Eurographics Symposium on Point-Based Graphics (2004)
M. Alexa, S. Rusinkiewicz, (Editors)

Topological estimation using witness complexes

Vin de Silva and Gunnar Carlsson[†]

Department of Mathematics, Stanford University, California, USA.

Abstract

This paper tackles the problem of computing topological invariants of geometric objects in a robust manner, using only point cloud data sampled from the object. It is now widely recognised that this kind of topological analysis can give qualitative information about data sets which is not readily available by other means. In particular, it can be an aid to visualisation of high dimensional data. Standard simplicial complexes for approximating the topological type of the underlying space (such as Čech, Rips, or α-shape) produce simplicial complexes whose vertex set has the same size as the underlying set of point cloud data. Such constructions are sometimes still tractable, but are wasteful (of computing resources) since the homotopy types of the underlying objects are generally realisable on much smaller vertex sets. We obtain smaller complexes by choosing a set of 'landmark' points from our data set, and then constructing a "witness complex" on this set using ideas motivated by the usual Delaunay complex in Euclidean space. The key idea is that the remaining (non-landmark) data points are used as witnesses to the existence of edges or simplices spanned by combinations of landmark points.

Our construction generalises the topology-preserving graphs of Martinetz and Schulten [MS94] in two directions. First, it produces a simplicial complex rather than a graph. Secondly it actually produces a nested family of simplicial complexes, which represent the data at different feature scales, suitable for calculating persistent homology [ELZ00, ZC04]. We find that in addition to the complexes being smaller, they also provide (in a precise sense) a better picture of the homology, with less noise, than the full scale constructions using all the data points. We illustrate the use of these complexes in qualitatively analyzing a data set of 3×3 *pixel patches studied by David Mumford et al [LPM03].*

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computing Methodologies]: Computer Graphics [Computational Geometry and Object Modeling]

1. Simplicial Approximation

Given a point-cloud dataset sampled from an underlying space X, it is often desirable to build a simplicial complex $\mathcal{S}$ approximating the geometric or topological structure of X. For example, a laser scanning device applied to a solid object might return the coordinates of thousands of points lying on the objects 2-dimensional surface. A standard problem is to build a triangular mesh from this unstructured collection of points, perhaps for visual rendering. Such a mesh should be a close *geometrical* approximation to X itself. Examples of provably successful algorithms for surface reconstruction can be found in the work of Amenta et al [ACDL02, AB99].

In this paper we focus on the analogous *topological* problem: how to find a representation of the data which can be used to compute topological invariants, robustly and efficiently. For example, the figure on the left is a (noisy) circle, and the figure on the right has three loop-shaped petals. How

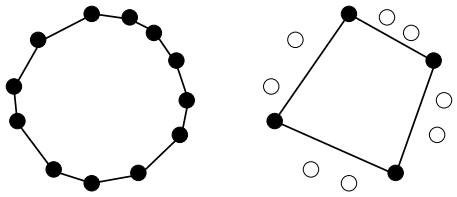
does one extract this kind of topological information automatically and reliably? There is increasing demand for such

[†] Both authors have been supported in part by NSF grant DMS-0101364.

techniques; for example, Carlsson et al [CCdS03] present algorithms for automatic feature-detection which depend on being able to make accurate topological calculations.

A natural approach is to represent the data by a simplicial complex $\mathcal{S}$, using the data points as vertices, and adding edges, triangles and higher-dimensional cells according to suitable rules. From $\mathcal{S}$ one can compute *Betti numbers* $b^k = b^k(\mathcal{S})$; this is a standard procedure in classical algebraic topology [Mas91] for counting the k-dimensional holes of a simplicial complex. If $\mathcal{S}$ is a faithful topological representation of X, then this effectively computes the numbers $b^k(X)$, by proxy. The figures underlying the examples shown above are distinguished by their first Betti number b^1: a circle has $b^1 = 1$, and a three-loop clover has $b^1 = 3$. The goal is to find an algorithm which produces simplicial complexes for these data sets which have the same properties.

The distinction between geometrical and topological approximation may be seen in the following pictures. Twelve



points have been sampled from a circle. On the left is a 12-vertex 12-edge simplicial complex which closely follows the contours of the hidden circle. On the right is a 4-vertex 4-edge simplicial complex which is geometrically only a crude approximation to the circle, but it still has the correct topology. When the goal is to calculate Betti numbers, this is a more efficient way of getting the right answer.

The purpose of this paper is to introduce, in systematic detail, the *witness complex* construction. In its simplest form, this builds a simplicial complex from a point-cloud data set Z and a choice of vertices $L \subset Z$ called *landmark points*, and no other input parameters. More precisely, the construction depends only on the matrix $D = D(L,Z)$ of distances between landmark points and data points. Any suitable metric can be used to define D, including data-dependent metrics such as the shortest-path distances in a graph [TdSL00].

Witness complexes can be regarded as approximations to the restricted Delaunay triangulation, but the construction sidesteps the curse of dimensionality associated with Delaunay computations; specifically, the extrinsic dimension of the data set plays no role in the complexity of the algorithm. The mechanism for doing this is explained by a new theorem [dS03] which gives an alternate definition of Delaunay triangulation, equivalent to the classical definition but which can be more easily adapted to the point-cloud data framework.

In addition to the parameter-free witness complex $\mathrm{W}(D)$, we also define three families of complexes $\mathrm{W}(D;R,\nu)$, where $\nu = 0,1,2$, dependent on a "feature size" parameter R. Any such family can be used to define *persistent homology*, which combines Betti number analysis with a notion of size ("persistence") for the holes that are detected. Thus we can exploit the powerful techniques of Edelsbrunner, Letscher and Zomorodian [ELZ00] to generate so-called persistence interval graphs for each family of complexes [ZC04]. The topological information carried in such an interval graph is richer and more robust than a single Betti number by itself.

