Improving Homology Estimates with Random Walks

Paul Bendich[1], **Taras Galkovskyi**[1], **and John Harer**[2]

[1] Department of Mathematics, Duke University, Durham, NC, 27708, USA
2 Departments of Mathematics and Computer Science, Center for Systems Biology, Duke University, Durham, NC, 27708, USA

E-mail:
bendich@math.duke.edu,galkovsky@gmail.com,harer@math.duke.edu

Abstract. This experimental paper makes the case for a new approach to the use of persistent homology in the study of shape and feature in datasets. By introducing ideas from diffusion geometry and random walks, we discover that homological features can be enhanced and more effectively extracted from spaces that are sampled densely and evenly, and with a small amount of noise. This study paves the way for a more theoretical analysis of how random walk metrics affect persistence diagrams, and provides evidence that combining topological data analysis with techniques inspired by diffusion geometry holds great promise for new analyses of a wide variety of datasets.

1. Introduction

The continually growing importance of analyzing massive datasets has prompted the development of a number of new techniques for finding shape and feature in point clouds. Well established methods from statistics are now augmented with approaches that focus on shape identification, inference, and dimension reduction using methods of geometry and topology, subjects that were largely regarded in the past as too abstract to be applied to data analysis. Computational Topology has emerged as a new discipline in the last ten years [12], and its most significant application to date has been to Topological Data Analysis (TDA) in which datasets are treated as point clouds, and persistent homology is calculated to look for both global and local features in these clouds. In parallel, diffusion geometry [9], [10] has developed into a powerful tool in dimension reduction and the analysis of high-dimensional data. While its methods are much more analytical than those of computational topology, it nonetheless shares the same goals and the same spirit of applying well established methods of geometry to applied problems.

Now that these subjects have come of age, the time has come to consider how they can be made to work together and to work with methods from statistics. This experimental article considers the question of how the methods of diffusion geometry can be combined with TDA. A related theoretical article [16] looks at the question of how to do statistics on the persistence diagrams that describe the results of applying TDA methods to repeated samples

from a geometric object. Along with some other recent papers, e.g. [6], these works suggest possible research directions in support of a future merger of techniques.

Main idea. The main idea that we explore here is that by using random walks on a dataset one can define metrics that enhance topological properties and increase the capability of persistent homology to capture them. A well-known example in diffusion is a curve in the plane which is topologically a circle, but lies close to a figure-8 (see Figure 2 below). When this curve is sampled densely enough, diffusion geometry recognizes the circle by finding a new embedding based on the eigenfunctions of an appropriate operator derived from the data.

Our approach is motivated by this idea. We use the diffusion metric D_m to define a distance between pairs of points in our dataset, and we then construct the Rips complexes using this distance and compute the corresponding persistence diagram. For reasons described below, we also consider our own ad-hoc definition of a different symmetric function ρ_m, defined below, on our point set and investigate the properties of its associated Rips complexes and persistence diagrams.

Results. Here we briefly summarize the experimental results, which are described more fully in Section 5. We ran two main types of experiments on points cloud sampled from the figure in Figure 2. When we sampled points densely without outliers, we found that the diffusion metric D_m caused the main cycle to separate very cleanly from the shorter cycle introduced by the bottle-neck near the origin. Here the advantage held by D_m over the traditional Euclidean distance was striking, while ρ_m did better than Euclidean distance but certainly not as well as diffusion.

We also considered what happens when varying amounts of noise, including a decent number of outliers, were introduced into our dataset. When the density of noisy samples was small relative to the sampling density on the underlying space, we found that our persistence diagrams for D_m were mostly unchanged, something which is impossible with the Euclidean metric. On the other hand, the performance of the diffusion metric rapidly broke down as the noise-to-signal ratio increased. For these very noisy samples, the ρ_m-diagrams continued to retain at least some of the topological structure of the underlying space.

To obtain persuasive results, we needed to take a large number of sample points, large enough so that the persistence diagram computation would have become prohibitively slow. To fix this problem, we employed a sub-sampling method broadly inspired by Witness Complex [11] methodology. This method seems interesting in its own right, and so we describe it fully in Section 4.

Prior work. A number of authors have begun to study methods for applying persistent homology to noisy datasets. In one of the first attempts [4], the space of natural images is denoised before computing a topological filtration that discovers a Klein bottle and leads to a new compression methodology for images. A more recent effort to preprocess data before computing persistence is in [15]. In a different direction, [6] defines the concept of distance from a distribution, which provides one of the first true inter-linkings of probabilistic

methods with topological ones and allows a new theoretical analysis of how well a noisy dataset represents the space from which it was sampled.

Outline. Necessary background on persistent homology is given in Section 2, while Section 3 introduces the method of using Euclidean distance to assess the homology of a point cloud. This method has obvious limitations, to which we propose our solution in Section 4. Experimental results, as well as some implementation details, are presented in Section 5. Finally, Section 6 concludes the paper with a discussion and ideas for future research directions.

2. Background

Persistent homology is described in great detail in the recent textbook [12]. In this section, we confine ourselves to a few brief definitions and examples. All homology groups are assumed to be taken over some field, usually $\mathbb{Z}/2\mathbb{Z}$, and are thus just vector spaces over that field. We assume familiarity with homology itself; see for example [18] for more background.

Sublevel set filtrations. Suppose that we have a topological space $\mathbb{M}$ equipped with a real-valued function $f : \mathbb{M} \to \mathbb{R}$. For each $r \leq s$ and each homological dimension p, the inclusion $\mathbb{M}_r \hookrightarrow \mathbb{M}_s$ of sublevel sets induces linear maps $\mathsf{H}_p(\mathbb{M}_r) \to \mathsf{H}_p(\mathbb{M}_s)$ between homology groups. In a nutshell, persistent homology uses these linear maps to track the evolution of the homology of $\mathbb{M}_r$, as r increases from negative to positive infinity, and then compactly displays this information in persistence diagrams; for an example of the latter, see the right side of Figure 2.

