The Annals of Applied Probability
2017, Vol. 27, No. 4, 2032–2060
DOI: 10.1214/16-AAP1232

MAXIMALLY PERSISTENT CYCLES IN RANDOM GEOMETRIC COMPLEXES

BY OMER BOBROWSKI[*,1], MATTHEW KAHLE[†,2] AND PRIMOZ SKRABA[‡,3]

Duke University, The Ohio State University† and*
Jozef Stefan Institute & University of Primorska‡

We initiate the study of persistent homology of random geometric simplicial complexes. Our main interest is in maximally persistent cycles of degree-k in persistent homology, for a either the Čech or the Vietoris–Rips filtration built on a uniform Poisson process of intensity n in the unit cube $[0,1]^d$. This is a natural way of measuring the largest "k-dimensional hole" in a random point set. This problem is in the intersection of geometric probability and algebraic topology, and is naturally motivated by a probabilistic view of topological inference.

We show that for all $d \geq 2$ and $1 \leq k \leq d-1$ the maximally persistent cycle has (multiplicative) persistence of order

$$\Theta\left(\left(\frac{\log n}{\log\log n}\right)^{1/k}\right),$$

with high probability, characterizing its rate of growth as $n \to \infty$. The implied constants depend on k, d and on whether we consider the Vietoris–Rips or Čech filtration.

1. Introduction. The study of topological properties of random graphs has a long history, dating back to classical results on the connectivity, cycles and largest components in Erdős–Renyi graphs [29, 30]. Generalizations have been developed in several directions. One direction is to consider different models of random graphs (see, e.g., [12, 44]). Another direction is to consider higher-dimensional topological properties, resulting in the study of *random simplicial complexes* rather than random graphs, where in addition to vertices and edges the structure consists also of triangles, tetrahedra and higher dimensional simplexes (see, e.g., [3, 37, 39, 41]). The study of random simplicial complexes focuses mainly on their homology, which is a natural generalization of the notions of connected components and cycles in graphs. Homology is an algebraic topology framework that is used to study cycles in various dimensions, where (loosely speaking) a k-dimensional

Received September 2015; revised May 2016.

[1] Supported by NSF Grants DMS-14-18261 and DMS-12-09155.

[2] Supported by DARPA #N66001-12-1-4226, NSF Grant CCF-1017182, NSF Grant DMS-1352386 and the Alfred P. Sloan Foundation.

[3] Supported by the EU projects FP7-612329 ProaSense and FP7-ICT-318493-STREP Toposys.

MSC2010 subject classifications. 60D05, 05C80, 62G32.

Key words and phrases. Persistent homology, topological inference, stochastic topology.

cycle can be thought of as the boundary of a $k+1$ dimensional solid (see Section 2 for more details).

In random *geometric* simplicial complexes, the vertices are generated by a random point process (e.g., Poisson) in a metric space, and then geometric conditions are applied to determine which of the simplexes should be included in the complex. The two most studied models are the random Čech and Vietoris-Rips complexes (see Section 2 for definitions). Several recent papers have studied various aspects of the topology of these complexes (see [6, 9, 11, 38, 40, 50, 51] and the survey [8]). These papers contain theorems which characterize the phase transitions where homology appears and disappears, estimates for the Betti numbers (the number of k-dimensional cycles), limiting distributions, etc. While this line of research presents a deep and interesting theory, it is also motivated by data analysis applications.

Topological data analysis (TDA) is a recently emerging field that focuses on extracting topological features from sampled data, and uses them as an input for various data analytic and statistical algorithms. The main idea behind it is that topological properties could help us understand the structure underlying the data, and provide us with a set of features that are robust to various types of deformations (cf. [16, 17, 33]). Geometric complexes play a key role in computing topological features from a finite set of data points. The construction of these complexes usually depends on one or more parameters (e.g., radius of balls drawn around the sample points), and the ability to properly extract topological features depends on choosing this parameter correctly. One of the most powerful tools in TDA is a multi-scale version of homology, called *persistent homology* (see Section 2), which was developed mainly to solve this problem of sensitive parameter tuning. In persistent homology, instead of finding the best parameter values, one considers the entire range of possible values. As the parameter values change, the observed topological features change (e.g., cycles are created and filled in). Persistent homology tracks these changes and provides a way to measure the significance of the features that show up in this process. One way to represent the information provided by persistent homology is via *barcodes*; see Figure 2. Here, every bar corresponds to a feature in the data and its endpoints correspond to the times (parameter value) where the feature was created and terminated. The underlying philosophy in TDA is that topological features that survive (or persist) through a long range of parameter values are significant and related to real topological structures in the data (or the "topological signal"), whereas ones with a shorter lifespan are artifacts of the finite sampling, and correspond to noise (see [31]). This approach motivates the following question: How long does a "long range" of parameters (or a long bar in the barcode) have to be in order to be considered significant? Phrased differently—how long should we expect this range to be, if the sample points were entirely random, without any underlying structure or features? This is the main question we try to answer in this paper.

To be more specific, in this paper we study the case where the data points are generated by a homogeneous Poisson process in the unit d-dimensional cube $[0,1]^d$ $(d > 1)$ with intensity n, denoted by $\mathcal{P}_n$. We consider the persistent homology of both the Čech complex $\mathcal{C}(\mathcal{P}_n, r)$ and the Rips complex $\mathcal{R}(\mathcal{P}_n, r)$, where the scale parameter r is the radius of the balls used to create these complexes (see Section 2). We denote by $\Pi_k(n)$ the maximal persistence of a cycle in the degree k persistent homology $(1 \le k \le d-1)$ of either the Čech or the Rips complex. Note that $\Pi_k(n)$ is a property of the persistent homology, where we consider all possible radii and, therefore, it does not depend r. Our main result shows that, with high probability,

$$\Pi_k(n) \sim \left(\frac{\log n}{\log\log n}\right)^{1/k},$$

in the sense that $\Pi_k(n)$ can be bounded from above and below by a matching term up to a constant factor. The precise definitions and statements are presented in Section 3. The proofs for the upper and lower bounds require very different techniques. To prove the upper bound, we present a novel "isoperimetric-type" statement (Lemma 4.1) that links the persistence of a cycle to the number of vertices that are used to form it. The lower bound proof uses an exhaustive search for a specific construction that guarantees the creation of a persistent cycle.

In addition to proving the theoretical result, in Section 7 we also present extensive numerical experiments confirming the computed bounds and empirically computing the implied constants. These results also suggest a conjectural law of large numbers. Finally, we note that while the results in this paper are presented for the homogeneous Poisson process on a d-dimensional cube, they should hold with minor adjustments also to nonhomogenous processes as well as for shapes other than the cube. We also predict that our statements will hold for more generic point processes (e.g., weakly sub-Poisson processes), using some of the statements made in [50]. The detailed analysis of these more generic cases is left as future work.

Earlier work: The study of the topology of random geometric complexes has been growing rapidly in the past decade. Most of the results so far are related to homology rather than persistent homology (i.e., fixing the parameter value). The study in [11, 38] focuses mainly on the phase transitions for appearance and vanishing of homology, which can be viewed as higher dimensional generalizations of the phase transition for connectivity in random graphs. In [6, 9, 40, 51], more emphasis was given to the distribution of the Betti numbers, namely the number of cycles that appear. Similar questions for more general point processes have also been considered in [50]. In [1, 43], simplicial complexes generated by distributions with an unbounded support were studied from an extreme value theory perspective. The recent survey [8] overviews recent progress in this area.

The study of random persistent homology, on the other hand, is at its very initial stages. Recall that the 0th homology represent the connected components in a space. Thus, the results in [2, 45] about the connectivity threshold in random geometric graphs could be viewed as related to the 0th persistence homology of either the Čech or the Rips complex. The first study of persistent homology in degree $k \geq 1$ for a random setting was for n points chosen uniformly i.i.d. on a circle by Bubenik and Kim [14]. In this setting, they used the theory of order statistics to describe the limiting distribution of the persistence diagram. Another direction of study is the persistence diagrams of random functions. In [7], the authors study the "persistent Euler characteristic" of Gaussian random fields.

Another line of research (see, e.g., [10, 19–23, 31]) focuses on statistical inference using persistent homology, and include results about confidence intervals, consistency and robustness for topological estimation, subsampling and bootstrapping methods and more.

Finally, we point out the earlier work in geometric probability [4], measuring the largest convex hole for a set of random points in a convex planar region R. A convex hole is generated when there is a subset of points for which the convex hull is empty (i.e., contains no other points from the set). The size of a convex hole is then measured combinatorially, as the number of vertices generating the hole. In [4], it is shown that the largest hole has $\Theta(\log n/\log\log n)$ vertices, regardless of the shape of the ambient convex region R. In this paper, we are also measuring the size of the largest hole, but in a very different sense. We are using the algebraic-topological notion of holes (via persistent homology), rather than combinatorial notion of counting vertices, so as far as we can tell the fact that these two ways of measuring the size of the largest hole have the same right of growth (when $d = 2$ and $k = 1$) is something of a coincidence.