Our long-term goal is to put topological data analysis on a sound, quantifiable footing. To this end we give two examples. The first example consists of points on the 2-sphere. We compare the performance of witness complexes to a standard construction, the Rips complex, in the task of obtaining the correct Betti numbers for the sphere. The second example comes from a natural image database provided by David Mumford [LPM03] which exhibits rather subtle statistical behaviour. We feel that these examples vindicate the use of witness complexes in topological data analysis.

We stress that our purpose in this paper is to provide a detailed, motivated description of a family of constructions we have found to be useful in deriving topological estimates from real data. We do not address the question of theoretical 'correctness' in this paper, nor do we give precise comparisons of the behaviour of the different complexes described here. Some of these questions seem quite difficult. That said, we can certainly articulate some of the immediate advantages of the witness complex construction.

- It produces much smaller complexes than other constructions. In fact, we can determine the size of the complex by our choice of the number of landmark points.
- Other than the number and choice of landmark points, there are no parameters that need to be set arbitrarily, expect for an optional "neighbourhood size" parameter that is used in one variation.
- They are defined for data sets in any metric space, not necessarily in Euclidean space. This is clearly an advantage in many settings.
- It provides a more robust calculation for homology than other methods, at least in the examples we have studied.

The remainder of Section 1 is taken up with background material, including a brief discussion of persistent homology. Section 2 motivates and describes witness complexes in some detail. The examples are discussed in Section 3.

1.1. Abstract simplicial complexes

An *abstract simplicial complex* $\mathcal{S}$ is specified by the following data:

- A vertex set Z.
- A rule specifying when a 'p-simplex' $\sigma = [z_0 z_1 \dots z_p]$ belongs to $\mathcal{S}$; here the vertices $z_0, z_1, \dots, z_p$ of σ are distinct

elements of Z, listed in some order which we fix once and for all.

- Each p-simplex σ has $p+1$ faces which are $(p-1)$-simplices; each face is obtained by deleting one of the vertices $z_0, z_1, \ldots, z_p$. The membership rule has the property that if σ belongs to $\mathcal{S}$ then all of its faces belong to $\mathcal{S}$.

We can compute the Betti numbers of an abstract simplicial complex using a standard linear algebra recipe. The details are not important here, so we refer the reader to any standard text in algebraic topology, such as [Mas91].

In order to calculate the Betti numbers of X correctly from an approximation $\mathcal{S}$, the technical condition is that *$\mathcal{S}$ has the same homotopy type as X*. We adopt this language without further comment.

1.2. Čech, Rips and α-shape complexes

We now discuss three well-known constructions. The Čech complex, and the closely related Rips complex, are the simplest constructions of an abstract simplicial complex from a point-cloud dataset Z. For $R \geq 0$, we define $\check{\mathrm{C}}\mathrm{ech}(Z,R)$, with vertex set Z, according to the following rule:

- the p-simplex $\sigma = [z_0 z_1 \ldots z_p]$ belongs to $\check{\mathrm{C}}\mathrm{ech}(Z,R)$ iff the closed Euclidean balls $B(z_j, R/2)$, $j = 0, 1, \ldots, p$, have non-empty common intersection.

In technical language, the Čech complex is the *nerve* [Spa66] of the collection of metric balls $\{B(z,R/2) : z \in Z\}$. In fact, $\check{\mathrm{C}}\mathrm{ech}(Z,R)$ has the same homotopy type as the union of these balls. If Z is sampled finely from a continuous space X then this union of balls, for a suitable value of R, often has the same homotopy type as X, which is exactly what we want.

The related complex $\mathrm{Rips}(Z,R)$, with vertex set Z, has a membership test which is much easier to evaluate:

- the p-simplex $\sigma = [z_0 z_1 \ldots z_p]$ belongs to $\mathrm{Rips}(Z,R)$ iff for every edge $[z_j z_k]$, $0 \leq j < k \leq p$, we have $|z_j - z_k| \leq R$.

$\mathrm{Rips}(Z,R)$ is the largest simplicial complex having the same 1-skeleton (i.e. vertices and edges) as $\check{\mathrm{C}}\mathrm{ech}(Z,R)$. It is convenient to implement, because one only needs to store the edges and vertices; a higher-dimensional simplex belongs to $\mathrm{Rips}(Z,R)$ iff all of its edges belong.

Remark 1 The definition of the Rips complex makes sense for points in an arbitrary metric space.

Both constructions tend to be extremely inefficient. Whenever k points form a cluster of diameter at most R, there is a corresponding $(k-1)$-dimensional simplex in $\check{\mathrm{C}}\mathrm{ech}(Z,R)$ and $\mathrm{Rips}(Z,R)$. In the language of graph theory, the vertices form a *clique*. This can lead to prohibitively large complexes, even when the underlying topology is very simple. An elegant solution to the problem of large cliques was found by Edelsbrunner [Ede95]. Each point $z \in Z$ is contained in a Voronoi cell V_z in the Voronoi diagram for Z. The α-shape complex $\mathrm{A}(Z,R)$, with vertex set Z, is defined by the following rule.

- the p-simplex $\sigma = [z_0 z_1 \ldots z_p]$ belongs to $\mathrm{A}(Z,R)$ iff the convex sets $B(z_j, R/2) \cap V_{z_j}$, $j = 0, 1, \ldots, p$ have non-empty common intersection.