We now give some more details as to what this means. As an aside, we note that there is nothing special about homology here, since persistence can be defined for any sequence of vector spaces connected by linear maps; see [5] for a detailed treatment.

Critical values. Although we ostensibly have infinitely many homology groups, one for each real number, it often turns out there are only a finite number of changes as we move from negative to positive infinity. More precisely, we say that r is a homological regular value (hrv) of f if there exists some $\epsilon > 0$ such that, for all positive $\delta < \epsilon$, the inclusions $\mathbb{M}_{r-\delta} \hookrightarrow \mathbb{M}_{r+\delta}$ induce homology isomorphisms in every dimension. If r does not satisfy this condition, then it is a homological critical value (hcv). For example, consider the round circle $\mathbb{X} \subset \mathbb{R}^2$ of radius r as drawn on the left side of Figure 1. This circle defines a distance function $d_{\mathbb{X}} : \mathbb{R}^2 \to \mathbb{R}$; in words, $d_{\mathbb{X}}(y)$ is the Euclidean distance between a point $y \in \mathbb{R}^2$ and its closest neighbor on $\mathbb{X}$. One can imagine the sublevel sets of $d_{\mathbb{X}}$ as being a sequence of thickenings of the circle $\mathbb{X}$ within its ambient space $\mathbb{R}^2$. In this case, $d_{\mathbb{X}}$ will have only two hcvs, 0 and r. On the other hand, the function $d_{\mathbb{Y}}$, where $\mathbb{Y}$ is the space on the left side of Figure 2, will have four hcv's: $0 < s < r < t$ (the right lobe is slightly larger than the left one). The smallest non-zero hcv of $d_{\mathbb{Y}}$ is called the *homological feature size* of $\mathbb{Y}$, e.g in Figure 2 it is s.

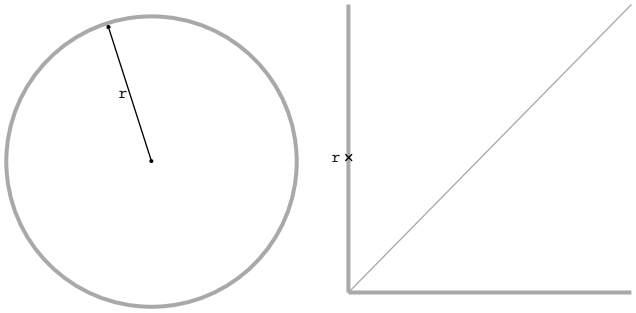


Figure 1: The persistence diagram $\mathrm{Dgm}_1(d_{\mathbb{X}})$ is on the right, where $\mathbb{X}$ is on the left.

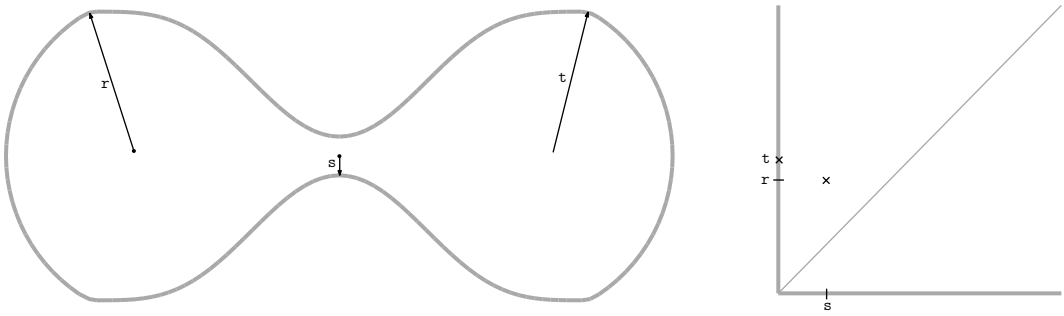


Figure 2: The persistence diagram $\mathrm{Dgm}_1(d_{\mathbb{Y}})$ is on the right, where $\mathbb{Y}$ is on the left.

Birth, death, diagrams. Returning to the general case, suppose we have a function $f : \mathbb{M} \to \mathbb{R}$ with a finite set of hcvs $a_1 < a_2 < \ldots < a_n$. A function which meets this assumption, along with the requirement that the homology groups of all sublevel sets be of finite rank, is called *tame*. We choose hrvs $\{b_i\}$ with $b_0 < a_1 < b_1 < \ldots < a_n < b_n$, and put $\mathbb{M}_i = \mathbb{M}_{r_i}$. Fixing a homological dimension p, we adopt the shorthand notation $\mathsf{H}_p^i = \mathsf{H}_p(\mathbb{M}_i)$, and we consider the sequence of homology groups connected by linear maps,

$$0 = \mathsf{H}_p^0 \to \mathsf{H}_p^1 \to \mathsf{H}_p^2 \to \ldots \to \mathsf{H}_p^n = \mathsf{H}_p(\mathbb{M}). \tag{1}$$

For each $i \leq j$, we let $\mathsf{f}_p^{i,j} : \mathsf{H}_p^i \to \mathsf{H}_p^j$ be the map induced on homology by the inclusion $\mathbb{M}_i \hookrightarrow \mathbb{M}_j$. A class $\alpha \in \mathsf{H}_p^i$ is then said to be *born* at $\mathbb{M}_i$ if $\alpha \notin \operatorname{im} \mathsf{f}_p^{i-1,i}$. Such an α *dies entering* $\mathbb{M}_j$ if $\mathsf{f}_p^{i,j}(\alpha) \in \operatorname{im} \mathsf{f}_p^{i-1,j}$, but $\mathsf{f}_p^{i,j-1}(\alpha) \notin \operatorname{im} \mathsf{f}_p^{i-1,j-1}$. The *persistence* of the class α is $b_j - b_i$; this real number gives a measure of how long α lives during the sublevel set filtration.