As far as we know, this article presents the first detailed probabilistic analysis for persistent kth homology of random geometric complexes, for $k \geq 1$.

The structure of the paper is as follows. In Section 2, we provide the topological and probabilistic building blocks we will use throughout the paper. In Section 3, we present the main result—the asymptotic behavior of maximally persistent cycles. In Sections 4 and 5, we provide the main parts of the proof for the random Čech complex (upper and lower bounds, respectively). Some parts of the proofs require more knowledge in algebraic topology than the others, and we present those in Section 6 (including the proof for the Rips complex). Finally, in Section 7 we present simulation results, complementing the main (asymptotic) result of the paper.

2. Background.

In this section, we provide a brief introduction to the topological and probabilistic notions used in this paper.

2.1. *Homology.*

We wish to introduce the concept of homology here in an intuitive rather than in a rigorous way. For a comprehensive introduction to homology, see [35] or [42]. Let X be a topological space, and $\mathbb{F}$ a field. The *homology* *of X with coefficients in* $\mathbb{F}$ is a set of vector spaces $\{H_k(X)\}_{k=0}^{\infty}$, which are topological invariants of X (i.e., they are invariant under homeomorphisms). We note that the standard notation is $H_k(X, \mathbb{F})$ where $\mathbb{F}$ denotes the coefficient ring, but we suppress the field and let $H_k(X)$ denote homology with $\mathbb{F}$ coefficients throughout this article.

The dimension of the zeroth homology $H_0(X)$ is equal to the number of connected components of X. For $k \geq 1$, the basis elements of the kth homology $H_k(X)$ correspond to k-dimensional "holes" or (nontrivial-) "cycles" in X. An intuitive way to think about a k-dimensional cycle is as the result of taking the boundary of a $(k+1)$-dimensional body. For example, if X a circle then $H_0(X) \cong \mathbb{F}$, and $H_1(X) \cong \mathbb{F}$. If X is a 2-dimensional sphere, then $H_0(X) \cong \mathbb{F}$ and $H_2(X) \cong \mathbb{F}$, while $H_1(X) \cong \{0\}$ (since every loop on the sphere can be shrunk to a point). In general, if X is a n-dimensional sphere, then

$$H_k(X) \cong \begin{cases} \mathbb{F}, & k = 0, n, \\ 0, & \text{otherwise.} \end{cases}$$

We will use $H_*(X)$ when making a statement that applies to all the homology groups simultaneously. In addition to providing information about spaces, homology is also used to study mappings between spaces. If $f : X \to Y$ is a map between two topological spaces, then it induces a map in homology $f_* : H_*(X) \to H_*(Y)$. This map is a linear transformation between vector spaces which tells us how cycles in X map to cycles in Y. These mappings are important when discussing persistent homology.

Finally, we say that two spaces X, Y are *homotopy equivalent*, denoted by $X \simeq Y$, if X can be continuously deformed to Y (loosely speaking). In particular, if $X \simeq Y$ then $H_*(X) \cong H_*(Y)$ (isomorphic). For example, a circle, an empty triangle and an annulus are all homotopy equivalent.

2.2. *The Čech and Vietoris–Rips complexes.* As mentioned earlier, the Čech and the Rips complexes are often used to extract topological information from data. These complexes are abstract simplicial complexes [35] and in our case will be generated by a set of points in $\mathbb{R}^d$. These complexes are tied together with the union of balls we define as

$$\mathcal{U}(\mathcal{P}, r) = \bigcup_{p \in \mathcal{P}} B_r(p), \tag{2.1}$$

where $\mathcal{P} \subset \mathbb{R}^d$, and $B_r(p)$ is a d-dimensional ball of radius r around p. Note that the set $\mathcal{P}$ does not have to be discrete, in which case we can think of $\mathcal{U}(\mathcal{P}, r)$ as a "tube" around $\mathcal{P}$. The definitions of the complexes are as follows.

DEFINITION 2.1 (Čech complex). Let $\mathcal{P} = \{x_1, x_2, \ldots, x_n\}$ be a collection of points in $\mathbb{R}^d$, and let $r > 0$. The Čech complex $\mathcal{C}(\mathcal{P}, r)$ is constructed as follows:

(1) The 0-simplices (vertices) are the points in $\mathcal{P}$.
(2) A k-simplex $[x_{i_0}, \dots, x_{i_k}]$ is in $\mathcal{C}(\mathcal{P}, r)$ if $\bigcap_{j=0}^{k} B_r(x_{i_j}) \neq \varnothing$.

DEFINITION 2.2 (Vietoris–Rips complex). Let $\mathcal{P} = \{x_1, x_2, \dots, x_n\}$ be a collection of points in $\mathbb{R}^d$, and let $r > 0$. The Vietoris–Rips complex $\mathcal{R}(\mathcal{P}, r)$ is constructed as follows:

(1) The 0-simplices (vertices) are the points in $\mathcal{P}$.
(2) A k-simplex $[x_{i_0}, \dots, x_{i_k}]$ is in $\mathcal{R}(\mathcal{P}, r)$ if $B_r(x_{i_j}) \cap B_r(x_{i_l}) \neq \varnothing$ for all $0 \leq j, l \leq k$.

Note that the Rips complex $\mathcal{R}(\mathcal{P}, r)$ is the flag (or clique) complex built on top of the geometric graph $G(\mathcal{P}, 2r)$, where two vertices x_i, x_j are connected if and only if $\|x_i - x_j\| \leq 2r$. The difference between the Čech and the Rips complexes, is that for the Čech complex we require all $k+1$ balls to intersect in order to include a face, whereas for the Rips complex we only require pairwise intersections between the balls. Figure 1 shows an example for the Čech and Rips complexes constructed from the same set of points and the same radius r, and highlights this difference.

Part of the importance of the Čech complex stems from the following statement known as the "Nerve lemma" (see [13]). We note that the original lemma is more general then stated here, but we will only be using it in the following special case.

LEMMA 2.3. *Let $\mathcal{P} \subset \mathbb{R}^d$ be a finite set of points. Then $\mathcal{C}(\mathcal{P}, r)$ is homotopy equivalent to $\mathcal{U}(\mathcal{P}, r)$, and in particular*

$$H_*(\mathcal{C}(\mathcal{P}, r)) \cong H_*(\mathcal{U}(\mathcal{P}, r)).$$

The Rips complex is commonly used in applications, as in some practical cases it requires less computational resources. In an arbitrary metric space, using the

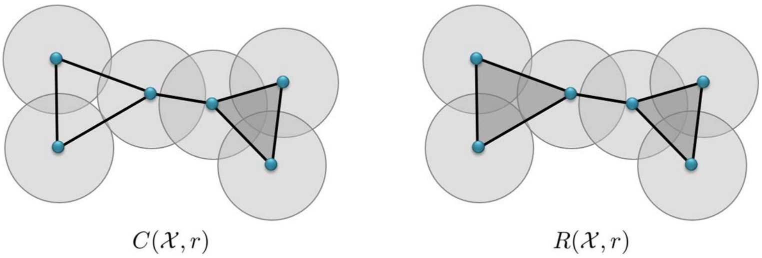


FIG. 1. *On the left—the Čech complex $\mathcal{C}(\mathcal{P}, r)$, on the right—the Rips complex $\mathcal{R}(\mathcal{P}, r)$ with the same set of vertices and the same radius. We see that the three left-most balls do not have a common intersection and, therefore, do not generate a 2-dimensional face in the Čech complex. However, since all the pairwise intersections occur, the Rips complex does include the corresponding face.*

triangle inequality we have the following inclusions of complexes:

$$\mathcal{C}(\mathcal{P}, r) \subset \mathcal{R}(\mathcal{P}, r) \subset \mathcal{C}(\mathcal{P}, 2r). \tag{2.2}$$

For subsets of Euclidean space, the constant 2 can be improved (see [25]).

2.3. *Persistent homology.* Let $\mathcal{P} \subset \mathbb{R}^d$, and consider the following indexed sets:

$$\mathcal{U} := \{\mathcal{U}(\mathcal{P}, r)\}_{r=0}^{\infty}, \qquad \mathcal{C} := \{\mathcal{C}(\mathcal{P}, r)\}_{r=0}^{\infty}, \qquad \mathcal{R} := \{\mathcal{R}(\mathcal{P}, r)\}_{r=0}^{\infty}.$$

These three sets are examples of "filtrations"—nested sequences of sets, in the sense that $\mathcal{F}_{r_1} \subset \mathcal{F}_{r_2}$ if $r_1 < r_2$ (where $\mathcal{F}$ is either $\mathcal{U}$, $\mathcal{C}$, or $\mathcal{R}$).