$\mathrm{A}(Z,R)$ has the same homotopy type as $\check{\mathrm{C}}\mathrm{ech}(Z,R)$, so we recover the same topological information. However, the definition implies that $\mathrm{A}(Z,R)$ is a subcomplex of the Delaunay triangulation $\mathrm{Del}(Z)$. This makes it considerably less wasteful of simplices than $\check{\mathrm{C}}\mathrm{ech}(Z,R)$. The main pitfall is that one needs to be able to compute Delaunay triangulations. There is a "curse of dimensionality" with respect to the space in which the data are embedded.

The motivation behind witness complexes is to find a construction which makes frugal use of simplices, but is nonetheless easily computed. What we lose in the bargain is the remarkable theoretical tractability of the Čech and α-shape complexes; much of our current understanding of witness complexes is heuristic. Of course, it would be nice to have some theoretical guarantees, but that is beyond the scope of this paper.

1.3. Persistent homology

The last piece of background knowledge, and an essential tool in our work, is persistent homology. Each of the three constructions in the preceding section produces a *nested* family of complexes. For example, whenever $R \leq R'$, we have an inclusion $\check{\mathrm{C}}\mathrm{ech}(Z,R) \subseteq \check{\mathrm{C}}\mathrm{ech}(Z,R')$. Algebraically, we can compute persistent Betti numbers $b^k(R,R')$ for every pair (R,R'). The interpretation is that $b^k(R,R')$ counts the number of k-dimensional holes in $\check{\mathrm{C}}\mathrm{ech}(Z,R)$ which remain open when we thicken the complex to $\check{\mathrm{C}}\mathrm{ech}(Z,R')$. We can do the same for the Rips and α-shape families of complexes.

Rapid calculation of persistent Betti numbers for all pairs (R,R') is possible, thanks to the definitive algorithm due to Edelsbrunner, Letscher and Zomorodian [ELZ00]. This produces interval graphs, which, for each dimension k, consist of a set of closed intervals lying above an axis parametrised by R. The presence of an interval $[R_0, R_1]$ indicates that a homology cycle ("hole") appears for the first time when R increases to R_0, and persists until $R = R_1$ at which point it closes up. Long intervals correspond to large holes, which may be regarded as genuine features. Small intervals indicate holes which close up almost as soon as they are formed; these may be regarded as noise. An algebraic analysis of the algorithm appears in [ZC04].

Persistent homology is undoubtedly the correct tool for estimating topological invariants from real data sets; individual Betti numbers by themselves are highly unstable. We therefore take the trouble to define nested families of witness complexes.

2. Witness complexes

Witness complexes are intended to behave like Delaunay triangulations computed in the intrinsic geometry of the data

set Z. A subset $L \subset Z$ of landmark points is chosen to be the vertex set, and the remaining points play a role in determining which simplices occur in the complex. However, it does not pay to be too pedantic about the interpretation of "Delaunay triangulation", particularly since we wish to define nested families for persistent homology; so the actual definitions may not be exactly as expected.

2.1. Definition of $\mathrm{W}(D)$

Let D be an $n \times N$ matrix of non-negative entries, regarded as the matrix of distances between a set of n landmarks and N data points. We define the (strict) witness complex $\mathrm{W}_\infty(D)$, with vertex set $\{1,2,\ldots,n\}$, as follows.

- the edge $\sigma = [ab]$ belongs to $\mathrm{W}_\infty(D)$ iff there exists a data point $1 \leq i \leq N$ such that $D(a,i)$ and $D(b,i)$ are the smallest two entries in the i-th column of D, in some order.
- by induction in p: suppose all the faces of the p-simplex $\sigma = [a_0 a_1 \ldots a_p]$ belong to $\mathrm{W}_\infty(D)$. Then σ itself belongs to $\mathrm{W}_\infty(D)$ iff there exists a data point $1 \leq i \leq N$ such that $D(a_0,i), D(a_1,i), \ldots, D(a_p,i)$ are the smallest $p+1$ entries in the i-th column of D, in some order.

In either case, we refer to i as a "witness" to the existence of σ.

Analogous to the Rips complex, there is a "lazy" version of the witness complex. We define $\mathrm{W}_1(D) \supseteq \mathrm{W}_\infty(D)$ formally as follows.

- $\mathrm{W}_1(D)$ has the same 1-skeleton as $\mathrm{W}_\infty(D)$.
- the p-simplex $\sigma = [a_0 a_1 \ldots a_p]$ belongs to $\mathrm{W}_1(D)$ iff all of its edges belong to $\mathrm{W}_1(D)$.

In other words, $\mathrm{W}_1(D)$ is the largest simplicial complex having the same vertices and edges as $\mathrm{W}_\infty(D)$. In practice we seldom use $\mathrm{W}_\infty(D)$ since its computation is fussier, and we write $\mathrm{W}(D)$ to mean $\mathrm{W}_1(D)$.

Remark 2 We are free to apply this construction to any distance matrix D, using the Euclidean or any other metric. An important alternative choice is the *intrinsic graph metric* D_G, which is defined by computing shortest paths in a suitable graph G on the set of all data points; for example the graph representing a relation "is a close neighbour of". In some situations this represents the intrinsic geometry of the data far better than the original D. See [TdSL00] for a major application of this idea.

2.2. The weak witnesses theorem

The strict witness complex $\mathrm{W}_\infty(D)$ can be motivated by comparing it with the the Delaunay triangulation in Euclidean space. A theorem is necessary to make the motivation complete.