There is one technical issue to address, which we do via the example of the space $\mathbb{Y}$ in Figure 2. Letting $\mathbb{M}_1$ and $\mathbb{M}_2$ be sublevel sets just before and just after the hcv s, respectively, we notice that $\mathbb{M}_1$ has the homotopy type of a circle, while $\mathbb{M}_2$ has that of a wedge of two circles. There are thus two 1-dimensional classes α, β born at $\mathbb{M}_2$, the ones represented by the two new circles. On the other hand, α and β both represent the same coset of $\mathsf{H}_1^2/\operatorname{im} \mathsf{f}_1^{1,2}$, since $\alpha + \beta$ is homologous to the generator of H_1^1. Both of these new classes die upon entering $\mathbb{M}_3$, a sublevel set just after the hcv r: one of them becomes homologous to zero, while the other one becomes homologous to the generator of H_1^1. This example illustrates the general fact that whenever a class α is born at some $\mathbb{M}_i$, an entire coset is born with it, but all classes

in this coset will die whenever α dies.

For each coset of p-dimensional classes born at $\mathbb{M}_i$ and dying entering $\mathbb{M}_j$, we draw the point (a_i, a_j) in the plane. The resulting multi-set, $\mathrm{Dgm}_p(f)$, is the p-th persistence diagram for the function f. Notice that the persistence of a class α is then just the vertical distance of its representative point from the major diagonal. For technical reasons that will soon become clear, we also add all major diagonal points (a, a) where $a \geq 0$, taken with infinite multiplicity, to every persistence diagram. As an example, the right side of Figure 2 displays $\mathrm{Dgm}_1(d_{\mathbb{Y}})$, where $\mathbb{Y}$ is the space drawn on the left side of the same figure. Notice that the "real" 1-dimensional homology of $\mathbb{Y}$ shows up as the point on the vertical axis, while the other non-zero persistence point in the diagram represents the new homology created by a slight thickening of $\mathbb{Y}$. In terms of the persistence diagram, the homological feature size of $\mathbb{Y}$ shows up as the minimum of the smallest death value of a point on the y-axis, and the smallest birth value of a point off the y-axis.

Diagram stability. It turns out that the persistence diagrams $\mathrm{Dgm}_p(f)$ and $\mathrm{Dgm}_p(g)$, where f and g are two "close" functions defined on the same domain, will also be "close," as measured by an appropriate metric. Indeed, given the inevitability of noisy or incomplete input, a statement of this kind really has to be true if persistence diagrams are to be a useful tool for any sort of data or shape analysis.

The bottleneck distance d_B between any two persistence diagrams D, D' is defined as follows:

$$d_B(D, D') = \inf_{\Gamma: D \to D'} \sup_{u \in D} ||u - \Gamma(u)||_\infty, \tag{2}$$

where Γ ranges over all bijections between the diagrams D and D'. Recall that every persistence diagram contains infinitely many copies of every major diagonal point, and thus bijections always exist. In words, the bottleneck distance is defined by considering all matchings between D and D', assigning the largest l_∞ distance between matched points, and then taking the minimum over all matchings. Note that the matching may not be unique.

Under this metric, the persistence diagram $\mathrm{Dgm}_p(f)$ is stable to small perturbations of f. Precisely, the following result was proven first in [7] and then given a wider context in [5]:

Theorem 1 (Diagram Stability Theorem)**.** *If $\mathbb{X}$ is a compact topological space, then for every homological dimension p and every pair of tame functions $f, g : \mathbb{X} \to \mathbb{R}$,*

$$d_B(\mathrm{Dgm}_p(f), \mathrm{Dgm}_p(g)) \leq ||f - g||_\infty.$$

Filtered simplicial complexes. In our experiments below, we will not strictly have a real-valued function on a topological space. Instead, we will consider some finite simplicial complex K, along with a monotonic function $f : K \to \mathbb{R}$ that is constant on each (open) simplex; here monotonic means that if τ is a face of σ, then $f(\tau) \leq f(\sigma)$. If we consider values $b_1 < b_2 < \ldots < b_n$ for f, we let K_i be the subcomplex of K consisting of all simplices

whose f-values are no larger than b_i. We then have inclusions $K_i \hookrightarrow K_j$, for $i \leq j$, and hence an application of the homology functor allows us to define persistence and draw persistence diagrams, exactly as above.

One example that we will make use of below is the following. Suppose that we are given a finite point set $\mathbb{U}$ and a symmetric function $\rho : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ on the edges of the complete graph G with vertex set $\mathbb{U}$, such that $\rho(u, u) = 0$ for every point u; for example, ρ could be a metric on $\mathbb{U}$, but it need not be. Given any subgraph $H \subseteq G$, we define $R(H)$ to be the largest simplicial complex which has H as its 1-skeleton; note that $R(H)$ is often called a clique complex in the literature and is akin to a Rips Complex. Since $\mathbb{U}$ is a finite set, ρ takes finitely many values. For each such value ρ_i, we let $G_i \subseteq G$ be the subgraph with vertex set $\mathbb{U}$ and edge set consisting of all edges with ρ-values less than or equal to ρ_i. Notice that $\mathbb{U} \subset G_0$. We can then filter the simplicial complex $R(G)$ with subcomplexes $R(G_i)$, leading to persistence diagrams in each homological dimension, which we denote $\mathrm{Dgm}_p(\mathbb{U}, \rho)$.

3. Homology of a Point Cloud

Consider the 1-dimensional homology of the point cloud $\mathbb{U}$ on the left side of Figure 3. Taken literally, this statement is of course absurd, since $\mathbb{U}$ is simply a discrete set of points and thus has homology only in dimension zero. On the other hand, a slight blurring of the point cloud may result in a clear one-cycle being formed, while a further blurring will pinch this cycle into two smaller ones. This can been seen in the persistence diagram $\mathrm{Dgm}_1(d_U)$ (right side of Figure 3), where the different blurrings manifest as sublevel sets of the distance function d_U. In this diagram, we see two highly persistent points, and then quite a few other points very close to the diagonal. Perhaps one could identify the homology classes associated to the high persistence points as representing, at least in some sense, the 1-dimensional homology of the point cloud; in general, it makes sense to think of classes which are born early and persist for a long time as being in some sense "real." More humbly, one could just report the entire persistence diagram as an answer to the point cloud homology question, and let the end-user make a thresholding decision as to which classes are the most important. In any case, any persistence-diagram approach will lead to difficulty in two different ways, as we now describe.

