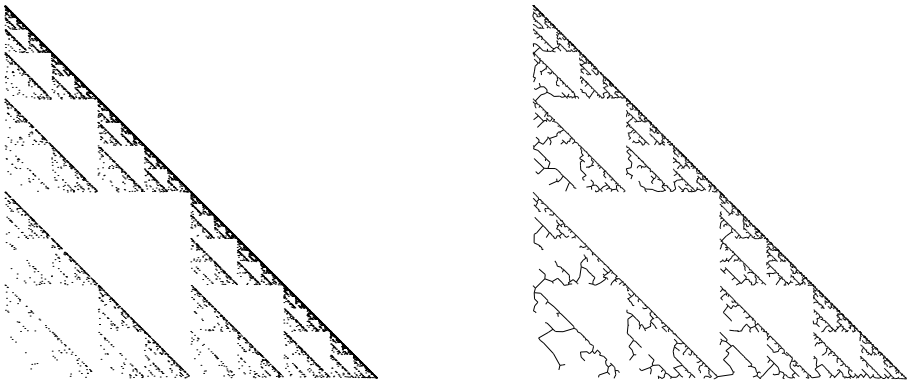


Figure 2.6: (a) Cutoff resolution, ρ, as a function of the number of points, $10^3 \leq N \leq 10^5$, covering the Sierpinski triangle for two values of p_1; ● marks data for the nonuniform distribution with $p_1 = 0.05$ and ▲ marks data for $p_1 = \frac{1}{3}$, i.e., a uniform distribution. (b) Cutoff resolution as a function of p_1 for 10^4 data points on the Sierpinski triangle. The error bars are the standard deviation about the mean of twenty calculations of ρ for each value of p_1.

We can lower the cutoff resolution by increasing the number of data points but, as shown in Figure 2.6(a), the rate at which ρ decreases is $\rho \approx N^{-0.23}$, significantly slower than that for the uniform data. This is an interesting result that deserves further investigation. The distribution of edge-lengths in the MST must depend on the distribution of points, and perhaps there is a way to formally relate the two distributions. For a connected set, ρ is the largest edge-length in the MST. Thus, given an analytic form for the distribution of edge-lengths, it may be possible to predict the scaling of ρ with the number of points. We arrive at similar questions in Chapter 4 when we consider point distributions on invariant circles of the standard map.

Finally, the graph in Figure 2.6(b) summarizes the variation of the cutoff resolution with the nonuniformity in the distribution of points. The measure of nonuniformity in the data is p_1, the probability of choosing f_1; we set $p_2 = p_3 = (1 - p_1)/2$. Twenty orbits of 10^4 points were generated for values of p_1 in the range 0.05 to 0.9. The cutoff resolution reaches a minimum at $p_1 = \frac{1}{3}$, i.e., for uniformly distributed data, as we expect. The other feature to note is that the standard deviation also depends on the distribution, and is greatest for highly nonuniform data. This is because when one function in the IFS is chosen with very low probability, there is greater variability in the way an orbit fills out the attractor; this in turn leads to greater variation in the edge-lengths of the MST.

We conclude that for moderate amounts of nonuniformity (for this example, $0.2 \leq p_1 \leq 0.5$) the cutoff resolution is at a level comparable to that for perfectly uniform data and our techniques are not adversely affected. For highly nonuniform coverings of an attractor, significantly more data points are needed to reach the same cutoff resolutions as for uniform data. The only effect this has is to generate inconclusive, rather than incorrect, diagnoses of the topology of the underlying set.



(a) (b)

Figure 2.7: (a) 10^4 points on the Sierpinski triangle generated by setting $p_1 = 0.05$ and $p_2 = p_3 = 0.475$. (b) The corresponding MST.

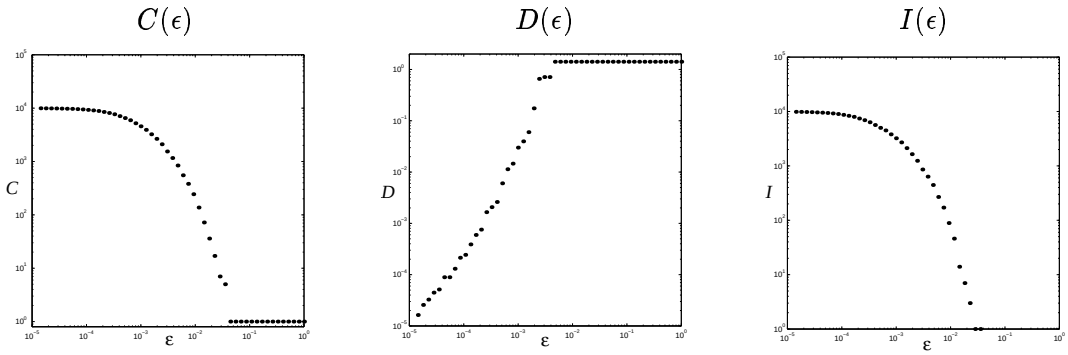
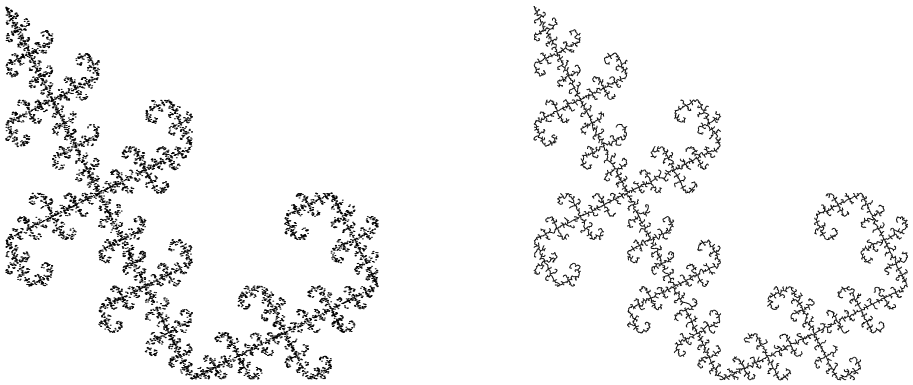