As the parameter r increases, the homology of the spaces $\mathcal{F}_r$ may change. The *persistent homology* of $\mathcal{F}$, denoted by $\mathrm{PH}_*(\mathcal{F})$, keeps track of this process. Briefly, $\mathrm{PH}_k(\mathcal{F})$ contains information about the kth homology of the individual spaces $\mathcal{F}_r$ as well as the mappings between the homology of $\mathcal{F}_{r_1}$ and $\mathcal{F}_{r_2}$ for every $r_1 < r_2$ (induced by the inclusion map). The *birth time* of an element (a cycle) in $\mathrm{PH}_*(\mathcal{F})$ can be thought of as the value of r where this element appears for the first time. The *death time* is the value of r where an element vanishes, or merges with another existing element.

Formally, we consider a filtration with parameter values from $[0, \infty)$, the birth and death times can be defined as the following.

DEFINITION 2.4. The *birth* of an element $\gamma \in \mathrm{PH}_k(\mathcal{F})$ is

$$\gamma_{\text{birth}} := \min\{r : \gamma \in H_k(X_r)\}.$$

DEFINITION 2.5. The *death time* of an element $\gamma \in \mathrm{PH}_k(\mathcal{F})$ is

$$\gamma_{\text{death}} := \min\{r : \gamma \in \ker(H_k(X_{\gamma_{\text{birth}}}) \to H_k(X_r))\}.$$

One useful way to describe persistent homology is via the notion of *barcodes* [33]. A barcode for the persistent homology of a filtration $\mathcal{F}$ is a collection of graphs, one for each order of homology group. A bar in the kth graph, starting at b and ending at d ($b \le d$) indicates the existence of an element of $\mathrm{PH}_k(\mathcal{F})$ (or a k-cycle) whose birth and death times are b and d, respectively. In Figure 2, we present the barcode for the filtration $\mathcal{U}$ where $\mathcal{P}$ is a set of 50 random points lying inside an annulus. The intuition is that the longest bars in the barcode represent "true" features in the data (e.g., the connected component and the 1-cycle in the annulus), whereas the short bars are regarded to as "noise." It can be shown that the pairing between birth and death times is sufficient to yield a unique barcode [52].

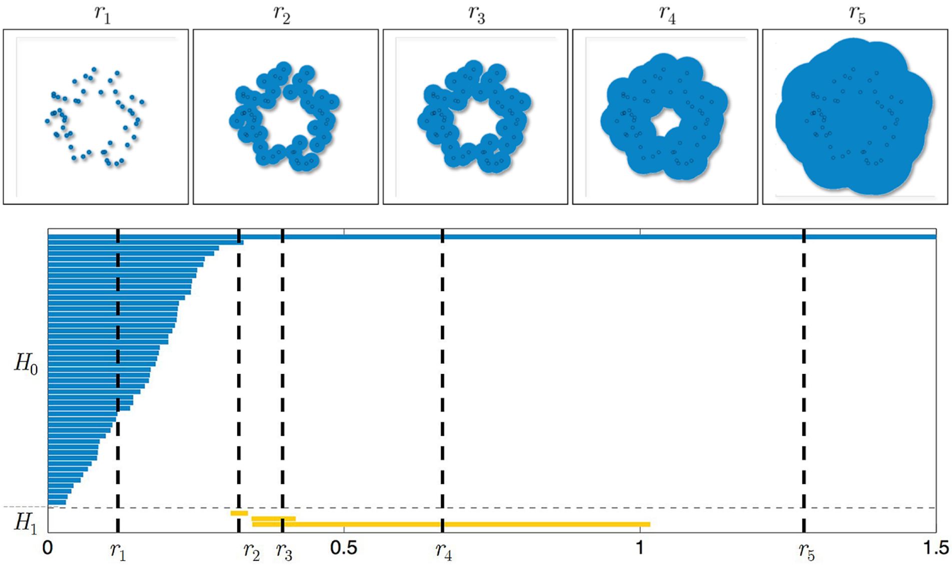


FIG. 2. (top) *$\mathcal{F}_r = \mathcal{U}_r$ is a union of balls of radius r around $\mathcal{P}$—a random set of $n = 50$ points, uniformly distributed on an annulus in $\mathbb{R}^2$. We present five snapshots of this filtration.* (bottom) *The persistent homology of the filtration $\mathcal{F}$. The x-axis is the radius r, and the bars represent the cycles that born and die. For H_0, we observe that at radius zero the number of components is exactly n and as the radius increases components merge (or die). The 1-cycles show up later in this process. There are two bars that are significantly longer than the others (one in H_0 and one in H_1). These correspond to the true features of the annulus.*

2.4. *The Poisson process.* In this paper, the set of points we use to construct either a Čech or a Rips complex will be generated by a Poisson process $\mathcal{P}_n$, which can be defined as follows. Let $X_1, X_2, \dots$ be an infinite sequence of i.i.d. (independent and identically distributed) random variables in $\mathbb{R}^d$. We will focus on the case where X_i is uniformly distributed on the unit cube $\mathcal{Q}^d = [0, 1]^d$. We note, however, that our results hold (with minor adjustments) for any distribution with a compact support and density bounded above and below. Next, fix $n > 0$, take $N \sim \text{Poisson}(n)$, independent of the X_i's, and define

$$\mathcal{P}_n = \{X_1, X_2, \dots, X_N\}. \tag{2.3}$$

Two properties characterizing the Poisson process $\mathcal{P}_n$ are:

(1) For every Borel-measurable set $A \subset \mathbb{R}^d$ we have that

$$|\mathcal{P}_n \cap A| \sim \text{Poisson}\big(n \operatorname{Vol}(A \cap \mathcal{Q}^d)\big),$$

where $|\cdot|$ stands for the set cardinality, and $\operatorname{Vol}(\cdot)$ is the Lebesgue measure.

(2) If $A, B \subset \mathbb{R}^d$ are disjoint sets, then $|\mathcal{P}_n \cap A|$ and $|\mathcal{P}_n \cap B|$ are independent random variables (this property is known as “spatial independence”).

The Poisson process $\mathcal{P}_n$ is closely related to the fixed-size set $\{X_1, \dots, X_n\}$. Note that the expected number of points in $\mathcal{P}_n$ is $\mathbb{E}\{N\} = n$. In fact, most results known for one of these processes apply to the other with very minor, or no, changes. This is true for the results presented in this paper as well. However, we choose to focus only on $\mathcal{P}_n$, mainly due to its spatial independence property.

In the following, we study asymptotic phenomena, when $n \to \infty$. In this context, if E_n is an event that depends on n, we say that E_n occurs *with high probability* (w.h.p.) if $\lim_{n\to\infty} \mathbb{P}(E_n) = 1$.

3. Maximally persistent cycles.

For the remainder of this paper, assume that $d \geq 2$ and $1 \leq k \leq d-1$ are fixed. Let $\mathcal{P}_n$ be the Poisson process defined above, and define

$$\mathcal{U}(n,r) := \mathcal{U}(\mathcal{P}_n, r), \qquad \mathcal{C}(n,r) := \mathcal{C}(\mathcal{P}_n, r), \qquad \mathcal{R}(n,r) := \mathcal{R}(\mathcal{P}_n, r).$$

Let $\mathrm{PH}_k(n)$ be the kth persistent homology of either of the filtrations for $\mathcal{U}$, $\mathcal{C}$, or $\mathcal{R}$ (it will be clear from the context which filtration we are looking at). Note that from the Nerve Lemma (2.3), we have that $\mathcal{U}(n,r) \simeq \mathcal{C}(n,r)$, so we will state the results only for $\mathcal{C}$ and $\mathcal{R}$. However, some of the statements we make are easier to prove for the balls in $\mathcal{U}$ rather than the simplexes in $\mathcal{C}$, and we shall do so.

For every cycle $\gamma \in \mathrm{PH}_k(n)$, we denote by $\gamma_{\mathrm{birth}}, \gamma_{\mathrm{death}}$ the birth and death times (radii) of γ, respectively. Commonly (see [16, 33]), the *persistence* of a cycle is measured by the length of the corresponding bar in the barcode, namely by the difference $\delta(\gamma) := \gamma_{\mathrm{death}} - \gamma_{\mathrm{birth}}$. In this paper, however, we choose to define the persistence of γ in a multiplicative way as

$$\pi(\gamma) := \frac{\gamma_{\mathrm{death}}}{\gamma_{\mathrm{birth}}}. \tag{3.1}$$

There are several reasons for defining the persistence of a cycle this way:

- This definition is equivalent to saying that we measure the difference in a logarithmic scale. Studying persistent homology in the logarithmic scale is common [15, 22, 36, 46, 49].
- This definition is scale invariant, which is desirable, since "topological significance" should focus on shape rather than size. For example, consider the cycles corresponding to γ_1, γ_2 in Figure 3. These two cycles are created by exactly the same configuration of points, just at a different scale. Therefore, we would like to say that these cycles are equally significant. Clearly, $\delta(\gamma_1) > \delta(\gamma_2)$, while $\pi(\gamma_1) = \pi(\gamma_2)$. Thus, our definition works better in this case.