Suppose $L \subset \mathbb{R}^D$ is a collection of points. Recall that the Delaunay triangulation $\mathrm{Del}(L)$ contains the p-simplex $\sigma = [\ell_0 \ell_1 \ldots \ell_p]$ precisely when there exists a point $x \in \mathbb{R}^D$ such that x is equidistant from the points $\ell_0, \ell_1, \ldots, \ell_p$ and has no nearer neighbour in L. We call x a *strong witness* to the existence of σ, with respect to L. When the set of allowed witnesses is discrete, there is no point looking for strong witnesses because they exist with probability 0. We say that $x \in \mathbb{R}^D$ is a *weak witness* for σ with respect to L iff $|x-\ell_i| \leq |x-\ell|$ for all $i = 0,1,\ldots,p$ and $\ell \in L \setminus \{\ell_0, \ell_1, \ldots, \ell_p\}$; in other words, if the $p+1$ nearest neighbours of x in L are $\ell_0, \ell_1, \ldots, \ell_p$ (in a sense that tolerates equality). Our definition of $\mathrm{W}_\infty(D)$ can be formulated in terms of weak witnesses: σ is a p-simplex of $\mathrm{W}_\infty(D)$ iff it has a weak witness and all of its cells have weak witnesses.

Theorem 3 Suppose $L \subset \mathbb{R}^D$ is a finite collection of points, and $\ell_0, \ell_1, \ldots, \ell_p \in L$. Then $\sigma = [\ell_0 \ell_1 \ldots \ell_p]$ has a strong witness with respect to L iff σ and all its cells have weak witnesses with respect to L.

In other words, instead of looking for a single strong witness, one looks for a whole constellation of weak witnesses. The case $p = 1$ was discussed by Martinetz and Schulten in [MS94], justifying the definition of the graph which forms the 1-skeleton of the complex $\mathrm{MS}_\infty(L,Z)$. The general result is due to de Silva [dS03].

In the light of this theorem, the definition of $\mathrm{W}_\infty(D)$ is a natural way to try to define the intrinsic Delaunay triangulation for a space represented only by point-cloud data.

2.3. Choosing the landmarks

We recommend obtaining the landmark set in one of two ways: randomly, or by *maxmin*. Both methods seem to give reasonable results. Maxmin is the following inductive procedure. Initialise by selecting $\ell_1 \in Z$ randomly. Inductively, if $\ell_1, \ell_2, \ldots, \ell_{i-1}$ have been chosen, let $\ell_i \in Z \setminus \{\ell_1, \ell_2, \ldots, \ell_{i-1}\}$ be the data point which maximises the function

$$z \mapsto \min\{D(z,\ell_1), D(z,\ell_2), \ldots, D(z,\ell_{i-1})\},$$

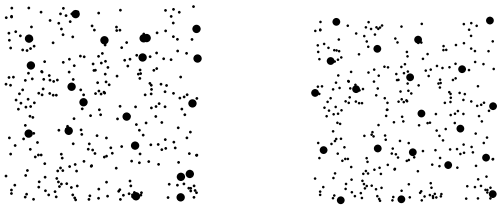
where D is the metric. Continue until the desired number of landmark points have been chosen.

Maxmin gives more evenly spaced landmarks, but it also has a tendency to pick out extremal points. Examples of both are shown in the next figure: random on the left, maxmin on the right.

The number of landmarks should be chosen by setting a lower bound on the ratio N/n. We do not have a systematic answer to what this lower bound should be, but $N/n \geq 20$ seems to work quite well for data sampled from a two-dimensional surface.

Remark 4 It may seem natural to regard the choice of landmarks as a clustering or vector quantisation problem, and to solve the problem using an iterative optimisation algorithm such as k-means clustering [Bis95]. Our experience

random: maxmin:



suggests that this is the wrong point of view, for several reasons. When the data set is large, this kind of optimisation is expensive. It is not clear that it is helpful to choose landmark points based on the pseudo-clustering that occurs in randomly chosen point samples, and in fact it may accentuate accidental features due to variation in sample density. On the other hand, the cheap-and-greedy maxmin algorithm does produce landmark points which cover the space and are locally well-separated. These seem to us to be the most pertinent qualities of a well-chosen landmark set.

2.4. Nested families

Suppose D is an $n \times N$ matrix of distances, as before. For each non-negative integer ν we construct a nested family of simplicial complexes $\mathrm{W}(D;R,\nu)$, where $R \in [0,\infty]$. The special cases $\nu = 0,1,2$ are of particular importance. The vertex set of $\mathrm{W}(D;R,\nu)$ is $\{1,2,\ldots,n\}$. Here is the definition.

- if $\nu = 0$, then for $i = 1,2,\ldots,N$ define $m_i = 0$.
- if $\nu > 0$, then for $i = 1,2,\ldots,N$ define m_i to be the ν-th smallest entry of the i-th column of D.
- the edge $\sigma = [ab]$ belongs to $\mathrm{W}(D;R,\nu)$ iff there exists a witness $i \in \{1,2,\ldots,N\}$ such that:

$$\max(D(a,i),D(b,i)) \leq R + m_i$$

- the p-simplex $\sigma = [a_0 a_1 \ldots a_p]$ belongs to $\mathrm{W}(D;R,\nu)$ iff all its edges belong to $\mathrm{W}(D;R,\nu)$; equivalently iff there exists a witness $1 \leq i \leq N$ such that:

$$\max(D(a_0,i),D(a_1,i),\ldots,D(a_p,i)) \leq R + m_i$$

To relate this to the previous construction, note the identity $\mathrm{W}(D;0,2) = \mathrm{W}(D)$.

Persistent homology groups over an interval $R \in [0,r]$ can now be computed using the algorithm from [ELZ00]. The preprocessing task is to generate a list of simplices (up to dimension $p+1$ for p-dimensional homology). For each simplex σ, one needs to identify its faces and determine its *time of appearance*, which is the smallest value $R = R_\sigma$ for which $\sigma \in \mathrm{W}(D,R)$. By definition, $R_\sigma = \max\{R_\tau : \tau \text{ is an edge of } \sigma\}$ and so we break up the task as follows:

1. Compute the $n \times n$ matrix E with off-diagonal entries $E(i,j) = R_{[ij]}$, which records the time of appearance of each edge.
2. Generate a list of simplices which appear by time r.
3. Compute the appearance time of each simplex as the maximum of the appearance times of its edges.