Figure 2.8: $C(\epsilon), D(\epsilon)$ and $I(\epsilon)$ for the nonuniformly distributed data set in Figure 2.7. All axes are logarithmic. The horizontal axis range is $10^{-5} < \epsilon < 1$.



(a) (b)

Figure 2.9: (a) 10^4 points on the set generated by (2.5) and (b) the corresponding MST.

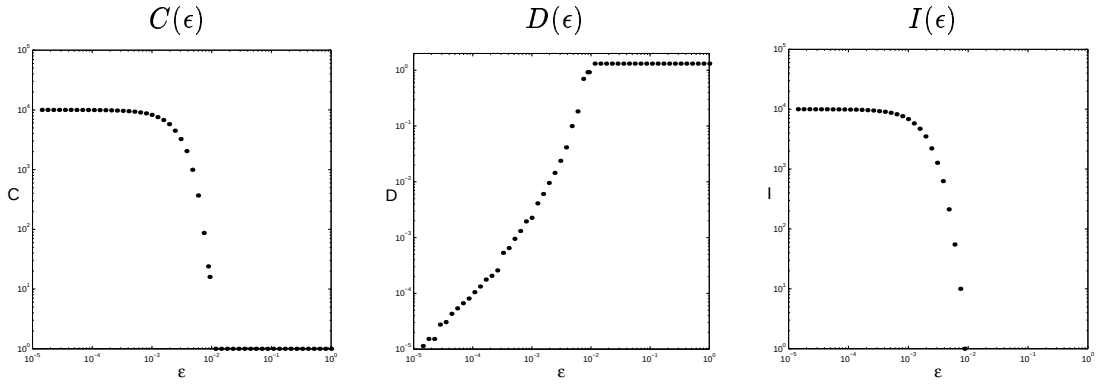
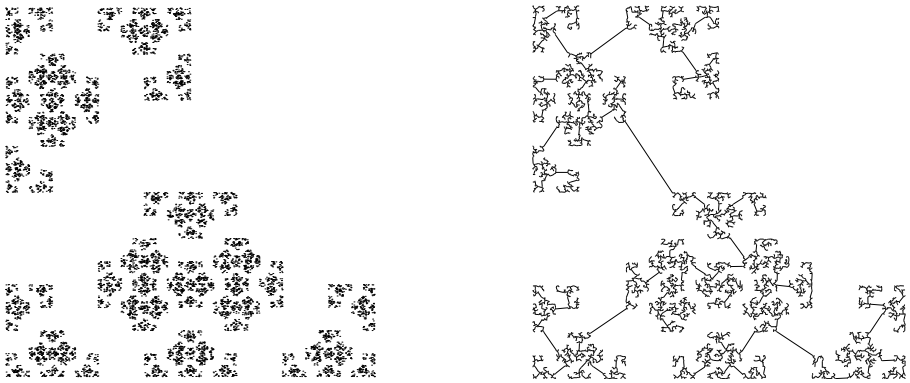


Figure 2.10: $C(\epsilon), D(\epsilon)$ and $I(\epsilon)$ for the simply connected gasket relative. The data is for 10^4 points on the set. All axes are logarithmic. The horizontal axis range is $10^{-5} < \epsilon < 1$.



(a) (b)

Figure 2.11: (a) 10^4 points on the Cantor set generated by (2.6) and (b) the corresponding MST.

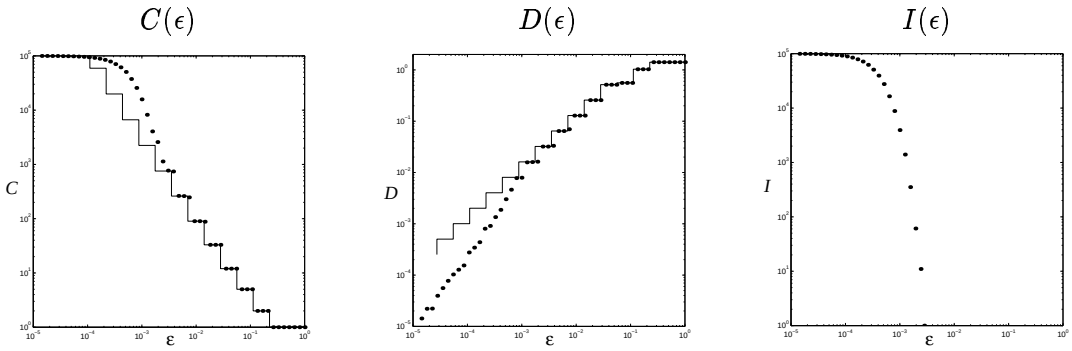


Figure 2.12: $C(\epsilon)$, $D(\epsilon)$ and $I(\epsilon)$ for 10^5 points uniformly distributed over the Cantor set triangle relative. All axes are logarithmic. The horizontal axis range is $10^{-5} < \epsilon < 1$. The solid lines represent $C(\epsilon)$ and $D(\epsilon)$ for ideal data; the dots are the computed values.