  In addition, this scale invariance guarantees that a linear change in the units used to measure the data (e.g., from inches to cm, or from degrees Celsius to Fahrenheit) will not affect the persistence value.
- One purpose of using a persistence measure is to differentiate between cycles that capture phenomena underlying the data, and those who are created merely

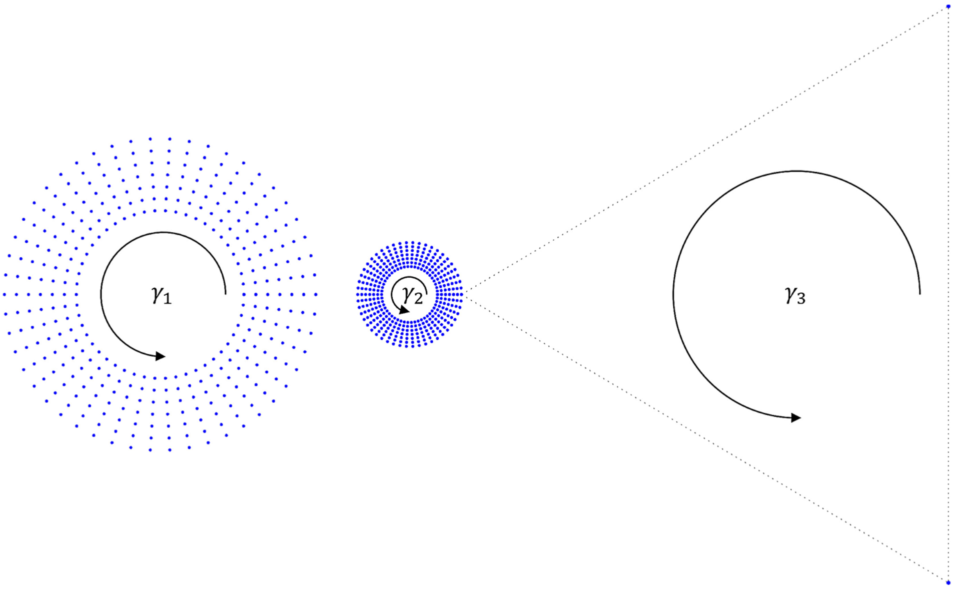


FIG. 3. *Multiplicative persistence as a significance measure. The dataset in this example consists of a few hundred points sampled from two annuli, and two outliers (on the right). We are interested in the* 1*-cycles that denoted by* $\gamma_1, \gamma_2, \gamma_3$*, that correspond to the two annuli and the triangle on the right.*

due to chance. To this end, the "physical size" of the cycle is not necessarily the correct measure. Consider, for example, the cycles corresponding to γ_2 and γ_3 in Figure 3. Intuitively, we would like to claim that γ_2 is more significant than γ_3, as the former is created by a very "stable" configuration of points, while the latter is created by outliers that clearly tell us nothing about the underlying structure. In this example, taking the "additive" persistence we will have that $\delta(\gamma_2) < \delta(\gamma_3)$, simply because the overall size of the annulus is much smaller than that of the triangle. However, taking multiplicative persistence yields $\pi(\gamma_2) > \pi(\gamma_3)$, which is more consistent with our intuition.

- Both the Čech and Vietoris–Rips complexes are important in TDA, and the natural relationship between these complexes is a multiplicative one [see (2.2)]. Because of this relationship, our results hold for both random Čech and Rips complexes, up to a constant factor (see Section 6.3). Indeed, the majority of approximation results for geometric complexes are multiplicative [18, 26, 48], making multiplicative persistence more relevant to existing stability guarantees.
- The argument from Section 5 of this paper suggests that there are many cycles γ for which $\gamma_{\text{birth}} = o(\gamma_{\text{death}})$. In this case, it is hard to differentiate between cycles by looking at $\gamma_{\text{death}} - \gamma_{\text{birth}} \approx \gamma_{\text{death}}$.

Our main interest is in the maximal persistence over all k-cycles, defined as

$$\Pi_k(n) := \max_{\gamma \in \mathrm{PH}_k(n)} \pi(\gamma). \tag{3.2}$$

More specifically, we are interested in the asymptotic behavior of $\Pi_k(n)$ as $n \to \infty$. The main result in this paper is that $\Pi_k(n)$ scales like the function $\Delta_k(n)$, defined by

$$\Delta_k(n) := \left(\frac{\log n}{\log\log n}\right)^{1/k}. \tag{3.3}$$

In particular, we have the following theorem.

THEOREM 3.1. *For fixed $d \geq 2$, and $1 \leq k \leq d-1$, let $\mathcal{P}_n$ be a Poisson process on the unit cube $[0,1]^d$ defined in* (2.3), *and let* $\mathrm{PH}_k(n)$ *be the kth persistent homology of either $\mathcal{C}$, or $\mathcal{R}$. Then there exist positive constants A_k, B_k such that*

$$\lim_{n\to\infty} \mathbb{P}\left(A_k \leq \frac{\Pi_k(n)}{\Delta_k(n)} \leq B_k\right) = 1.$$

REMARKS.

(1) The constants A_k and B_k depend on k (the homology degree), d (the ambient dimension), and on whether we consider the Čech or the Rips complex. We conjecture that a law of large numbers holds, namely that $\Pi_k(n)/\Delta_k(n) \to C_k$ for some $A_k \leq C_k \leq B_k$. For some evidence for this conjecture, see the experimental results in Section 7. In the following sections, we will prove Theorem 3.1.

(2) The additive persistence $\delta(\gamma)$ can be bounded naively by the result on the contractibility of the Čech complex in [38]. More concretely, Theorem 6.1 states that if $r \geq c(\frac{\log n}{n})^{1/d}$ then the Čech complex is contractible (w.h.p.). This implies that for every cycle γ we have $\delta(\gamma) \leq \gamma_{\mathrm{death}} \leq c(\frac{\log n}{n})^{1/d}$. Similar statements can be made about PH_0 using the connectivity radius in [2, 45] (which is of the same $(\log n/n)^{1/d}$ scale). However, these are only crude upper bounds on the additive persistence, that do not differentiate between the different cycles in persistent homology, or even between different degrees of homology (note that these bounds do not depend on k).

(3) The study in [38] suggests the following upper bound for $\Pi_k(n)$. As mentioned before, we know that $\gamma_{\mathrm{death}} \leq c(\frac{\log n}{n})^{1/d}$ for all γ. In addition, the analysis in [38] shows that if $n^{k+1}r^{dk} \to 0$ then $H_k(\mathcal{C}(n,r)) = 0$, which implies that $\gamma_{\mathrm{birth}} \geq c' n^{-\frac{k+2}{d(k+1)}}$ for some $c' > 0$. Therefore, we have that $\pi(\gamma) = O((\log n)^{1/d} n^{\frac{1}{d(k+1)}})$. However, as we shall see later, this is a very crude upper bound.

4. Proof—upper bound. For this section and the next one, consider the Čech complex only. We want to prove the upper bound in Theorem 3.1. That is, we need to show that there exists a constant $B_k > 0$ depending only on k and d, so that with high probability

$$\Pi_k(n) \le B_k \Delta_k(n) = B_k \left(\frac{\log n}{\log\log n}\right)^{1/k}.$$

The main idea in proving the upper bound in Theorem 3.1 is to show that large cycles require the formation of a large connected component in $\mathcal{C}(n,r)$ at a very early stage (small radius r). To this end, we will provide two bounds: (1) a lower bound for the size of the connected component supporting a large cycle (Lemma 4.1), and (2) an upper bound for the size of connected components in $\mathcal{C}(n,r)$ for small values of r (Lemma 4.2).

LEMMA 4.1. *Let $\gamma \in \mathrm{PH}_k(n)$, with $\gamma_{\mathrm{birth}} = r$ and $\pi(\gamma) = p$. Then there exists a constant C_1 such that $\mathcal{C}(n,r)$ contains a connected component with at least $m = C_1 p^k$ vertices. The constant C_1 depends on k, d only.*

The proof for this lemma requires more working knowledge in algebraic topology than the rest of this paper, and we defer it to Section 6. At this point, we would like to suggest an intuitive explanation. Suppose that $\mathcal{C}(n,r)$ contains a k-cycle such that all the points generating it lie on a k-dimensional sphere of radius R, and such that there are no points of $\mathcal{P}_n$ inside the sphere. In that case, the death time of the cycle will be R and then $\pi(\gamma) = p \ge R/r$. The minimum number of balls of radius r required to cover a k-dimensional sphere of radius R is known as the "covering number" and is proportional to $(R/r)^k = p^k$. The cycle created is then a part of a connected component of $\mathcal{C}(n,r)$ containing at least $C \times p^k$ vertices. Intuitively, creating a cycle with the same birth and death times in any other way (i.e., not necessarily around a sphere) will require coverage of an area larger than the k-dimensional sphere, and therefore larger connected components. To make this statement precise, in Section 6 we present an isoperimetric-type inequality for k-cycles. Note that this statement is completely deterministic (i.e., nonrandom).

The following lemma bounds the number of vertices in a connected component of the Čech complex $\mathcal{C}(n,r)$, for small values of r.