Step 1 can be expressed algebraically as a kind of 'min-max' matrix product: $E = D \odot D^*$. Here $\odot$ represents the operation

$$[A \odot B](i,j) = \min_k \max(A(i,k),B(k,j))$$

which is easily implemented.

For Step 2, a list of edges which appear by time r can be used to generate higher-dimensional cells inductively: for example the simplex $[a_0 \ldots a_p]$ occurs by time r iff the three lower-dimensional simplices $[a_1 \ldots a_p]$, $[a_0 \ldots a_{p-1}]$ and $[a_0 a_p]$ all occur by time r. Step 3 can be carried out concurrently with Step 2.

2.5. Comments on $\nu = 0,1,2$

We make some brief observations on the three different classes of persistent witness complex, $\nu = 0,1,2$.

$\nu = 0$: The family of complexes $\mathrm{W}(D;R,0)$ is closely related to the family $\mathrm{Rips}(L;R)$. Specifically, there are inclusions:

$$\mathrm{W}(D;R,0) \subseteq \mathrm{Rips}(L;2R) \subseteq \mathrm{W}(D;2R,0)$$

Relations between the persistent homology groups of the two families can be deduced. In practice, we find that the interval graphs for $\mathrm{W}(D;R,0)$ and $\mathrm{Rips}(L;R)$ look similar to one another.

$\nu = 1$: In some ways this is the best motivated of the three families, since it can be interpreted as arising from a family of coverings of the space X by Voronoi-like regions surrounding each landmark point, which overlap increasingly as $R \to \infty$.

$\nu = 2$: Although the persistent family is not as well motivated as in the previous case, we do have the following identity at $R = 0$:

$$\mathrm{W}(D;0,2) = \mathrm{W}(D)$$

In practice $\nu = 2$ families seem to give very clean persistent interval graphs, with surprisingly little "noise". The explanation suggested by this identity is that the complex is essentially already correct at $R = 0$; or at any rate only a small increase in R is necessary.

3. Examples

3.1. The sphere $S^2 \subset \mathbb{R}^3$

We applied several simplicial complex approximations to the task of recovering the correct Betti numbers of the sphere $S^2 \subset \mathbb{R}^3$. In each trial, 500 points were generated uniformly randomly on the unit sphere, by sampling points from a spherically symmetric Gaussian distribution and projecting

radially onto the unit sphere. From these, 12 landmark points were chosen randomly and by maxmin. Seven constructions were applied to each of these data sets: the Rips complex (on the landmark points alone), and witness complexes for the Euclidean and graph metrics, for each of $\nu = 0, 1, 2$. The calculation was organised so as to determine the Betti numbers b^0, b^1 and b^2, for all possible values of the feature-size parameter R. No persistent homology groups were computed.

How often was the correct profile $(b^0, b^1, b^2) = (1, 0, 1)$ obtained? A selection of statistics is presented in Figure 1. We ran 100 trials for each method of generating the landmarks. For each trial and each construction, we determined four constants R_0, R_1, K_0 and K_1. These are defined as follows. R_0 and R_1 are chosen so that $(b^0, b^1, b^2) = (1, 0, 1)$ for $R \in [R_0, R_1)$, agreeing with the 2-sphere; but $(b^0, b^1, b^2) \neq (1, 0, 1)$ for $R \geq R_1$ and for $R = R_0 - \varepsilon$. In other words, $[R_0, R_1)$ is the rightmost contiguous interval over which the homology of S^2 is correctly recovered. At $R = K_0$ the Betti profile changes permanently to $(1, 0, 0)$, indicating that the data have coalesced into a single contractible blob. Finally, $R = K_1$ marks the time when the complex becomes the complete simplex on 12 vertices; all possible cells have been included.

In the tables, "% success" indicates the number of trials (out of 100) where the homology of S^2 is correctly recovered for some interval of values of R, no matter how small. For each successful trial, *relative dominance* and *absolute dominance* are defined to be $(R_1 - R_0)/K_0$ and $(R_1 - R_0)/K_1$ respectively. Relative dominance compares the lengths of the successful interval $[R_0, R_1)$ and the interval $[0, K_0]$ of homological activity. Absolute dominance compares the successful interval with the interval $[0, K_1]$ of *cellular* activity. If either of these quantities is large (that is, close to 1), this indicates that the Betti profile $(1, 0, 1)$ can be taken seriously *a priori*, and not just because we know the correct answer.

The last three rows of the tables give median values of these statistics, and of the total number of cells (up to dimension 3) at $R = R_0$. The median is taken over successful trials only. For unsuccessful trials R_0 and R_1 are not defined; although both dominances may be taken to be 0 in those cases.

We make several observations.

1. The Rips construction can be grouped with the $\nu = 0$ witness complexes; their behaviour is very similar.
2. Choosing landmark points by maxmin is enormously better than random choice for $\nu = 0$. For $\nu = 1, 2$ there is still an improvement, but it it much less significant. In a sense, these latter constructions have a built-in robustness to irregular sampling.
3. High dominances indicate stable results. The $\nu = 1, 2$ algorithms considerably outperform the $\nu = 0$ algorithms.
4. A mystery: the $\nu = 2$ cases have extremely high relative dominances, but much lower absolute dominances. Which should be taken more seriously? The underlying cause of the difference is that K_0/K_1 is small; in other words homological activity dies down long before cellular activity, as R increases. A more sophisticated understanding is called for.