A simply connected relative

The fractal shown in figure 2.9 is generated by:

$$\begin{aligned} f_1(x,y) &= \tfrac{1}{2}(-x+1,-y+1) \\ f_2(x,y) &= \tfrac{1}{2}(-y+2,x) \\ f_3(x,y) &= \tfrac{1}{2}(x,y+1). \end{aligned} \tag{2.5}$$

This set is simply connected, in contrast to the Sierpinski triangle which has infinitely many holes. Despite this difference, we see that data for $C(\epsilon)$ and $D(\epsilon)$ are almost identical to those for the triangle. The problem of detecting holes is addressed in Chapter 3.

A Cantor set relative

Figure 2.11 shows the attractor for the iterated function system generated by

$$\begin{aligned} f_1(x,y) &= \tfrac{1}{2}(-y+1, x) \\ f_2(x,y) &= \tfrac{1}{2}(y+1, x) \\ f_3(x,y) &= \tfrac{1}{2}(y, -x+2). \end{aligned} \tag{2.6}$$

This fractal is a Cantor set, so we should see $C(\epsilon) \to \infty$ and $D(\epsilon) \to 0$ as $\epsilon \to 0$. We can derive analytic expressions for both functions, since we know that the set is three copies of itself at half the size. We call the region that disconnects two ϵ-components a "gap". Its "width" is the metric distance between the two components. Clearly, the set has a single component whenever ϵ is greater than the width of the largest gap, g_0. As ϵ decreases, subsequent components are resolved at gap sizes $g_n = g_0/2^n$. The diameters of the components follow the same pattern after some initial transient behavior. Thus for $n \geq 3$,

$$D(\epsilon) = D(g_2)/2^{n-2} \quad \text{for} \quad g_{n+1} < \epsilon \leq g_n. \tag{2.7}$$

This gives $\delta = 1$, confirming that the set is totally disconnected.

The number of components, $C(\epsilon)$, is a little harder to determine. Since $C(\epsilon)$ is just one more than the total number of gaps, we calculate it by first deriving an expression for the latter. Let N_n be the number of gaps of size g_n. Since the fractal contains three copies of itself, one might think that $N_n = 3^n$. By careful inspection of the IFS template we find that this is not the whole story — some gaps merge into one. In fact, with $N_0 = 1$ we have the recursion:

$$N_n = \begin{cases} 3N_{n-1} & \text{if } n \text{ is odd,} \\ 3N_{n-1} - 2 \cdot 3^{n/2-1} & \text{if } n \text{ is even.} \end{cases}$$

These can be solved to find:

$$N_n = \begin{cases} 2 \cdot 3^{n-1} + 3^{(n-1)/2} & \text{if } n \text{ is odd,} \\ 2 \cdot 3^{n-1} + 3^{n/2-1} & \text{if } n \text{ is even.} \end{cases}$$

We then have that for $g_{n+1} < \epsilon \leq g_n$,

$$C(\epsilon) = 1 + \sum_{j=0}^{n} N_j = \begin{cases} 3^n + 2 \cdot 3^{(n-1)/2} & \text{if } n \text{ is odd,} \\ 3^n + 3^{n/2} & \text{if } n \text{ is even.} \end{cases}$$

The leading power is the same for all n, so we may use either case to evaluate the limit:

$$\gamma = \limsup_{n\to\infty} \frac{\log(C(g_n))}{\log(1/g_n)} = \lim_{n\to\infty} \frac{\log[3^n + 3^{n/2}]}{\log[2^n/g_0]} = \frac{\log 3}{\log 2}\,.$$

Since there are no isolated points, the attractor is perfect and therefore a Cantor set, as claimed.

We can see in Figure 2.12 that the numerical calculations agree very well with the theory down to the cutoff resolution $\rho \approx 0.003$. When $\epsilon < \rho$, the computed values of $C(\epsilon)$ are larger than the predicted values because isolated points are counted as extra components. For still-smaller values of ϵ, every point is resolved as an isolated point and the $C(\epsilon)$ curve levels off. The meaningful portion of the data — between these extremes — shows a staircase periodicity about a linear trend. The slope of the linear trend is an estimate of γ. We determine γ numerically,

using a least-squares fit, to be 1.41 ± 0.05. This is lower than the true limiting value given above (1.58) because of second-order effects at the relatively large values of ϵ for which the $C(\epsilon)$ data are valid. We estimate the slope of the true curve over the same range to be 1.48, which is closer to the value computed above.