LEMMA 4.2. *Let $\alpha > 0$ be fixed. There exists a constant $C_2 > 0$ depending only on α and d such that if*

$$nr^d \le \frac{C_2}{(\log n)^\alpha}$$

and

$$m \ge \alpha^{-1} \frac{\log n}{\log\log n},$$

then with high probability $\mathcal{C}(n,r)$ has no connected components with more than m vertices.

PROOF OF LEMMA 4.2. Let $N_m(r)$ be the number of subsets of $\mathcal{P}_n$ with m vertices that are connected in $\mathcal{C}(n,r)$. We can write $N_m(r)$ as

$$\sum_{\mathcal{Y}\subset\mathcal{P}_n} \mathbb{1}\{\mathcal{C}(\mathcal{Y},r) \text{ connected}\},$$

where the sum is over all sets $\mathcal{Y}$ of m vertices. We will show that choosing r and m as the lemma states, we have $\mathbb{P}(N_m(r) > 0) \to 0$ which implies the statement of the lemma.

By Palm theory (see, e.g., Theorem 1.6 of [44]) we have that

$$\mathbb{E}\{N_m(r)\} = \frac{n^m}{m!}\mathbb{P}\bigl(\mathcal{C}(\{X_1,\dots,X_m\},r) \text{ is connected}\bigr),$$

where $X_i \sim U([0,1]^d)$ are i.i.d. variables. If $\mathcal{C}(\{X_1,\dots,X_m\},r)$ is connected, then the underlying graph must contain a subgraph isomorphic to a tree on m vertices. Suppose that Γ is a labelled tree on the vertices $\{1,\dots,m\}$. Assuming that vertex 1 is the root, for $2 \le i \le m$ let $\mathrm{par}(i)$ be the parent of vertex i in the tree. Suppose also that the vertices are ordered so that $\mathrm{par}(i) < i$. If $\mathcal{C}(\{X_1,\dots,X_m\},r)$ contains Γ then every X_i must be connected to $X_{\mathrm{par}(i)}$ which implies that $X_i \in B_{2r}(X_{\mathrm{par}(i)})$. Therefore,

$$\begin{aligned}\mathbb{P}\bigl(\mathcal{C}(\{X_1,\dots,X_m\},r) \text{ contains } \Gamma\bigr) &\le \mathbb{P}\bigl(X_i \in B_{2r}(X_{\mathrm{par}(i)}), \forall 2 \le i \le m\bigr)\\ &\le \int_{[0,1]^d}\int_{B_{2r}(x_{\mathrm{par}(2)})}\cdots\int_{B_{2r}(x_{\mathrm{par}(m)})} dx_m\cdots dx_1\\ &= \bigl(\omega_d 2^d r^d\bigr)^{m-1}.\end{aligned}$$

The second inequality is due to the effect of the boundary of cube. The same bound holds for any ordering of the vertices. It is known that the total number of labelled trees on m vertices is m^{m-2} and, therefore, we have

$$\mathbb{E}\{N_m(r)\} \le \frac{n^m}{m!} m^{m-2}\bigl(\omega_d 2^d r^d\bigr)^{(m-1)}.$$

From Stirling's approximation, we have that $m! \ge (m/e)^m$ and, therefore,

$$\mathbb{E}\{N_m(r)\} \le n^m e^m m^{-2}\bigl(\omega_d 2^d r^d\bigr)^{(m-1)} = e\frac{n}{m^2}\bigl(e\omega_d 2^d n r^d\bigr)^{m-1}.$$

Defining $C_2 = \frac{1}{2}(e\omega_d 2^d)^{-1}$, if $nr^d \le C_2(\log n)^{-\alpha}$ then

$$\mathbb{E}\{N_m(r)\} \le e\frac{n}{m^2} e^{-(m-1)(\alpha\log\log n + \log 2)}.$$

If $m \geq \alpha^{-1} \frac{\log n}{\log \log n}$, we therefore have (for n large enough):

$$\mathbb{E}\{N_m(r)\} \leq \frac{e}{m^2},$$

and $e/m^2 \to 0$ as $n \to \infty$.

Finally, by Markov's inequality, $\mathbb{P}(N_m(r) > 0) \leq \mathbb{E}\{N_m(r)\}$, and therefore we have that $\mathbb{P}(N_m(r) > 0) \to 0$ which completes the proof. □

With these two lemmas, we can prove the upper bound in Theorem 3.1.

PROOF OF THEOREM 3.1—UPPER BOUND. Fix a value $\alpha > 0$, and consider two kinds of k-cycles: The *early-born* cycles are the ones created at a radius r satisfying $nr^d \leq C_2(\log n)^{-\alpha}$ (see Lemma 4.2). The *late-born* cycles are all the rest.

If $\gamma \in \mathrm{PH}_k(n)$ is an early-born cycle, then according to Lemma 4.2 it is part of a connected component with $m < \alpha^{-1} \frac{\log n}{\log \log n}$ vertices. If $\pi(\gamma) = p$, then from Lemma 4.1 we have that $C_1 p^k \leq m$. Combining these two statements, we have that with high probability,

$$\pi(\gamma) \leq (C_1\alpha)^{-1/k} \left(\frac{\log n}{\log \log n}\right)^{1/k}.$$

Therefore, $\pi(\gamma) \leq B_k \Delta_k(n)$, with $B_k = (C_1\alpha)^{-1/k}$.

Suppose now that $\gamma \in \mathrm{PH}_k(n)$ is a late-born cycle. This implies that $\gamma_{\mathrm{birth}} = r$ where $nr^d > (\log n)^{-\alpha}$, or in other words that $\gamma_{\mathrm{birth}} > (\frac{1}{n(\log n)^\alpha})^{1/d}$. Next, in [38] it is shown (see Theorem 6.1) that there exists $C > 0$ such that if $r \geq C(\frac{\log n}{n})^{1/d}$ then with high probability $\mathcal{C}(n, r)$ is contractible (i.e., can be "shrunk" to a point and, therefore, has no nontrivial cycles). In particular, this implies that $\gamma_{\mathrm{death}} \leq C(\frac{\log n}{n})^{1/d}$ for every cycle γ. Thus, for late-born cycles γ

$$\pi(\gamma) < C(\log n)^{(1+\alpha)/d}.$$

Thus, for any $\alpha < d/k - 1$, we have that with high probability the persistence of late-born cycles γ satisfies

$$\pi(\gamma) = o\left(\left(\frac{\log n}{\log \log n}\right)^{1/k}\right).$$

□

5. Proof—lower bound. In this section, we prove the lower bound part of Theorem 3.1 for the Čech complex $\mathcal{C}(n, r)$, namely that there exists $A_k > 0$ (depending on k and d), such that with high probability,

$$\Pi_k(n) \geq A_k \Delta_k(n) = A_k \left(\frac{\log n}{\log \log n}\right)^{1/k}.$$

In other words, we need to show that with a high probability there exists $\gamma \in \mathrm{PH}_k(n)$ with $\pi(\gamma) \geq A_k \Delta_k(n)$.

To show that, we take the unit cube $Q = [0,1]^d$ and divide it into small cubes of side $2L$. The number of small cubes we can fit in Q denoted by M satisfies $M \geq C_3 L^{-d}$ for some $C_3 > 0$. Denoting the small cubes by $Q_1, \dots, Q_M$, we want to show that at least one of these cubes contains a large cycle. Let Q_i be one of these cubes, and think of it as centered at the origin, so that $Q_i = [-L, L]^d$. Let $\ell < L/4$, denote $\hat{L} = \lfloor L/\ell \rfloor \times \ell$, and define

$$S_i^{(1)} = [-\hat{L}/2, \hat{L}/2]^{k+1} \times [-\ell/2, \ell/2]^{d-k-1},$$
$$S_i^{(2)} = [-\hat{L}/2 + \ell, \hat{L}/2 - \ell]^{k+1} \times [-\ell/2, \ell/2]^{d-k-1},$$
$$S_i = S_i^{(1)} \backslash S_i^{(2)}.$$

In other words, S_i is a "thickened" version of the boundary of a $k+1$ dimensional cube of side $\hat{L} \approx L$ (see Figure 4).

We will show that if the balls of radius r around $\mathcal{P}_n$ cover S_i but leave most of Q_i empty then $\mathcal{C}(n,r)$ would have a k-dimensional cycle. Choosing L and ℓ properly we can make sure that this cycle has the desirable persistence. More specifically, take S_i and split it into m cubes of side ℓ, denoted by $S_{i,1}, S_{i,2}, \dots, S_{i,m}$ (see Figure 4). The number of boxes m is almost proportional to the ratio of the volumes of S_i and the $S_{i,j}$-s and, therefore, $m \leq C_4 (L/\ell)^k$ for some $C_4 > 0$. The following lemma uses the process $\mathcal{P}_n$ but is in fact nonrandom, and provides a lower-bound to the persistence of the cycles we are looking for.