Homological feature size. It would seem reasonable to demand that our guess at the homology of $\mathbb{U}$ should in some sense match with the homology of any topological space $\mathbb{Y}$ from which $\mathbb{U}$ was sampled, at least for sufficiently good sampling. But this can never work in general. For example, our point cloud $\mathbb{U}$ was in fact sampled, with very little noise and absolutely no outliers, from the space $\mathbb{Y}$ on the left side of Figure 2. But $\mathsf{H}_1(\mathbb{Y})$ is rank one, since $\mathbb{Y}$ is topologically a circle.

To be more precise, $\mathbb{U}$ is an ϵ-sample of $\mathbb{Y}$, where $\epsilon = ||d_Y - d_U||_\infty$ is the Hausdorff distance between the space and the point cloud. Appealing to the above theorem, this means the two persistence diagrams $\mathrm{Dgm}_1(d_Y)$ and $\mathrm{Dgm}_1(d_U)$ can be no more than ϵ apart in the bottleneck metric. Recall that the actual first homology of $\mathbb{Y}$ is represented by the

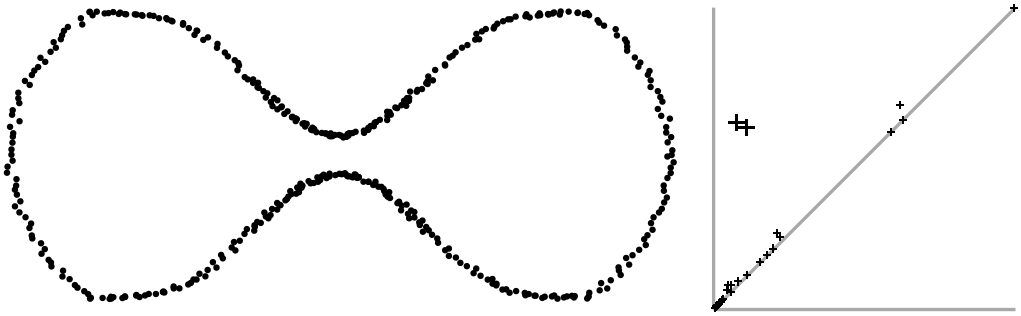


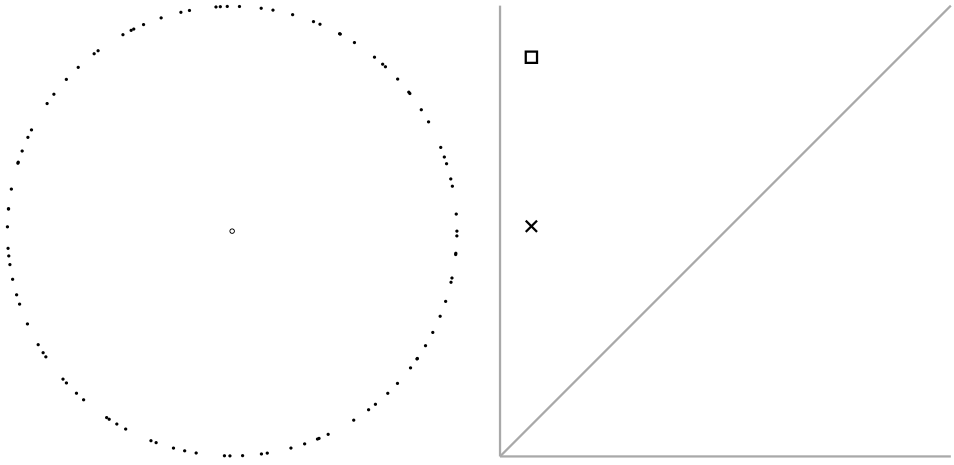
Figure 3: On the left, a sampled version $\mathbb{U}$ of the space $\mathbb{Y}$ shown in figure 2. The persistence diagram $\mathrm{Dgm}_1(d_U)$ is on the right; points of higher persistence are drawn larger for clarity.

point directly on the y-axis of $\mathrm{Dgm}_1(d_Y)$, while the other point represents the new one-cycle formed by a slight thickening of $\mathbb{Y}$. Although the optimal bijection between the two diagrams certainly matches the two large-persistence points in $\mathrm{Dgm}_1(d_U)$ with these two points, sending all other points to the major diagonal, we cannot hope to tell which $\mathrm{Dgm}_1(d_U)$-point matched to the y-axis point, or indeed how many y-axis points there really were. In a nutshell, the problem is that the homological feature size of $\mathbb{Y}$ is very small, and so nothing but the very densest of samples has any hope of picking out the one true cycle without also suggesting the extra cycle.

Outliers. There is a second, and in some sense far graver, problem: persistence diagrams are stable to small changes in input set, but not to a small amount of very large noise. Namely, suppose that the vast majority of points in a cloud $\mathbb{U}$ seem to trace out a clear space $\mathbb{Y}$, but that there are also a few very bad sampling errors here and there. More formally, suppose that there is a tiny subset $\mathbb{U}_0$ of $\mathbb{U}$ such that $\mathbb{U} - \mathbb{U}_0$ is an ϵ-sample of $\mathbb{Y}$, for some very small ϵ, but that the points in $\mathbb{U}_0$ lie quite far away from $\mathbb{Y}$. In this case, the Hausdorff distance between $\mathbb{U}$ and $\mathbb{Y}$ will be quite high, and thus the Diagram Stability Theorem gives us no useful guarantees about the relationship between the respective diagrams.