The numerically calculated values of $D(\epsilon)$ also show a staircase periodicity about a linear trend. The data have a systematic bias for the jumps at slightly larger values of ϵ than predicted by theory. This is due to finite-data effects: the points are not filling out the corners, so edges in the MST are a little longer than the "true" gaps and the diameters of the ϵ-components are a little less. The data points fall below the true curve when $\epsilon < \rho$ because inter-point distances are comparable to the inter-component distances at these resolutions. Since $D(\epsilon)$ measures the largest diameter, the flat tails of the $D(\epsilon)$ data in Figure 2.12 are due to the presence of at least one triple/pair combination within a distance ϵ of each other, whilst almost all the other points have become isolated. The slope of the linear trend of $D(\epsilon)$ is an estimate of δ. We estimate the slope over the range $\rho < \epsilon < 0.06$ because the first few steps are shallower than the limiting trend. Using a least squares fit, again, we calculate δ to be 1.00 ± 0.03, as predicted above.

The cutoff resolution for data from this IFS varies with the number of points and their distribution in exactly the same manner as the data from the Sierpinski triangle. For moderately nonuniform data, estimates of γ and δ are the same as those given above. When the data is very unevenly distributed, ρ increases significantly and the ϵ range may be too small to allow an estimate of the slope. Again, this leads to inconclusive results rather than incorrect ones.

A relative with infinitely many connected components

A third triangle relative, shown in Figure 2.13, is generated by the following similarities:

$$\begin{aligned} f_1(x,y) &= \tfrac{1}{2}(x,y) \\ f_2(x,y) &= \tfrac{1}{2}(y+1,-x+1) \\ f_3(x,y) &= \tfrac{1}{2}(x,y+1). \end{aligned} \tag{2.8}$$

The attractor for this system has infinitely many connected components, yet is not totally disconnected because the components have positive diameters. We therefore expect to see $C(\epsilon) \to \infty$ and $D(\epsilon) \to 1$ as $\epsilon \to 0$. Again, the gaps decrease simply as $g_n = g_0/2^n$. This time, the number of components is just

$$C(\epsilon) = \tfrac{1}{2}(3^{n+1}+1) \quad \text{for} \quad g_{n+1} < \epsilon \le g_n, \tag{2.9}$$

giving $\gamma = \log 3/\log 2$. The largest ϵ-component always contains the line segment $x = 0$, $0 \le y \le 1$, so $D(\epsilon) \to 1$.

The graphs of $C(\epsilon)$ and $D(\epsilon)$ in Figure 2.14 reflect this; the former has characteristics similar to those of the Cantor set above, but $D(\epsilon)$ looks like that for the Sierpinski triangle. The slope of the $C(\epsilon)$ staircase is estimated from the data for a 10^5 point approximation, giving a value of $\gamma = 1.55 \pm 0.03$. This is in very close agreement with the theoretical value of $\gamma = \log 3/\log 2 \approx 1.585$.

2.4.2 Cantor sets in the plane

One of our objectives is to use our techniques to identify and characterize phase-space structures in dynamical systems. Cantor sets are often present in chaotic dynamical systems, so it is useful to examine some simple Cantor set examples to gain a better understanding of the different

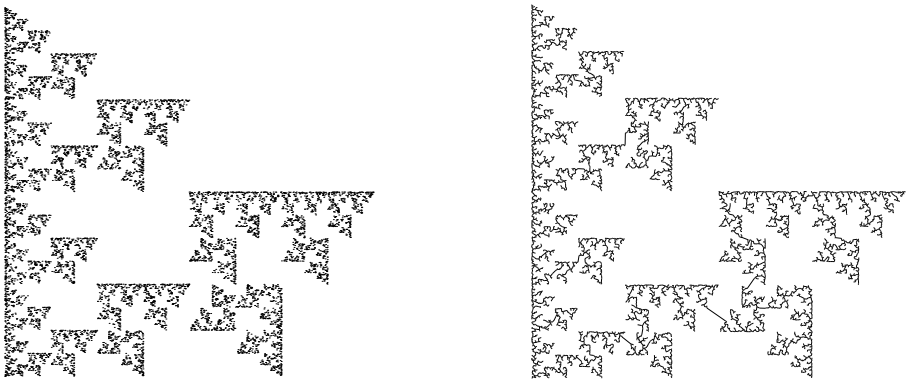


Figure 2.13: (a) 10^4 points on the fractal generated by (2.8) and (b) the corresponding MST.

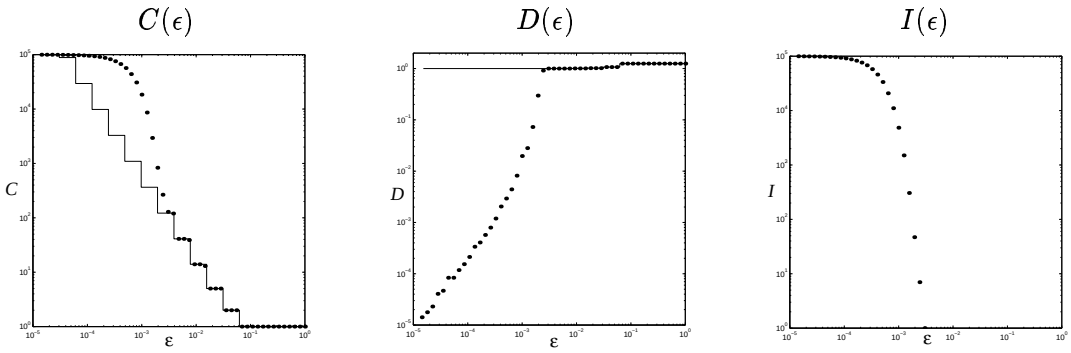


Figure 2.14: $C(\epsilon)$, $D(\epsilon)$ and $I(\epsilon)$ for a triangle relative with infinitely many components. Again, the data is for 10^5 points uniformly distributed on the set. All axes are logarithmic. The horizontal axis is $10^{-5} < \epsilon < 1$. The solid line represents $C(\epsilon)$ and $D(\epsilon)$ for ideal data; the dots are the computed values.