LEMMA 5.1. *Suppose that for every $1 \leq j \leq m$ we have $|S_{i,j} \cap \mathcal{P}_n| = 1$, and $|Q_i \cap \mathcal{P}_n| = m$. Then there exists $\gamma \in \mathrm{PH}_k(n)$ with $\pi(\gamma) \geq \frac{1}{4\sqrt{d}} \times \frac{L}{\ell}$.*

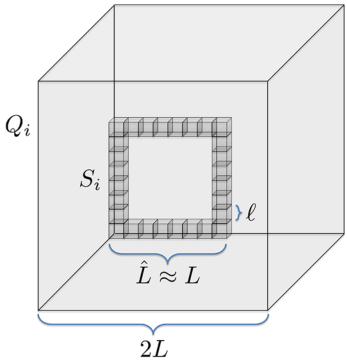


FIG. 4. *The construction we are examining to find a maximal cycle, for $d = 3$ and $k = 1$. Q_i is the big box of side $2L$, and S_i is construction made of small boxes in the middle of it, which is homotopy equivalent to a circle.*

The proof of this lemma also requires some working knowledge in algebraic topology and, therefore, we postpone it to Section 6. Intuitively, the assumptions of the lemma guarantee that for every $r \in [r_1, r_2]$, where $r_1 = \sqrt{d}\ell$ and $r_2 = L/4$, the union of balls $\mathcal{U}(\mathcal{P}_n \cap Q_i, r)$ covers S_i, and is disconnected from the rest of the balls. Therefore, its shape is "similar" to S_i and forms a nontrivial k-cycle. Since this cycle exists through the entire range $[r_1, r_2]$, its persistence is greater than $r_2/r_1 = L/4\sqrt{d}\ell$.

Following Lemma 5.1, we define the event

$$E_i = \{\forall 1 \le j \le m : |S_{i,j} \cap \mathcal{P}_n| = 1, \text{ and } |Q_i \cap \mathcal{P}_n| = m\},$$

then $E = E_1 \cup E_2 \cup \cdots \cup E_M$ is the event that at least one of the Q_i cubes contains a large cycle. Lemma 5.1 suggests that to prove there exists a large cycle it is enough to show that E occurs with high probability. We start by bounding the probability of the complement event. The next lemma shows that given the right choice of $L = L(n)$ and $\ell = \ell(n)$ we can guarantee that $E = E^{(n)}$ satisfies $\mathbb{P}(E) \to 1$.

LEMMA 5.2. *Let $n\ell^d = (\log n)^{-\alpha}$ such that $\alpha > d/k$, and let $L = \tilde{A}_k \Delta_k(n)\ell$ where $\tilde{A}_k \le (C_4\alpha)^{-1/k}$. Then*

$$\lim_{n\to\infty} \mathbb{P}(E) = 1.$$

PROOF. We start with the probability of E_i. By the spatial independence property of the Poisson process, we have

$$\mathbb{P}(E_i) = (n\ell^d)^m e^{-n(2L)^d}$$

and, therefore,

$$\mathbb{P}(E^c) = \prod_{i=1}^{M} (1 - \mathbb{P}(E_i)) = (1 - (n\ell^d)^m e^{-n(2L)^d})^M \le e^{-M(n\ell^d)^m e^{-n(2L)^d}}.$$

Thus, in order to prove that $\mathbb{P}(E) \to 1$ it is enough to show that

$$\mathcal{E} := M(n\ell^d)^m e^{-n(2L)^d} \to \infty.$$

Recall that $M \ge C_3 L^{-d}$ and that $m \le C_4(L/\ell)^k$. Assuming that $n\ell^d < 1$, we have

$$\mathcal{E} \ge C_3 L^{-d} (n\ell^d)^{C_4(L/\ell)^k} e^{-2^d nL^d} = C_3 L^{-d} e^{C_4(L/\ell)^k \log(n\ell^d) - 2^d nL^d}.$$

Now, if $n\ell^d = (\log n)^{-\alpha} < 1$ for some $\alpha > 0$ and $L = \tilde{A}_k \Delta_k(n)\ell$ for some $\tilde{A}_k > 0$, then

$$nL^d = \tilde{A}_k^d \Delta_k^d(n) \cdot n\ell^d = \tilde{A}_k^d \frac{(\log n)^{d/k-\alpha}}{(\log\log n)^{d/k}}.$$

Taking $\alpha > d/k$ yields that $nL^d \to 0$ and, therefore,

$$\mathcal{E} \geq Cn \frac{(\log\log n)^{d/k}}{(\log n)^{d/k-\alpha}} e^{-C_4 \tilde{A}_k^k \alpha \log n},$$

for some constant C. Choosing $\tilde{A}_k$ such that $C_4 \tilde{A}_k^k \alpha < 1$ we have $\mathcal{E} \to \infty$ which completes the proof. □

PROOF OF THEOREM 3.1—LOWER BOUND. From Lemma 5.2, we have that if $n\ell^d = (\log n)^{-\alpha}$ and $L/\ell = \tilde{A}_k \Delta_k(n)$ then with high probability E occurs. From Lemma 5.1, this implies that with high probability we have a "cubical" cycle γ with $\pi(\gamma) \geq \tilde{A}_k \Delta_k(n)/4\sqrt{d}$. Taking $A_k = \tilde{A}_k/4\sqrt{d}$ completes the proof. □

6. Proofs for topological lemmas.

As mentioned above, the proofs for Lemmas 4.1 and 5.1 require some working knowledge in algebraic topology. In particular, we will be making use of the definitions of chains, cycles, boundaries and induced maps in both simplicial and singular homology. For more background, see [35] or [42]. To make reading the paper fluent for readers who are less familiar with the subject, we deferred these proofs to this section. Also included in this section is the translation of Theorem 3.1 from the Čech to the Rips complex.

6.1. *Proof of Lemma 4.1.*

First, we restate the lemma.

LEMMA 4.1. *Let $\gamma \in \mathrm{PH}_k(n)$, with $\gamma_{\mathrm{birth}} = r$ and $\pi(\gamma) = p$. Then there exists a constant C_1 such that $\mathcal{C}(n, r)$ contains a connected component with at least $m = C_1 p^k$ vertices. The constant C_1 depends on k, d only.*

For the sake of simplicity, we will be using homology with coefficients in $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$. Nevertheless, Lemma 4.1 holds using coefficients over any field.

For every two spaces $S_1 \subset S_2$, we denote $i : S_1 \hookrightarrow S_2$ as the inclusion map, and the induced map in homology will be $i_* : H_*(S_1) \to H_*(S_2)$. For any finite set $\mathcal{P} \subset [0, 1]^d$ and every $r > 0$, by the Nerve Lemma 2.3 the spaces $\mathcal{C}(\mathcal{P}, r)$ and $\mathcal{U}(\mathcal{P}, r)$ are homotopy equivalent. Therefore, there are natural maps $h : \mathcal{U}(\mathcal{P}, r) \to \mathcal{C}(\mathcal{P}, r)$ and $j : \mathcal{C}(\mathcal{P}, r) \to \mathcal{U}(\mathcal{P}, r)$ such that the induced maps $h_* : H_*(\mathcal{U}(\mathcal{P}, r)) \to H_*(\mathcal{C}(\mathcal{P}, r))$ and $j_* : H_*(\mathcal{C}(\mathcal{P}, r)) \to H_*(\mathcal{U}(\mathcal{P}, r))$ are isomorphisms.

The explicit construction of j is as follows. Each vertex in $\mathcal{C}(\mathcal{P}, r)$ is sent to the center of the corresponding ball. The map is then extended to every simplex by mapping it to the convex hull of the points its vertices are mapped to. Each simplex is a convex set and it is straightforward to check that in Euclidean space, the image of each simplex lies within the union of balls $\mathcal{U}(\mathcal{P}, r)$. This way for every k-simplex $\sigma \in \mathcal{C}(\mathcal{P}, r)$ we can define its volume $\mathrm{Vol}_k(\sigma)$ to be the k-dimensional Lebesgue measure of $j(\sigma) \subset \mathbb{R}^d$.

With the volume of a simplex defined, we can now define the volume of a chain. If $\gamma \in C_k(\mathcal{C}(\mathcal{P}, r))$ is a k-chain of the form $\gamma = \sum_i \alpha_i \sigma_i$ ($\alpha_i \in \{0, 1\}$), then $\mathrm{Vol}_k(\gamma) := \sum_i \alpha_i \mathrm{Vol}_k(\sigma_i)$. In other words, the volume of a chain is defined to be the sum of the volumes of the simplexes it contains.

To prove Lemma 4.1, we will be using an isoperimetric inequality related to singular cycles in $\mathcal{U}(\mathcal{P}, r)$ (see Theorem 6.2), rather than work directly with the simplicial cycles. To try to avoid confusion, we will use γ to refer to simplicial cycles, and η for singular cycles. Recall that a singular k-simplex in $\mathbb{R}^d$ is a actually map $\sigma : \Delta^k \to \mathbb{R}^d$, where Δ^k is the standard k-simplex. For brevity, we will identify every singular simplex σ with its image $\mathrm{Im}(\sigma) \subset \mathbb{R}^d$, and every k-chain $\eta = \sum_i \alpha_i \sigma_i$ with the union $\bigcup_{i:\alpha_i \neq 0} \mathrm{Im}(\sigma_i) \subset \mathbb{R}^d$. We will also need to define the volume of a singular k-chain. Such a definition exists (cf. [32]); however, we will be looking only at chains that are of the form $\eta = j(\gamma)$ where γ is a simplicial k-chain in $\mathcal{C}(\mathcal{P}, r)$, and for those we can simply define $\mathrm{Vol}_k(\eta) := \mathrm{Vol}_k(\gamma)$.