For example, suppose $\mathbb{U}$ is the point cloud shown on the left side of Figure 4, which seems to be a very faithful sample of a perfectly round circle $\mathbb{X}$. The persistence diagram $\mathrm{Dgm}_1(d_U)$ for this sample consists of the box on the right of the same figure. Compare this to the right side of Figure 1, which shows the persistence diagram $\mathrm{Dgm}_1(d_X)$; we see that the two diagrams are quite close in the bottleneck metric.

On the other hand, suppose that we add one solitary outlier directly in the center, producing the new point cloud $\mathbb{V}$, which is also shown on the left of Figure 4. Looking at the cross on the right side of the same figure, we find that $\mathrm{Dgm}_1(d_V)$ is markedly different from $\mathrm{Dgm}_1(d_U)$, as the persistence of the main cycle has diminished by a factor of $\sqrt{3}$. The reason is obvious: suppose the circle has radius r_0; for $\mathbb{U}$, this main cycle dies when we reach the side length of an equilateral triangle inscribed in the circle, which is $\sqrt{3}r_0$, while it dies at r_0 for the sample with just one outlier.



(a) Samples (b) Diagrams

Figure 4: Left: dataset $\mathbb{U}$ of size 100 sampled from a circle. The noisy dataset $\mathbb{V}$, consisting of $\mathbb{U}$ plus the central point (hollow point) is also depicted. Right: persistence diagrams. The symbols $\square$ and $\times$ mark persistent homology classes corresponding to 'clean' and 'noisy' datasets, respectively.

Summary. Looking at the above paragraphs in a slightly different light, we see that the idea of doing homology inference by thickening a point sample $\mathbb{U}$ into gradually growing Euclidean balls and then computing the resulting persistence diagram has two major flaws. First, the space itself might have very small homological feature size. We can get around this by taking an extremely dense point sample, but this will come at a high computational cost. More seriously, any realistic sample will of course have a few bad outliers, and increasing the density will never hope to eliminate them.

4. Random Walk

Here we define the two symmetric functions, each based on random walk, that we use in our experiments. We also describe the witness-like method we employ to achieve a massive increase in computation speed.

Diffusion metric. There are various definitions of the diffusion metric to be found in the literature; we use the definition given in [8]. First, we imagine that we are standing at a particular point $u \in \mathbb{U}$ and are going to choose a random step to another point $v \in \mathbb{U}$, where the probability of choosing v is proportional to some kernel function $k(u, v)$ which is a decreasing function of the Euclidean norm $||u - v||$; most of the literature uses a kernel function of the form $k(x, y) = exp(-\frac{||x-y||^2}{\alpha})$, where α is some positive tuning parameter.

That is, we choose v with probability

$$\phi(u, v) = \frac{k(u, v)}{\Sigma_z k(u, z)}, \tag{3}$$

where the sum in the denominator runs over all points in $z \in \mathbb{U}$ except u itself. We also set $\phi(u, u) = 0$ for all $u \in \mathbb{U}$.

The assignments $\phi(u, v)$ place probability weights on the edges of the directed complete graph G, and we then run random walks on this graph with these weights. Choosing a positive integer m, let $\phi_m(u, v)$ be the probability that a random walk of exactly m steps starting at u will end at v. We then define the diffusion distance $D_m : U \times U \to \mathbb{R}$ by:

$$D_m(x, y) = \sqrt{\Sigma_u[\phi_m(x, u) - \phi_m(y, u)]^2}, \tag{4}$$

where the sum runs over all points $u \in \mathbb{U}$.

Now that we have a symmetric function on the edges of G, we can compute the persistence diagrams $\mathrm{Dgm}_p(\mathbb{U}, D_m)$ in each homological dimension, as described at the end of section 2. Note that the integer m is of course a parameter that we must select in advance; our preliminary experiments indicate that values of m somewhere in the $10-30$ range provide the best results, but we cannot say anything more conclusive than that at this time.

Random walk function. For an alternative set of experiments, we also consider the following symmetric function $\rho_m : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$. We again run random walk on $\mathbb{U}$, but this time taking a step from u to v with probability $\phi(u, v)$ proportional to the kernel function $\kappa(u, v) = ||u - v||^{-2}$. Choosing a positive integer m, let $\phi_{\leq m}(u, v)$ be the probability that a random walk of less than or equal to m steps which starts at u will end at v; more precisely, we set

$$\phi_{\leq m}(u, v) = \sum_{k \leq m} \phi_k(u, v),$$

where ϕ_k is defined above. Finally, we define:

$$\rho_m(u, v) = \frac{1}{\min\{\phi_{\leq m}(u, v), \phi_{\leq m}(v, u)\}}, \tag{5}$$

for $u \neq v$ and $\rho_m(u, u) = 0$, and then we compute the persistence diagrams $\mathrm{Dgm}_p(\mathbb{U}, \rho_m)$ in each homological dimension, as described at the end of section 2.

Our reasons for defining and testing this new function were as follows. First of all, we wanted to use a kernel function which did not depend on some parameter α; in our experiments with D_m, we were able to tune α correctly by exploiting prior knowledge of the data, but this cannot be done in reality. It also seemed more natural to use the random-walk probabilities ϕ_m directly, rather than summing over intermediate points as in the definition of D_m. This was especially appealing in the case of outliers: if a data point v lies far way from the rest of the data points u, then the numbers $\phi_m(u, v)$ will all be quite small, and hence $\rho_m(u, v)$ will be large. As we see below, our intuitions were partially but not fully justified by our experimental results.

Sub-sampling. In order to produce a persistence diagram from a point cloud endowed with a metric, one first has to construct the filtered Rips complex and then feed the resulting simplices to a matrix reduction algorithm [13] which runs in worst-case cubic time in the number of input simplices. Thus, assuming we start with N points, and compute only the d-skeleton of the Rips complex (for example, if we are only interested in persistent homology up to dimension $(d-1)$), the entire computation takes $O(N^{3(d+1)})$ time; hence a large number of

points will certainly lead to a very slow diagram computation. On the other hand, any random-walk-based metric must require a fairly large number of points sampled from the underlying space: if one wishes to "see" the true geometry, then one must be able to choose from a large number of paths that lie along the space.