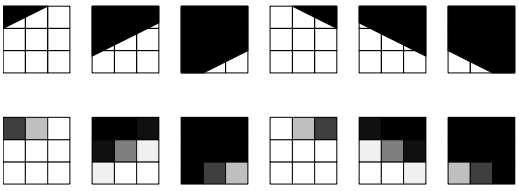
The overall message is reasonably clear: the $\nu = 1, 2$ witness complexes give topological approximations which are more reliable, use fewer cells, and are statistically more defensible than Rips complexes.

3.2. Natural image statistics

In this section we demonstrate one approach to studying the topological features of a noisy point-cloud data set. We apply our techniques to an example derived from natural image data, provided by David Mumford, which has a known topological feature that we seek to identify. Our methods detect this feature, as well as some secondary features.

The data set in question is described by Lee, Pederson and Mumford in [LPM03]. They extracted 4.2×10^6 high-contrast 3×3 optical image patches from van Hateren's still image collection [vHvdS98]. Each patch is normalised twice: first by subtracting the mean intensity, then by rescaling to unit length in a suitable metric. After these normalisations, the data can be represented by points on the unit sphere in $\mathbb{R}^8$. For our analysis, we randomly selected a much smaller subset of 5×10^4 points and regarded that as our primary source.

A edge feature in a natural image can be idealised as a perfectly straight boundary between two homogeneous regions of different brightness levels. Within a single patch, the family of edge features can be parametrised by angle and distance from the center. Here are some examples, before (top row) and after (bottom row) pixelisation:



The parametrisation by angle and distance implies that the family of idealised edges has the topology of an annulus. This annulus is naturally embedded in the unit sphere of $\mathbb{R}^8$, where the normalised patches live. Since edges are common feature of high-contrast regions in natural images, one expects to find a strong concentration of data points on or near this annulus. See [LPM03] for a detailed discussion.

Can we detect this annulus using topological methods only? A direct application of simplicial complex approximation to the 5×10^4 data points is destined to fail, since there are points distributed all over the sphere and not just in the high-density regions. To extract a high-density sample, we threshold on a simple density function

$$\rho_K(x) = |x - x_K|$$

12 LANDMARK POINTS CHOSEN RANDOMLY							
	Rips	Witness: Euclidean metric			Witness: graph metric		
		$\nu = 0$	$\nu = 1$	$\nu = 2$	$\nu = 0$	$\nu = 1$	$\nu = 2$
% success	54	51	99	99	53	100	97
in cases where a successful reconstruction exists for some R							
median relative dominance	0.038	0.059	0.620	0.808	0.062	0.600	0.798
median absolute dominance	0.034	0.047	0.347	0.163	0.046	0.318	0.152
median number of cells	208	199	86	94	208	92	92

12 LANDMARK POINTS CHOSEN BY SEQUENTIAL MAXMIN							
	Rips	Witness: Euclidean metric			Witness: graph metric		
		$\nu = 0$	$\nu = 1$	$\nu = 2$	$\nu = 0$	$\nu = 1$	$\nu = 2$
% success	100	100	100	100	100	100	100
in cases where a successful reconstruction exists for some R							
median relative dominance	0.184	0.215	0.752	0.924	0.216	0.744	0.922
median absolute dominance	0.161	0.162	0.519	0.252	0.153	0.466	0.209
median number of cells	74	78	66	79	82	66	80

Figure 1: *Recovering the homology profile of the sphere $S^2 \subset \mathbb{R}^3$ using 14 different constructions*

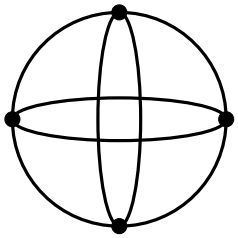
where x_K is the K-th nearest neighbour of x, for some K.

The choice of K affects the results qualitatively, as can be seen in Figure 2, which shows different cuts of the data projected onto the first two coordinates of $\mathbb{R}^8$. The nine panels show the 10%, 20% and 30% of points having the smallest values of ρ_{15}, ρ_{100} and ρ_{300}. The 30% cut with $K = 300$ appears to be concentrated entirely on an annulus. With $K = 15$, on the other hand, there is a cross-like feature that is already present once the cut is large enough for the annulus to be fully formed.

Remark 5 This kind of behaviour can be explained by the following simple model. Suppose that the data are concentrated along various strata of different dimensions, with a uniform density within each stratum. For a point x in a stratum of dimension d, we have the approximate behaviour $\rho_K(x) \propto K^{1/d}$. It follows from this formula that small values of K emphasise lower-dimensional strata, whereas large values of K emphasise higher-dimensional strata.

Figure 3 shows the persistence interval graphs for Betti 1, computed for witness complexes having 50 vertices chosen by maxmin, using the Euclidean metric. The number of long intervals in each of the bottom six graphs matches what we see in the 2-dimensional plots. In the cases with cut = 10%, there is an interval near the end of the range for R. This arises when the four clusters link up, briefly, to form a ring. The ring gets filled in very quickly as R increases, so the interval is short.

Each of the persistence graphs for $K = 15$ has five intervals of noticeable length. At the 30% cut these intervals are long enough to be regarded, without any doubt, as stable features. This is not evident from the two-dimensional projection, which appears to show four holes. Guided by the evidence that $b^1 = 5$, we are led to the following 'three circles' model shown in the figure. The data are clustered along three



circles in $\mathbb{R}^8$, namely the unit circles in the e_1–e_2, e_1–e_3 and e_2–e_4 planes. There are four points of intersection; the second and third circles do not intersect at all. Once we have it, it is comparatively easy to verify this interpretation by a closer inspection of the data.