Next, we define the *filling radius* of a singular k-cycle. Intuitively, the filling radius of a cycle measures how much we need to "inflate" the cycle to get it filled in (so it becomes trivial). Formally, we have the following.

DEFINITION 6.1. Let η be a compactly supported singular cycle in $\mathcal{U}(\mathcal{P}, r)$. A *filling* of η is a $(k + 1)$-chain in $\mathbb{R}^d$ such that $\partial\Gamma = \eta$. The *filling radius* $R_{\mathrm{fill}}(\eta)$ is defined as

$$R_{\mathrm{fill}}(\eta) = \inf\{\rho > 0 : \exists\Gamma \text{ such that } \eta = \partial\Gamma \text{ and } \Gamma \subset \mathcal{U}(\eta, \rho)\}.$$

In other words, $R_{\mathrm{fill}}(\eta)$ is the smallest ρ such that the "ρ-thickening" of η contains some filling Γ.

The workhorse of our proof of Lemma 4.1 is the following general isoperimetric inequality due to Federer and Fleming [32]. For a proof, see either the original article or Section 3 of Guth's expository notes on Gromov's systolic inequality [34].

THEOREM 6.2 (Volume to filling radius, isoperimetric inequality). *Let η be a singular k-cycle, such that* $\mathrm{Vol}_k(\eta) = V$. *Then the filling radius of η satisfies*

$$R_{\mathrm{fill}}(\eta) \leq CV^{1/k},$$

for some constant C (depending on k, d).

Recall that as in Definition 6.1, η is a k-cycle in $\mathcal{U}(\mathcal{P}, r)$. However, it is worth noting that for any k-cycle $\gamma \in \mathcal{C}(\mathcal{P}, r)$, there is a canonical inclusion into $\mathcal{U}(\mathcal{P}, r)$. This is the geometric realization of η (although it need not be embedded). Hence, this result also holds for cycles in the Čech complex.

To prove Lemma 4.1, we will thus need to take two steps: (1) bound the volume of a cycle η, and (2) bound death time of η using the filling radius $R_{\mathrm{fill}}(\eta)$. We start with the following definition.

DEFINITION 6.3. Let X be a set in $\mathbb{R}^d$. For $\varepsilon > 0$ the set S is called an ε-net of X if:

(1) $S \subseteq X$,
(2) $X \subset \mathcal{U}(S, \varepsilon)$, that is, X is covered by the balls of radius ε around S, and
(3) for every $p_1, p_2 \in S$, $\|p_1 - p_2\| \geq \varepsilon$.

In other words, an ε-net is both an ε-cover and an ε-packing.

ε-nets are a standard construction in computational geometry and exist for any metric space [24]. They can be constructed incrementally using the following algorithm: (1) Initialize S to be the empty set. (2) Select any uncovered point in X and add it to S. (3) Mark all points of distance less than ε from the selected point as covered. (4) Repeat 2–3 until there are no uncovered points.

Next, let $\mathcal{P} = \{x_1, x_2, \dots, x_m\} \subset \mathbb{R}^d$ and let $S \subset \mathcal{P}$ be an ε-net of $\mathcal{P}$. By the definition of ε-nets, the following holds:

$$\mathcal{P} \subset \mathcal{U}(S, \varepsilon), \tag{6.1}$$

$$\|p_i - p_j\| \geq \varepsilon \qquad \forall p_i, p_j \in S. \tag{6.2}$$

Using (6.1) and the triangle inequality, we also have

$$\mathcal{U}(\mathcal{P}, \varepsilon) \subset \mathcal{U}(S, 2\varepsilon) \subset \mathcal{U}(\mathcal{P}, 2\varepsilon). \tag{6.3}$$

We will use the intermediate construction $\mathcal{U}(S, 2\varepsilon)$ to bound the volume of cycles. In particular, we will need the following lemma. We use $[\cdot]$ to denote the equivalence class in homology of a corresponding cycle.

LEMMA 6.4. *Let $\mathcal{P}$ and S be as defined above, and let γ be a k-cycle in $\mathcal{C}(S, 2\varepsilon)$. Then $\mathrm{Vol}_k(\gamma) \leq C_5 m \varepsilon^k$, where C_5 depends only on k, d. Consequently, for every (singular) cycle η in $\mathcal{U}(S, 2\varepsilon)$ there exists a homologous cycle η' such that $[\eta] = [\eta']$ and such that $\mathrm{Vol}_k(\eta') \leq C_5 m \varepsilon^k$.*

PROOF. The k-dimensional volume of γ is the sum of the k-volumes of the simplexes in γ. This can be bounded by the maximal volume induced by any one simplex multiplied by the number of simplexes in γ.

To bound the number of simplexes, first observe that γ is supported on S. By (6.2), every pair of vertices $p_1, p_2 \in S$ are at distance $\|p_1 - p_2\| \geq \varepsilon$. So the balls centered at points in S of radius $\varepsilon/2$ are disjoint. This implies, by a sphere packing bound, that every vertex in S has only a bounded number of neighboring vertices in $\mathcal{C}(S, 2\varepsilon)$, namely the maximum number of disjoint balls of radius $\varepsilon/2$ that can fit in a ball of radius 4ε. This sphere packing number is clearly bounded above by 8^d, the ratio of the volumes of these spheres. This implies that every vertex is contained in at most $\binom{8^d}{k}$ k-dimensional faces and since by assumption there are at most m vertices in $\mathcal{P}$ and hence S, there are at most $m\binom{8^d}{k}$ k-dimensional faces total.

To bound the maximal volume of the single simplexes, observe that the longest edge in any simplex of γ has length at most 4ε. Therefore, for every simplex σ in γ we have $\mathrm{Vol}_k(\sigma) \le (4\varepsilon)^k$ (the volume of a cube of side 4ε).

To conclude, we have shown that γ has at most $m\binom{8^d}{k}$ simplexes, the volume of each of them is bounded by $(4\varepsilon)^k$. Therefore, $\mathrm{Vol}_k(\gamma) \le C_5 m\varepsilon^k$ where $C_5 = 4^k\binom{8^d}{k}$.

Next, let η be a cycle in $\mathcal{U}(\mathcal{S}, 2\varepsilon)$. Since the map $j_* : H_*(\mathcal{C}(S, 2\varepsilon)) \to H_*(\mathcal{U}(S, 2\varepsilon))$ is an isomorphism, we can look at the homology class $j_*^{-1}([\eta])$, and take a representative cycle γ. Defining $\eta' = j(\gamma)$ then $[\eta'] = j_* \circ j_*^{-1}([\eta]) = [\eta]$, so η and η' are homologous. In addition, since γ is a cycle in $\mathcal{C}(S, 2\varepsilon)$ and $\eta' = j(\gamma)$, we have that $\mathrm{Vol}_k(\eta') = \mathrm{Vol}_k(\gamma) \le C_5 m\varepsilon^k$. That completes the proof. □

For the next lemma, consider the following sequence of maps in homology (induced by inclusion maps):

$$H_k\bigl(\mathcal{U}(\mathcal{P}, \varepsilon)\bigr) \xrightarrow{i_*} H_k\bigl(\mathcal{U}(S, 2\varepsilon)\bigr) \xrightarrow{i_*} H_k\bigl(\mathcal{U}(\mathcal{P}, 2\varepsilon)\bigr).$$

LEMMA 6.5. *Vertices to volume] Let $\mathcal{P} = \{x_1, x_2, \dots, x_m\} \subset \mathbb{R}^d$. Suppose that η is an arbitrary k-cycle in $\mathcal{U}(\mathcal{P}, \varepsilon)$, and let $i \circ i(\eta)$ be its image in $\mathcal{U}(\mathcal{P}, 2\varepsilon)$. Then there exists a k-cycle η' in $\mathcal{U}(\mathcal{P}, 2\varepsilon)$, homologous to $i \circ i(\eta)$, such that* $\mathrm{Vol}_k(\eta') \le C_5 m\varepsilon^k$ *for some constant $C_5 > 0$ depending only on k and d.*

PROOF. Let $i(\eta)$ be the inclusion of η into $\mathcal{U}(S, 2\varepsilon)$. From Lemma 6.4, we have that there exists a cycle η'' in $\mathcal{U}(S, 2\varepsilon)$ such that $[\eta''] = [i(\eta)]$ and such that $\mathrm{Vol}_k(\eta'') \le C_5 m\varepsilon^k$. Defining $\eta' = i(\eta'')$, then $[\eta'] = i_*([\eta'']) = i_*([i(\eta)]) = [i \circ i(\eta)]$, and since the inclusion does not change the volume we have $\mathrm{Vol}_k(\eta') = \mathrm{Vol}_k(\eta'') \le C_5 m\varepsilon^k$. That completes the proof. □

Finally, we relate the filling radius to the persistence of the cycles.