types of scaling that can occur in the $C(\epsilon)$ and $D(\epsilon)$ graphs. In Figure 2.15, Figure 2.17, and Figure 2.19, we show five Cantor sets in the plane. In each case, the orbit has 50000 points. Four of these have zero Lebesgue measure and one (Figure 2.17(b)) has positive measure, so it is termed a fat Cantor set (this is analogous to the term "fat fractal" for fractals with positive measure [22]). All are attractors of iterated function systems of the form

$$S = f[S] = f_1[S] \cup f_2[S] \cup f_3[S] \cup f_4[S].$$

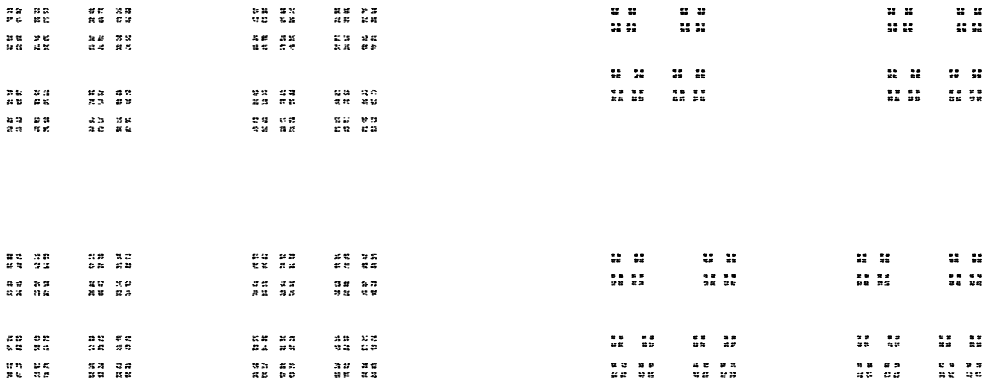
The generating functions, f_i, become increasingly complex in this series of examples. The three simplest involve only affine transformations and a fourth uses conformal functions; the functions that generate the fat Cantor set cannot be written in closed form. The geometric structure of each set is reflected in the type of staircase seen in the graphs of $C(\epsilon)$ and $D(\epsilon)$. For the four examples with zero Lebesgue measure, we expect to see $\gamma = \dim_B$, the box-counting dimension; this is supported by our results, summarized in Table 2.1. For the Cantor set with positive measure, the value of γ is significantly different from the dimension. We again observe, for all of the examples, that the cutoff resolution, ρ, is well approximated by the ϵ-value where the number of isolated points ceases to be zero. Again, this is because all of the underlying sets are perfect.

We start with a simple example where each f_i is a similarity transformation with contraction ratio $\frac{1}{3}$, as shown in Figure 2.15(a). The numerical calculations of $C(\epsilon)$, $D(\epsilon)$, and $I(\epsilon)$ are presented in the top row of Figure 2.16. These graphs show staircase scaling behavior similar to the triangle-relative Cantor set presented in Section 2.4.1; this makes sense because both are generated by iterated function systems of similarity transformations. For this example, the jumps in $C(\epsilon)$ and $D(\epsilon)$ are at $\epsilon_n = 1/3^n$; because there are four self-similar copies at one third the size, we see $C(\epsilon_n) = 4^n$ and $D(\epsilon_n) = 1/3^n$, which gives the theoretically determined limits of $\gamma = \log 4/\log 3 \approx 1.262$ and $\delta = 1$. This is in close agreement with our numerical estimate of $\gamma = 1.23 \pm 0.05$ and $\delta = 0.97 \pm 0.02$.

The second example, Figure 2.15(b), is also generated by similarities. This time, the lower two have a contraction ratio of $\frac{1}{3}$ and the upper two have a ratio of $\frac{1}{4}$. As can be seen from the second row of Figure 2.16, this leads to a more complicated staircase pattern in the $C(\epsilon)$ and $D(\epsilon)$ graphs. Jumps in these graphs occur at values of ϵ corresponding to edge lengths in the MST. The structure from the IFS means these edge lengths are of the form $l(\frac{1}{3})^m(\frac{1}{4})^n$, for all integers m and n, where l is one of the two longest edges. Values of γ and δ, presented in Table 2.1, are again very close to the expected values.

To generate the set in Figure 2.17(a), more-general affine transformations are used, each contracting by $\frac{1}{3}$ horizontally and $\frac{1}{4}$ vertically. The corresponding graphs of $C(\epsilon)$ and $D(\epsilon)$ in Figure 2.18 show the now-familiar staircase scaling pattern. Compared to the second Cantor set example, the larger steps in these graphs reflect the more regular geometric structure of the set.