To understand the significance of these clusters, we con-

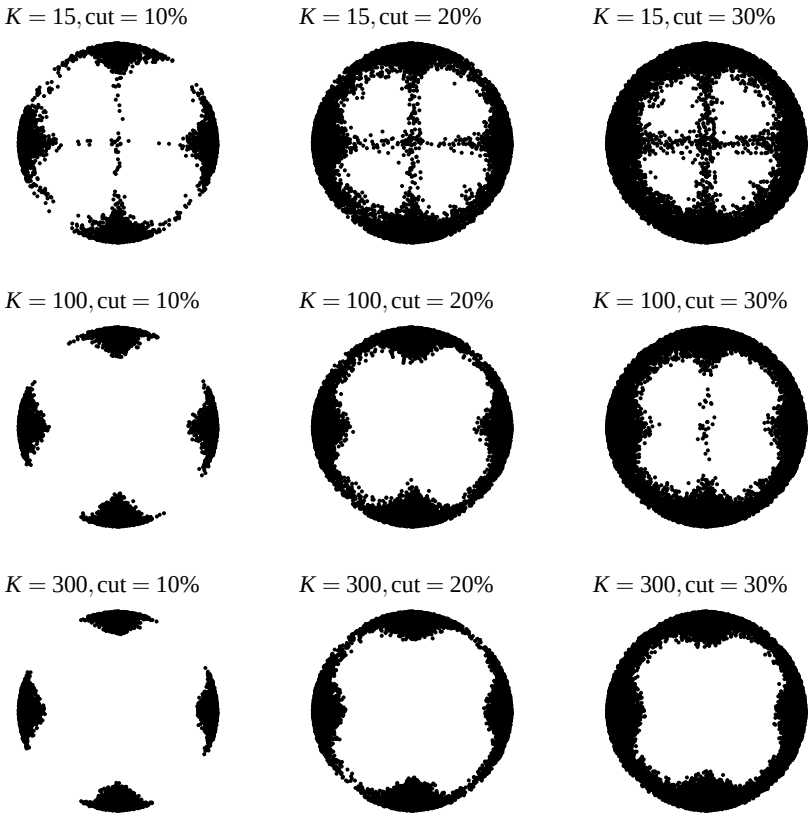
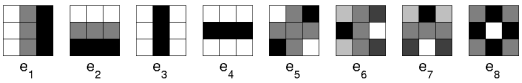


Figure 2: *First two coordinates of the Mumford data: different cuts*

sult the basis vectors themselves. The e_1–e_2 circle corresponds to linear gradients, parametrised by angle. Each linear gradient can be regarded as the mean of a family of edge features sharing the same angle, and located nearby in patch space. The two fainter circles indicate the prevalence of vertically symmetric and horizontally symmetric patches, respectively. It is unclear whether this is more an artifact of the choice of localisation (square patches with vertical and horizontal sides) than a symptom of the bias for vertical and horizontal features observed in natural image statistics [CY03]. It is likely that both factors play a part.



Remark 6 In constructing the witness complexes for these examples, we have taken $\nu = 1$. In practice, the three choices $\nu = 0, 1, 2$ lead to persistence interval graphs of quite different visual character, even if they convey the same message. Figure 4 illustrates the three different cases for a fixed cut, consisting of the 25% of points having the lowest values of ρ_{125}. The case $\nu = 0$ typically presents the most ambiguity; the 'true' long interval does not appear immediately, and there are several short 'noise' intervals which appear at different values of R. The case $\nu = 1$ is clear-cut; there is a much larger number of noise intervals at $R = 0$, but these disappear en masse very quickly. When $\nu = 2$ the picture is shockingly clean; this supports our claim that the weak witness complex $\mathrm{W}(D; 0, 2) = \mathrm{W}(D)$ is immediately a good approximation to the underlying space. We find that these traits are quite consistent across different data sets, and certainly they invite further investigation.

4. Concluding remarks

Modern statistical analysis increasingly calls for the use of nonlinear techniques, capable of resolving the underlying structure of a data set. The modern theory of nonlinear dimensionality reduction (NLDR) gives several examples of such techniques [TdSL00, RS00]. These tend to be restricted to data manifolds whose topology is comparatively simple. On the other hand, it is clear that many naturally occuring data sets exhibit non-trivial topology. We believe that the estimation of topological invariants is a necessary part of the analysis of such data sets. In other areas of research, there is a growing body of algorithms which exploit topological