To resolve this conflict, we propose the following technique. From an initial large sample $\mathbb{U}$ of size N, we draw a much smaller sub-sample $\mathbb{U}_0$ of size $n << N$. We compute all pairwise distances $D_m(u, v)$ for all pairs from the large sample, but we then build a filtered Rips complex using only the $\mathbb{U}_0$-points as vertices and compute the resulting persistence diagrams. The complexity is then much better: the entire algorithm from point cloud to metric ρ_m/D_m to d-skeleton of Rips complex to persistence diagram runs in $O(n^{3(d+1)} + mN^3)$.

Equivalently, one may think of defining a new metric on the space $\mathbb{U}_0$ which is based on random walk, but where the random walker is allowed to explore within a much larger set of available paths. In this way, we satisfy two goals: a small filtered simplicial complex, but one with edge weights derived from a much larger sample that more faithfully reflects the underlying space.

5. Experiments

At the end of this section, we present our main experimental results, and we argue that they provide strong evidence for random-walk methods being a solution to the two problems discussed at the end of Section 3. We begin the section with details on the implementation of the algorithms used to compute persistent homology, as well as the computation of the numbers ϕ_m, which then lead directly to the D_m and ρ_m edge weights.

5.1. Implementation

To compute persistent homology, we used the algorithm library Dionysus [17]. In particular, we took advantage of the implementations of the persistent homology algorithm from [13] and the Bron-Kerbosch algorithm [3] for the construction of the Rips complex. We also produced a MATLAB wrapper of Dionysus to streamline the design and execution of these and future experiments [14].

Probabilities. Here we describe the computation of the metrics D_m and ρ_m given a dataset of witnesses and a sub-set of landmarks. In each case, we represent the metric as a matrix of all pair-wise distances between all points in the dataset, without making a distinction between witness and landmark points. We then construct the filtered Rips complex using only the distances between landmark points.

The probability $\phi_m(u, v)$ of a random walk of exactly m steps from a point u to a point v is recursively defined using the value of $\phi_{m-1}(u, v)$ as follows:

$$\phi_m(u, v) = \sum_z \phi_{m-1}(u, z)\phi(z, v) \tag{6}$$

Thinking of each ϕ_m as an $n \times n$ matrix, where n is the total number of points, we can rewrite (6) in matrix form:

$$\phi_m = \phi_{m-1} \circ \phi, \tag{7}$$

$$\phi_{\leq m} = \sum_{k \leq m} \phi_k \tag{8}$$

Implementing (8) in a straightforward way gives us a total complexity of $O(n^3 m)$ to compute our ad-hoc metrics ρ_m. However, for the diffusion metrics D_m, we need only compute random walks for step size exactly m. In this case, rising the matrix to the power m can be done by a dichotomy algorithm that computes the result in $\log m$ multiplications, giving a total complexity of $O(n^3 \log m)$. Because of the relatively small values of m that were used in experiments, numerical precision errors due to matrix multiplication are small and can be ignored.

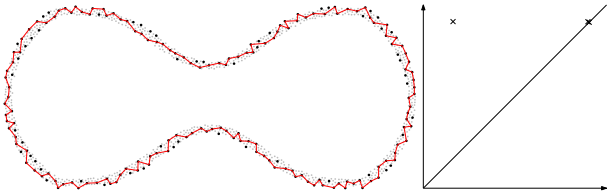
Sub-sampling. We now describe the procedure that we used to sub-sample a witness set $\mathbb{U}_0$ from a initial point sample $\mathbb{U}$. In a nutshell, we iteratively draw one landmark point from a very dense sampling of the underlying space. For each landmark point we define a no-inclusion ball of a radius δ. Each point from the dataset that is inside this ball will not be considered later, and could not be drawn in future iterations. It is obvious that no two landmark points drawn by this algorithm will lie closer than δ.

5.2. Results

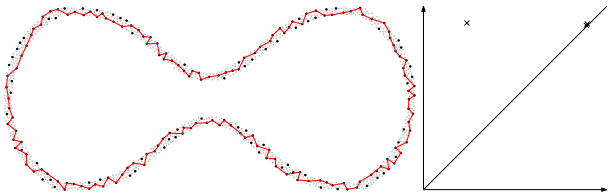
Our experiments can be divided into two main groups. First we investigated the performance of our random-walk metrics versus the Euclidean metric on several sets of points densely sampled without noise from the nearly figure-eight space; our goal was to see which metric did the best job amplifying the true topological signal from the underlying space. Then we began to progressively add noise to our sample to see which metrics preserved which type of signal.

Amplifying the signal. For this experiment we drew samples from three versions of the nearly figure eight space, with neck size varying from large to small, and for each sample we computed the one-dimensional persistence diagrams associated to the diffusion metric D_m, our metric ρ_m, and the Euclidean metric. The goal was to explore how easy it would be to distinguish signal from noise on the persistence diagram. In this synthetic case, we can of course precisely define the terms signal and noise, since we have a priori information about the original space.

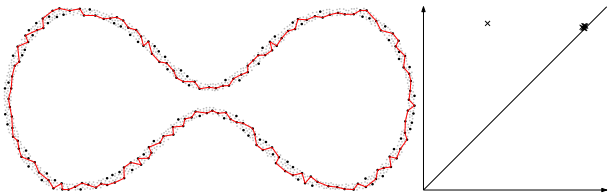
The point clouds consisted of 1800 points total on average, each with about 225 landmark points. For subselecting landmarks from the whole point cloud we used the algorithm as defined in section 5.1, with a higher value of δ than for the initial sampling of the underlying space cloud. Our primary criterion for choosing δ was to draw approximately 225 landmark points each time.