The fourth example is a Cantor set with positive Lebesgue measure and therefore a dimension of 2. It is possible to represent this set as the attractor of an iterated function system of the general form above. The functions involved, however, are limits of piecewise linear approximations and it is not possible to write them in closed form. Instead, we generate the set as the cross product of two positive measure Cantor subsets of the unit interval. These sets are constructed as follows: at each level, $n \geq 1$, 2^{n-1} gaps of length $a/2^{pn-1}$ are removed from the center of an interval remaining from level $n-1$. The sum of the gap lengths is $a/(2^{p-1}-1)$; choosing p and a to make this length less than one ensures the Cantor set has positive measure. It is easy to recursively generate the end points of the gaps (down to some level) and these points are used as the finite point-set approximation. For the set in figure 2.17(b), we set $a = 2/3$ and $p = 2$ for



(a) (b)

Figure 2.15: Cantor sets generated by iterated function systems of four similarity transformations. Both sets have 50000 points. (a) Similarities with contraction ratio $\frac{1}{3}$. (b) The upper two similarities have ratio $\frac{1}{4}$ and the lower two have ratio $\frac{1}{3}$.

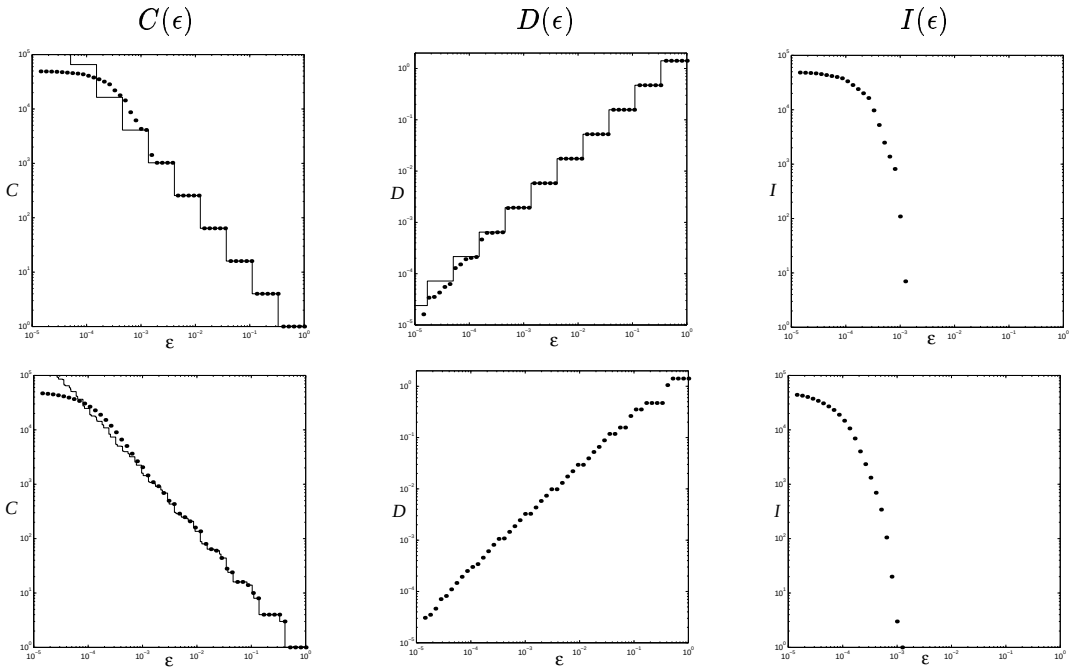
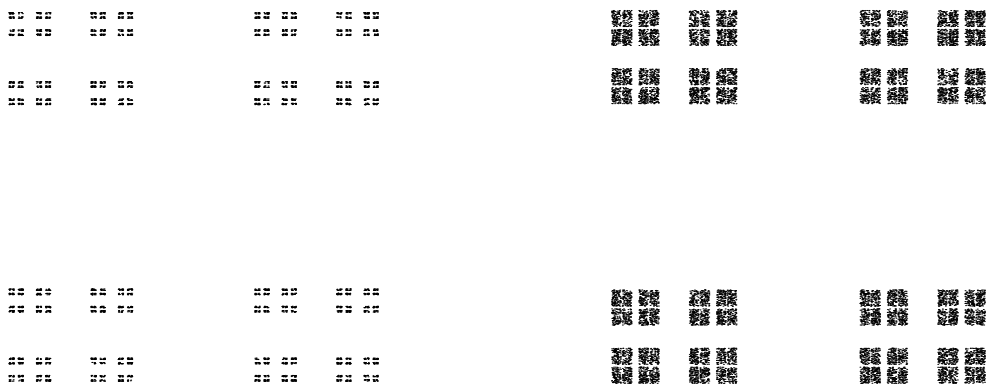


Figure 2.16: $C(\epsilon), D(\epsilon)$ and $I(\epsilon)$ for the Cantor sets in figure 2.15. The top row is data for Figure 2.15(a); the second row is for Figure 2.15(b). All axes are logarithmic. The horizontal axis range is $10^{-5} < \epsilon < 1$. The solid lines represent $C(\epsilon)$ and $D(\epsilon)$ for ideal data; the dots are the computed values.



(a) (b)

Figure 2.17: Two Cantor sets with largest gaps of $1/2$ and $1/3$. (a) A set generated by an IFS of four affine transformations with horizontal contraction of $\frac{1}{3}$ and vertical contraction of $\frac{1}{4}$. (b) A fat Cantor set, generated as the cross product of two Cantor sets of positive measure in the real line.