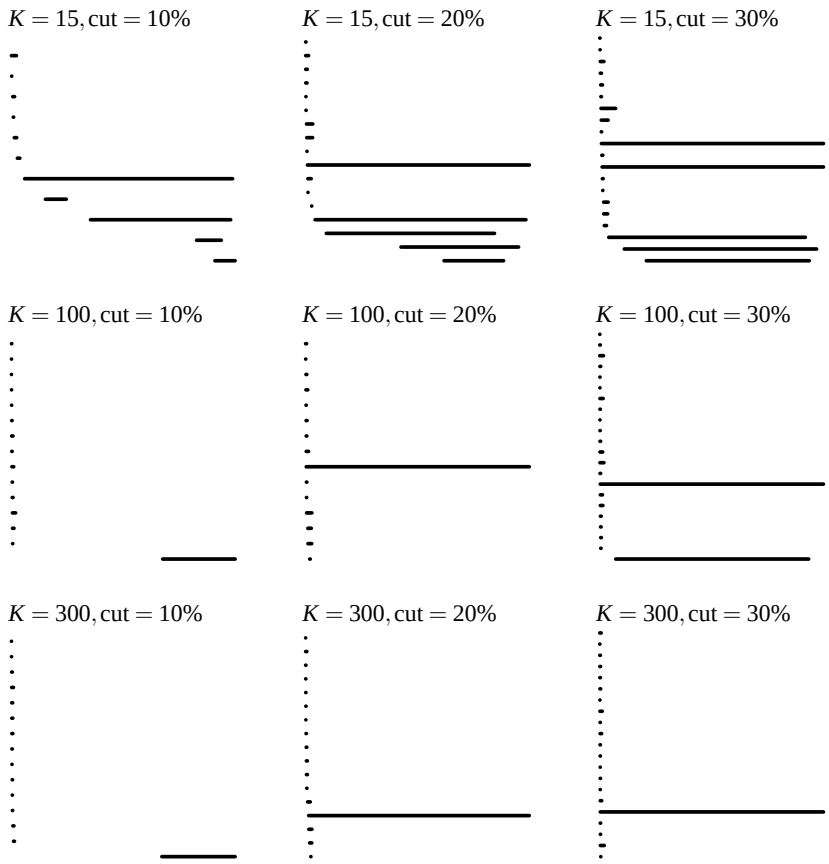


Figure 3: *Betti 1 persistence intervals of witness complexes for the Mumford data: varying the cut, and keeping fixed 50 vertices, Euclidean metric,* $\nu = 1$.

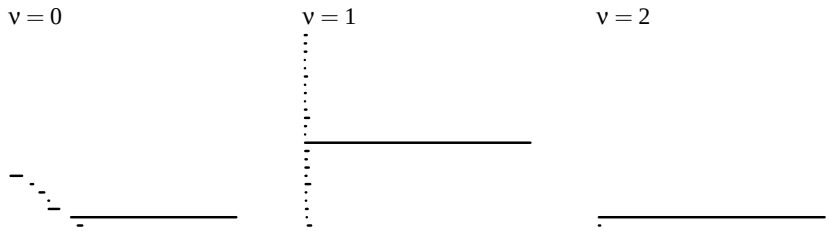


Figure 4: *Betti 1 persistence intervals of witness complexes for the Mumford data: varying* ν*, and keeping fixed* $K = 125$*, cut = 25%, 50 vertices, intrinsic graph metric (with* $k = 125$*).*

information carried in point-cloud data. Rapid topological profiling is an essential tool in these developments.

In this paper we present a robust, efficient tool for carrying out these tasks. Witness complexes have several advantages over existing methods; they are easily computed, they are adaptable to arbitrary metrics, they use only a small number of cells, and they do not suffer from curse of dimensionality. As shown by the examples, the combination of witness complexes with persistent homology is highly effective in practice, even on noisy data. There is a long way to go before we have truly flexible and robust tools for topological data analysis. We hope that this paper represents a useful step in that direction.

Acknowledgements

We gratefully acknowledge the support of the NSF, through grant DMS-0101364.This work was carried out at the Department of Mathematics, Stanford University.

We wish to thank several individuals: Afra Zomorodian, for numerous discussions and for use of his code for persistent homology calculations; David Mumford, for making available his database of 3×3 patches; Debashis Paul, for his considerable assistance with preliminary investigations of the Mumford data; and Leo Guibas, for helpful comments and encouragement. VdS also thanks Josh Tenenbaum, for generously sharing his insights into nonlinear statistics and for emphasising the importance of landmark-based techniques; and Carrie Grimes, for advice on the experimental studies in this paper.

References

[AB99] AMENTA N., BERN M.: Surface reconstruction by Voronoi filtering. *Discrete and Computational Geometry 22* (1999), 481–504. 1

[ACDL02] AMENTA N., CHOI S., DEY T. K., LEEKHA N.: A simple algorithm for homeomorphic surface reconstruction. *International Journal of Computational Geometry and Applications 12*, 1-2 (2002), 125–141. 1

[Bis95] BISHOP C. M.: *Neural Networks for Pattern Recognition*. Oxford University Press, 1995. 4

[CCdS03] CARLSSON E., CARLSSON G., DE SILVA V.: An algebraic topological method for feature identification. [submitted], 2003. 2

[CY03] COUGHLAN J. M., YUILLE A. L.: Manhattan world: orientation and outlier detection by bayesian inference. *Neural Computation 15* (2003), 1063–1088. 8

[dS03] DE SILVA V.: A weak definition of Delaunay triangulation. [submitted], 2003. 2, 4

[Ede95] EDELSBRUNNER H.: The union of balls and its dual shape. *Discrete & Computational Geometry 13*, 3-4 (1995), 415–440. 3

[ELZ00] EDELSBRUNNER H., LETSCHER D., ZOMORODIAN A.: Topological persistence and simplification. In *IEEE Symposium on Foundations of Computer Science* (2000), pp. 454–463. 1, 2, 3, 5

[LPM03] LEE A. B., PEDERSEN K. S., MUMFORD D.: The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision 54*, 1-3 (2003), 83–103. 1, 2, 6

[Mas91] MASSEY W.: *A Basic Course in Algebraic Topology*. Springer-Verlag, New York, 1991. 2, 3

[MS94] MARTINETZ T., SCHULTEN K.: Topology representing networks. *Neural Networks 7*, 3 (1994), 507–522. 1, 4

[RS00] ROWEIS S., SAUL L.: Nonlinear dimensionality reduction by locally linear embedding. *Science 290* (December 2000), 2323–2326. 8

[Spa66] SPANIER E. H.: *Algebraic Topology*. McGraw-Hill Book Co., 1966. 3

[TdSL00] TENENBAUM J. B., DE SILVA V., LANGFORD J. C.: A global geometric framework for nonlinear dimensionality reduction. *Science 290* (December 2000), 2319–2323. 2, 4, 8

[vHvdS98] VAN HATEREN J. H., VAN DER SCHAAF A.: Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B 265* (1998), 359–366. 6

[ZC04] ZOMORODIAN A., CARLSSON G.: Computing persistent homology. In *20th ACM Symposium on Computational Geometry* (2004). 1, 2, 3