(a) Large neck size, $m = 10$

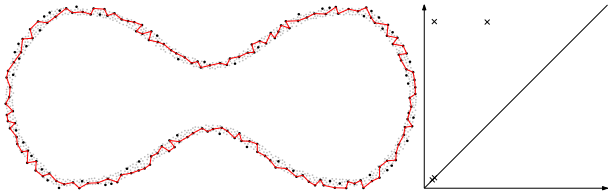


(b) Medium neck size, $m = 10$

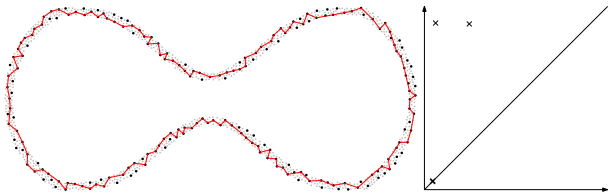


(c) Small neck size, $m = 10$

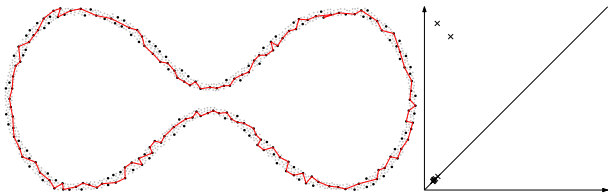
Figure 5: D_m metric



(a) Large neck size, $m = 20$.



(b) Medium neck size, $m = 20$.



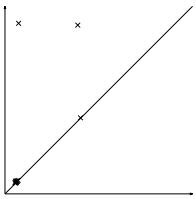
(c) Small neck size, $m = 20$.

Figure 6: ρ_m metric

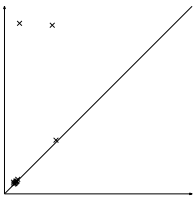
Figure 7: Witnesses appear as gray points, with landmarks as solid black points. The overlaid cycles are the canonical representatives for the most persistent homology classes. The persistence diagram for each point cloud appears to its immediate right.

We then ran a series of computations using various choices for parameter values (step size m for both metrics D_m and ρ_m; α for D_m only), and we found that these choices greatly affected the diagram outcome. Moreover, for different point clouds, the "best" parameter values differed. A table which summarizes the experimental results for all chosen parameter values can be found at `http://www.math.duke.edu/~elimgta/metrics/` [2]. Here we show only the results for the best choice of parameters.

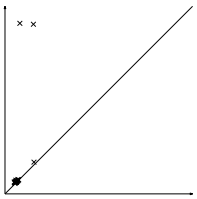
Our results for D_m and ρ_m are presented in Figure 7, and the results for Euclidean distance appear in Figure 8. As we can see, the D_m-diagrams closely resemble those of



(a) Large neck size



(b) Medium neck size

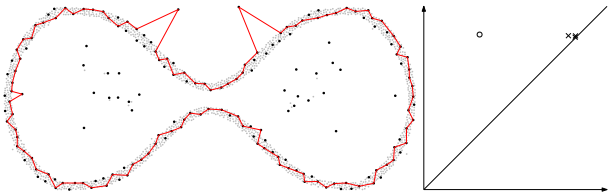


(c) Small neck size

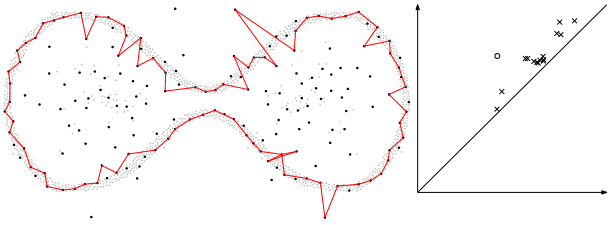
Figure 8: Persistence diagrams computed for the same point clouds as in figure 7, this time with the Euclidean metric.

a round circle (compare with figure 1) in that there is only one point of non-negligible persistence; this point does move closer to the diagonal as the neck size decreases, but the topological signal remains quite clear, even for the smallest neck, since no other classes appear in the diagram. The Euclidean diagrams, on the other hand, all feature two high-persistence points. As the neck size decreases, these points move closer to one another, making the topological signal harder to discern. A nearly identical situation holds for the ρ_m-diagrams, although the statistics in [2] show that the signal/noise ratio is in fact marginally better for ρ_m versus Euclidean in most cases.

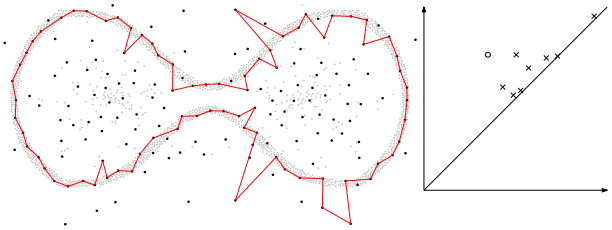
It is worth noting that one of the biggest problems that we have stumbled upon was the proper choice for α in the kernel used in the D_m metric. As can be seen in [2], "bad" choices of α lead either to incorrect homology estimates, or to signal/noise ratios close to that of the Euclidean metric. For this particular series of figure-eight datasets, we found that the best choice of α was always around one-tenth of the neck size. In general, though, we have an open question: what strategy should be used to choose the α value for datasets where we have no a priori information about the underlying topological space.



(a) Small level of noise, $m = 20$.

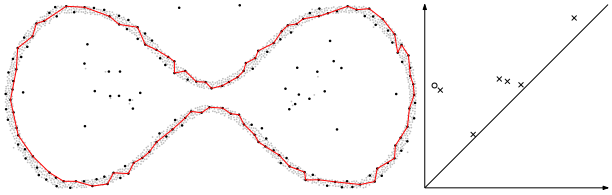


(b) Medium level of noise, $m = 20$.

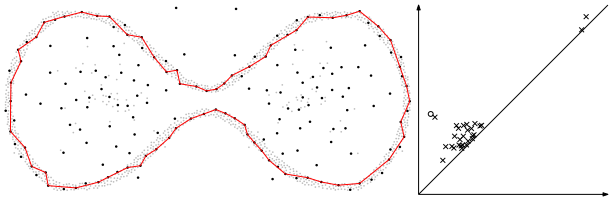


(c) High level of noise, $m = 30$.