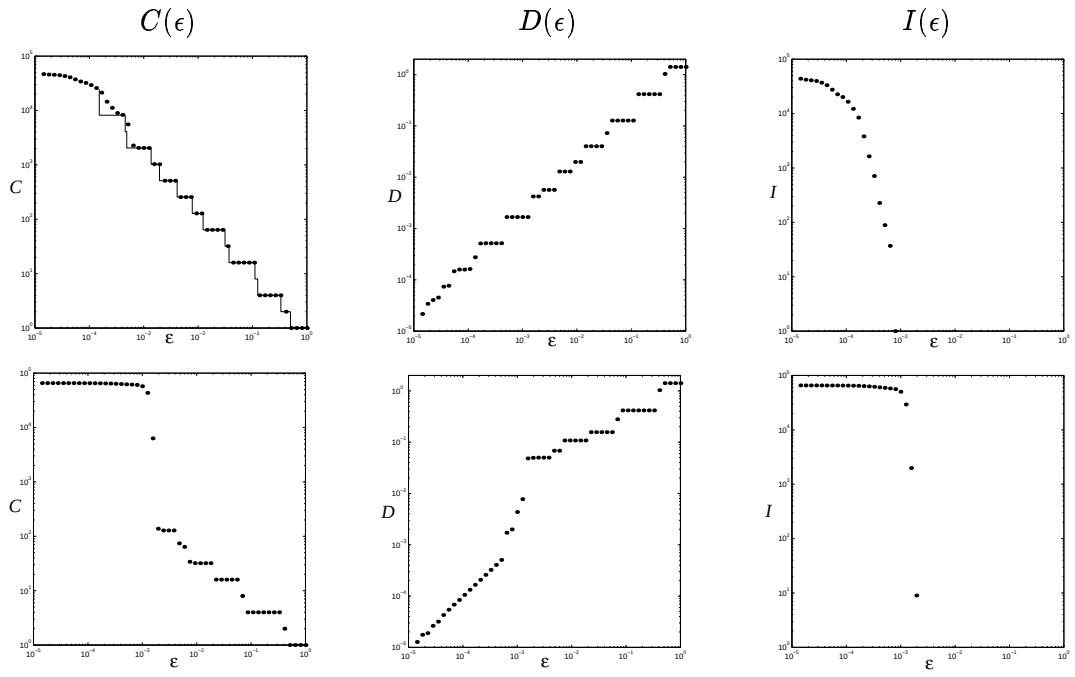
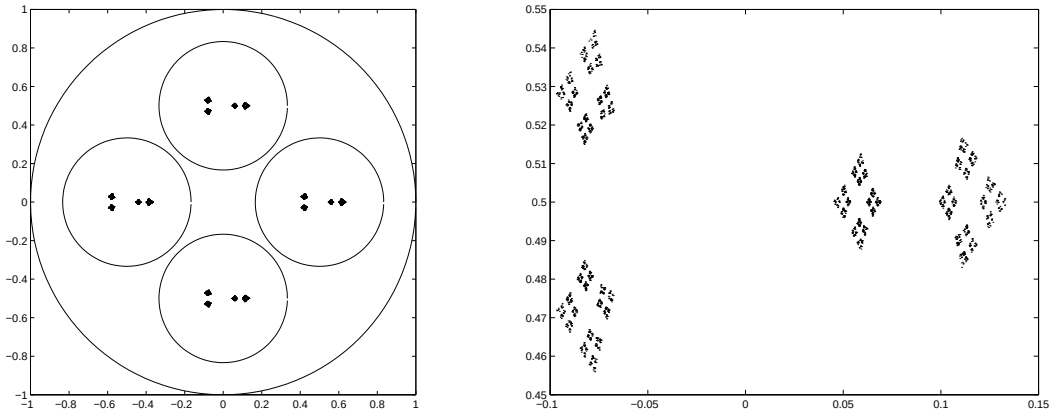


Figure 2.18: $C(\epsilon), D(\epsilon)$ and $I(\epsilon)$ for the 2-D Cantor sets in Figure 2.17. The top row is data for Figure 2.17(a); the second row for Figure 2.17(b), the fat Cantor set; All axes are logarithmic. The horizontal axis range is $10^{-5} < \epsilon < 1$.



(a) (b)

Figure 2.19: A Cantor set generated by an IFS consisting of four nonlinear affine transformations, each mapping the unit circle into a circle of radius $\frac{1}{3}$. (a) The data set with circle boundaries. (b) A close up of one of the four clusters.