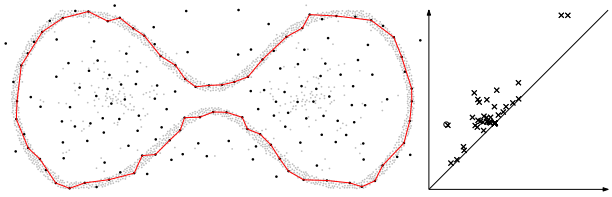
Figure 9: D_m metric



(a) Small level of noise, $m = 20$.



(b) Medium level of noise, $m = 20$.



(c) High level of noise, $m = 30$. The cycle shown here is NOT the most persistent cycle, but the second most persistent one. The most persistent one includes almost solely noise points and has very late birth time

Figure 10: ρ_m metric

Figure 11: Witnesses appear as gray points, with landmarks as solid black points. The overlaid cycles are the canonical representatives for the homology classes corresponding to the points marked by circles in the diagrams. The persistence diagram for each point cloud appears to its immediate right.

Exploring robustness. As mentioned above, persistence diagrams computed using the Euclidean metric are extremely sensitive to even a small amount of very bad noise. In a second set of experiments, we investigated the robustness of the D_m and ρ_m diagrams in the presence of a increasing number of outliers.

For these experiments, we drew three different samples from the same nearly figure-eight space; each sample contained some outliers, with the number of outliers increasing with each draw. Each time, the goal was the same as for the first set of experiments: draw out the true topological signal.

The sampling and sub-sampling procedures were the same as before: the point clouds consisted of 1800 points on average, from which we drew out around 225 landmark points. Note that most of the outlier points became landmarks, a clear consequence of the sub-sampling procedure described above.

As before, we ran a series of computations using various parameter values, and summarized all results in a table [2]. The results for the best parameter choices are shown in Figure 11. In the left column, we see the results for the D_m-metric. These diagrams show a weaker topological signal than in the no-noise experiments. The canonical representative for the most persistent class hints at the underlying space, but also contains quite a few outliers; Reading the left column top-to-bottom, we see that the persistence of spurious classes increases with the number of outliers.

The right column of the same figure displays the ρ_m-results. Again, there is certainly no crystal-clear topological signal, at least compared to the no-noise experiments. However, one feature of the ρ_m metric that inspired us to publish its results is the fact that its diagrams do contain a point with non-negligible persistence whose canonical representative cycle traces out the underlying space without including any outliers. This compares well with the cycles for D_m, which always seem to pick out outliers.

6. Discussion

The ideas and experiments described above demonstrate how the use of methods on random walks rather than the usual Euclidean distance provides a more robust and powerful way to use topological data analysis (TDA) in making measurements of dataset shape. The approach we have put forward stands in sharp contrast to methods where noise is only addressed in a pre-processing step. Instead it integrates the process of de-noising and homological calculation, and incorporates the important ideas put forward in [9] that use local proximity information in the determination of global structure. Combined with other new approaches of this form, e.g. [6], we believe that TDA is moving to a new level.

Most real world datasets have some level of noise, and the challenge now is to fine tune the random walk metric to different types of data. We are particularly interested in discovering stratified structure, based on the ideas of [1].

We sum up with the visual image of a triangle, with vertices corresponding to TDA, diffusion geometry (DG) and statistical methodology (S); see Figure 12. This paper makes a first stab at the edge from TDA to DG, and the papers [6] and [16] make steps towards the

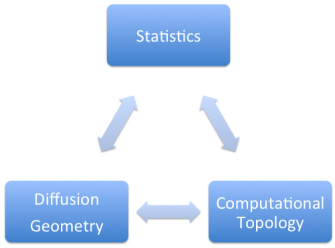


Figure 12: A triangle of future research directions.

edge from TDA to S. These methods can be used together to make homological inference the effective shape analysis tool that it promises to be.

References

[1] P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov. Inferring local homology from sampled stratified spaces. In *Proceedings 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 536–546, 2007.

[2] Paul Bendich, Taras Galkovskyi, and John Harer. Experimental results for diffusion and random walk metrics. `http://www.math.duke.edu/~elimgta/metrics/`, 2011.

[3] C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575577, 1973.

[4] G. Carlsson and T. Ishkhanov. A topological analysis of the space of natural images. *Preprint*, 2011.

[5] F. Chazal, D. Cohen-Steiner, M. Glisse, L. Guibas, and S. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings 25th Annual Symposium on Computational Geometry*, pages 237–246, 2009.

[6] F. Chazal, D. Cohen-Steiner, and Q. Merigot. Geometric inference for measures based on distance functions. *Research Report 6930, INRIA*, 2010.

[7] D. Cohen-Steiner, H. Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, **37**:103–120, 2007.

[8] R. Coifman and S. Lafon. Diffusion maps. *Applied and Comput. Harm. Analysis*, 21:5–30, 2006.

[9] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data. part i: Diffusion maps. *Proc. of Nat. Acad. Sci*, 102:7426–7431, 2005.

[10] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data. part ii: Multiscale methods. *Proc. of Nat. Acad. Sci*, 102:7432–7438, 2005.

[11] V. de Silva and G. Carlsson. Topological estimation using witness complexes. *Symposium on Point-Based Graphics*, pages 157–166, 2004.

[12] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[13] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, **28**:511–533, 2002.

[14] Taras Galkovskyi. Kapitoshka: Matlab wrappers for dionysus. `http://www.math.duke.edu/~elimgta/kapitoshka/`.

[15] J. Kloke and G. Carlsson. Topological de-noising: Strengthening the topological signal. *Preprint*, 2011.

[16] Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Preprint*, 2011.

[17] D. Morozov. Dionysus: C++ library for computing persistent homology. `http://www.mrzv.org/software/dionysus/`.

[18] J. Munkres. *Elements of algebraic topology*. Addison-Wesley, Redwood City, California, 1984.