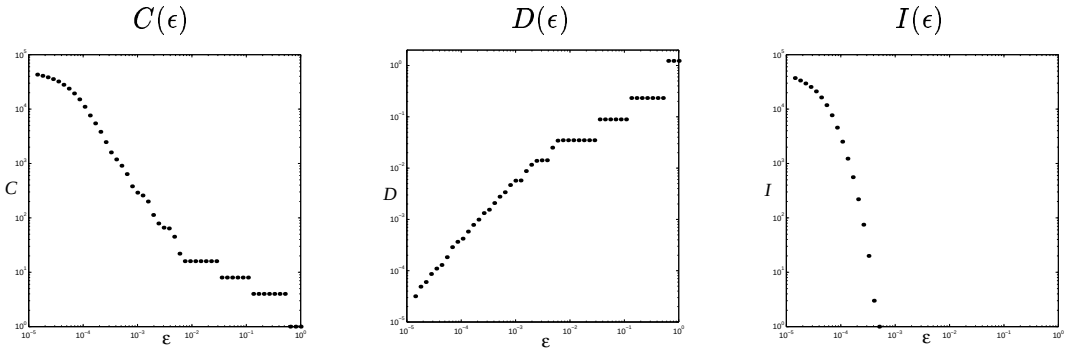


Figure 2.20: $C(\epsilon), D(\epsilon)$ and $I(\epsilon)$ for the nonlinear Cantor set of Figure 2.19. All axes are logarithmic. The horizontal axis range is $10^{-5} < \epsilon < 1$.

the horizontal cross-section, and $a = 2$, $p = 3$ for the vertical one. The behavior of $C(\epsilon)$ and $D(\epsilon)$, shown in the bottom row of Figure 2.18, is not unlike that of the previous example. The slopes are significantly shallower, however, because the gaps are decreasing at a faster rate than the component diameters, giving $\delta \approx 0.46$. The component growth rate, γ, is approximately 0.80, which is clearly distinct from the box-counting (and Hausdorff) dimension of 2.

The final IFS example uses nonlinear, conformal transformations. A function, F, is conformal if its derivative matrix at each point, $DF(x)$ is a similarity transformation. For the set illustrated in Figure 2.19, the functions are of the form $f_k(z) = z^2/3 + c_k$, where $z = x + iy$, and the translations, c_k for $k = 1, \ldots, 4$, take the values $\{\pm\frac{1}{2}, \pm\frac{i}{2}\}$. Notice that although we choose the f_i with equal probability, the nonlinearity introduces a nonuniformity to the distribution of points over the Cantor set. The cutoff resolution $\rho \approx 5 \times 10^{-4}$ is nevertheless comparable to the previous examples of uniformly distributed data. Scaling in the graphs of $C(\epsilon)$ and $D(\epsilon)$ occurs in two distinct ϵ intervals; see Figure 2.20. For $0.005 < \epsilon < 1$, there are three shallow steps reflecting the large-scale structure that is visible in figure 2.19(a). The

Table 2.1: A summary of the values of γ and δ for Cantor subsets of the plane. The numbers are estimated using a least-squares linear fit to logarithmic plots of $C(\epsilon)$ and $D(\epsilon)$, respectively; the error margins are estimated by varying the scaling range. The second column gives the box-counting dimension, $\mathrm{d}_{\mathrm{box}\,B}$, for each set; these numbers are computed using formulas from Falconer [23].

Data Set	$\dim_B$	γ	δ
Fig. 2.15(a)	1.262	1.23 ± 0.02	0.96 ± 0.04
Fig. 2.15(b)	1.126	1.11 ± 0.02	1.00 ± 0.03
Fig. 2.17(a)	1.131	1.13 ± 0.01	0.98 ± 0.03
Fig. 2.17(b)	2	0.80 ± 0.05	0.46 ± 0.05
Fig. 2.19	$1.21 < \dim_B < 1.34$	1.36 ± 0.03	0.95 ± 0.05

second portion of the data, for $\rho < \epsilon < 0.005$, has a steeper slope, corresponding to the limiting small-scale structure of the set. The values of γ and δ given in Table 2.1 are slopes of the $C(\epsilon)$ and $D(\epsilon)$ over the interval $\rho < \epsilon < 0.005$. We find, as for the previous zero-measure Cantor sets, that γ is close to the box-counting dimension and $\delta \approx 1$.

2.5 Concluding remarks

Our results demonstrate that the minimal spanning tree of a finite set of points can provide accurate information about the topology of the underlying set, down to a numerically computable resolution $\rho > 0$. In particular, we are able to identify sets that are connected, totally disconnected, or have infinitely many connected components with non-zero diameter. Confidence in the extrapolation can be increased by sampling more points in order to get a better approximation to the underlying set, but of course we are still ultimately restricted by the machine precision. Connected sets have disconnectedness index $\gamma = 0$, and discreteness index $\delta = 0$. Based on the examples in the chapter, we conjecture that Cantor sets with zero Lebesgue measure have γ equal to the box-counting dimension and $\delta = 1$. Results in this direction are given in Chapter 5. The fat Cantor set example shows that γ and dimension are not the same when the set has positive Lebesgue measure.

The points in most of the examples of this chapter are fairly evenly distributed over the underlying set. However, it is often the case that orbits cover an attractor in a highly nonuniform way. Previous work on characterizing fractals has dealt with this by introducing a concept of dimension for measures, and developing a theory of multifractals [24]. The results of Section 2.4.1 show that our techniques are most effective for uniformly distributed data; for a fixed number of data points, the cutoff resolution, ρ, is minimal when the points are evenly spread over the attractor. Our techniques can still give valid results for highly nonuniform data; the difference is that ρ may be too high in these cases to make strong statements about the topology of the underlying set. It may be possible to weight the edges of the MST by the density of the data distribution and thereby lower ρ in these situations.

A natural question that arises from studying the Sierpinski triangle relatives is how to distinguish between simply connected sets and ones with holes. This involves reformulating concepts from homology theory by introducing a resolution parameter, similar to the way we treat the definition of connectedness. This is the subject of the next chapter.