LEMMA 6.6 (Filling radius to persistence). *If η is a cycle in $\mathcal{U}(\mathcal{P}, r)$, with a filling radius $R_{\mathrm{fill}}(\eta) = R$, then $\eta_{\mathrm{death}} \le R + r$.*

PROOF. Since η is a cycle in $\mathcal{U}(\mathcal{P}, r)$, then by the triangle inequality we have that $\mathcal{U}(\eta, R) \subset \mathcal{U}(\mathcal{P}, r + R)$. By the definition of R_{fill} (see Definition 6.3), this implies that there exists a $(k+1)$ cycle Γ in $\mathcal{U}(\mathcal{P}, R+r)$ such that $\eta = \partial\Gamma$. Therefore, in $\mathcal{U}(\mathcal{P}, R+r)$ the cycle η is already trivial which implies that $\eta_{\mathrm{death}} \le R + r$. □

We are now ready to prove Lemma 4.1.

PROOF OF LEMMA 4.1. Let $\gamma \in \mathrm{PH}_k(n)$ with $\gamma_{\mathrm{birth}} = r$. Suppose that the simplexes constructing γ are contained in a connected component with m vertices of $\mathcal{C}(n,r) = \mathcal{C}(\mathcal{P}_n, r)$. Let $\mathcal{P} \subset \mathcal{P}_n$ be the set of vertices in this connected component, then necessarily γ is also a cycle in $\mathcal{C}(\mathcal{P}, r)$.

Next, take the corresponding cycle $\eta = j(\gamma)$ in $\mathcal{U}(\mathcal{P}, r)$. According to Lemma 6.5, there exists a cycle η' in $\mathcal{U}(\mathcal{P}, 2r)$, homologous to $i \circ i(\eta)$, such that $\mathrm{Vol}_k(\eta') \le C_5 m r^k$, and from Theorem 6.2 this implies that $R_{\mathrm{fill}}(\eta') \le C(C_5 m r^k)^{1/k} = C' m^{1/k} r$. Using Lemma 6.6, we then have that $\eta'_{\mathrm{death}} \le r(C' m^{1/k} + 2)$. Since η' and $i \circ i(\eta)$ are homologous, then η and η' share the same death time, which in turn implies that γ and η' share the same death time. Therefore, $\pi(\gamma) \le C' m^{1/k} + 2 \le C'' m^{1/k}$. In other words, if $\pi(\gamma) = p$ then we have that $p^k \le m(C'')^k$. Taking $C_1 = 1/(C'')^k$ completes the proof. □

6.2. *Proof of Lemma* 5.1. We first restate the lemma.

LEMMA 5.1. *Suppose that for every* $1 \le j \le m$ *we have* $|S_{i,j} \cap \mathcal{P}_n| = 1$, *and* $|Q_i \cap \mathcal{P}_n| = m$. *Then there exists* $\gamma \in \mathrm{PH}_k(n)$ *with* $\pi(\gamma) \ge \frac{1}{4\sqrt{d}} \times \frac{L}{\ell}$.

PROOF. Let $r_1 = \sqrt{d}\ell$ and $r_2 = L/4$. The assumptions that $|S_{i,j} \cap \mathcal{P}_n| = 1$ for every $1 \le i \le m$ and $|Q_i \cap \mathcal{P}_n| = m$ assure that:

- For every $r \ge r_1$ the set $\mathcal{U}(\mathcal{P}_n \cap Q_i, r)$ is connected and covers S_i.
- For every $r \le r_2$ the sets $\mathcal{U}(\mathcal{P}_n \cap Q_i, r)$ and $\mathcal{U}(\mathcal{P}_n \backslash Q_i, r)$ are disjoint.

In other words, for every $r \in [r_1, r_2]$ the set $\mathcal{U}(\mathcal{P}_n \cap Q_i, r)$ is a connected component of $\mathcal{U}(n, r)$. We will show that this component contains the desired cycle.

Defining $S_i^{(r)} = \mathcal{U}(S_i, r)$, for every $r \in [r_1, r_2]$ we have

$$S_i \subset \mathcal{U}(\mathcal{P}_n \cap Q_i, r) \subset S_i^{(r)}.$$

In addition, for every $r \in [r_1, r_2]$, the inclusion $S_i \hookrightarrow S_i^{(r)}$ is a homotopy equivalence and both spaces are homotopy equivalent to a k-dimensional sphere, and in particular have a nontrivial k-cycle. A standard argument in algebraic topology (using the induced maps in homology) yields that $\mathcal{U}(\mathcal{P}_n \cap Q_i, r)$ must have a nontrivial k-cycle as well. Intuitively, since the k-cycle in S_i "survives" the inclusion into $S_i^{(r)}$, it must also be present in the intermediate set $\mathcal{U}(\mathcal{P}_n \cap Q_i, r)$. Now consider the following sequence induced by the inclusion maps

$$H_k(S_i) \xrightarrow{i_*} H_k(\mathcal{U}(\mathcal{P}_n \cap Q_i, r_1)) \xrightarrow{i_*} H_k(\mathcal{U}(\mathcal{P}_n \cap Q_i, r_2)) \xrightarrow{i_*} H_k(S_i^{(r_2)}).$$

Let η be a nontrivial cycle in S_i, then $i_* \circ i_* \circ i_*([\eta]) \neq 0$ since by assumption $i_* \circ i_* \circ i_*(\eta)$ is a nontrivial cycle in $S_i^{(r_2)}$ as well. Consequently, we must have $i_*([\eta]) \neq 0$ and $i_* \circ i_*([\eta]) \neq 0$. Next, define $\gamma = h \circ i(\eta)$—a cycle in $\mathcal{C}(\mathcal{P}_n, r_1)$, then γ is nontrivial and so does $i(\gamma)$ in $\mathcal{C}(\mathcal{P}_n, r_2)$. Therefore, $\gamma_{\text{birth}} \le r_1$ and $\gamma_{\text{death}} \ge r_2$, and then

$$\pi(\gamma) = \frac{\gamma_{\text{death}}}{\gamma_{\text{birth}}} \ge \frac{r_2}{r_1} = \frac{1}{4\sqrt{d}} \times \frac{L}{\ell},$$

this completes the proof. □

6.3. *Proof of Theorem 3.1 for the Vietoris–rips filtration.*

PROOF. Let $r_2 > 2r_1$, and consider the following sequences of maps induced by the inclusions in (2.2):

$$H_k\big(\mathcal{C}(n, r_1)\big) \xrightarrow{i_*} H_k\big(\mathcal{R}(n, r_1)\big) \xrightarrow{i_*} H_k\big(\mathcal{R}(n, r_2/2)\big) \xrightarrow{i_*} H_k\big(\mathcal{C}(n, r_2)\big).$$

Suppose there exists a cycle γ in $\mathcal{C}(n, r_1)$ with $\gamma_{\text{death}} \ge r_2$. Then necessarily $i_* \circ i_* \circ i_*([\gamma]) \ne 0$, which implies that both $i_*([\gamma]) \ne 0$ and $i_* \circ i_*([\gamma]) \ne 0$. Therefore, there exists a nontrivial cycle $\gamma' = i(\gamma)$ in $\mathcal{R}(n, r_1)$ such that $\gamma'_{\text{death}} \ge r_2/2$, and consequently $\pi(\gamma') \ge r_2/2r_1$. Thus,

$$\mathbb{P}\big(\Pi_k^{\mathcal{C}}(n) \ge A_k \Delta_k(n)\big) \le \mathbb{P}\big(\Pi_k^{\mathcal{R}}(n) \ge A_k \Delta_k(n)/2\big). \tag{6.4}$$

On the other hand, we can look at the following sequence for $r_2 > 2r_1$:

$$H_k\big(\mathcal{R}(n, r_1)\big) \xrightarrow{i_*} H_k\big(\mathcal{C}(n, 2r_1)\big) \xrightarrow{i_*} H_k\big(\mathcal{C}(n, r_2)\big) \xrightarrow{i_*} H_k\big(\mathcal{R}(n, r_2)\big).$$

Suppose that there exists a cycle γ in the Rips filtration with $\gamma_{\text{birth}} \le r_1$ and $\gamma_{\text{death}} \ge r_2$. Then there exists a cycle γ' in the Čech filtration with $\gamma'_{\text{birth}} \le 2r_1$ and $\gamma'_{\text{death}} \ge r_2$, and therefore, $\pi(\gamma') \ge r_2/2r_1$. Thus,

$$\mathbb{P}\big(\Pi_k^{\mathcal{C}}(n) \le B_k \Delta_k(n)\big) \le \mathbb{P}\big(\Pi_k^{\mathcal{R}}(n) \le 2B_k \Delta_k(n)\big). \tag{6.5}$$

To conclude, we have that

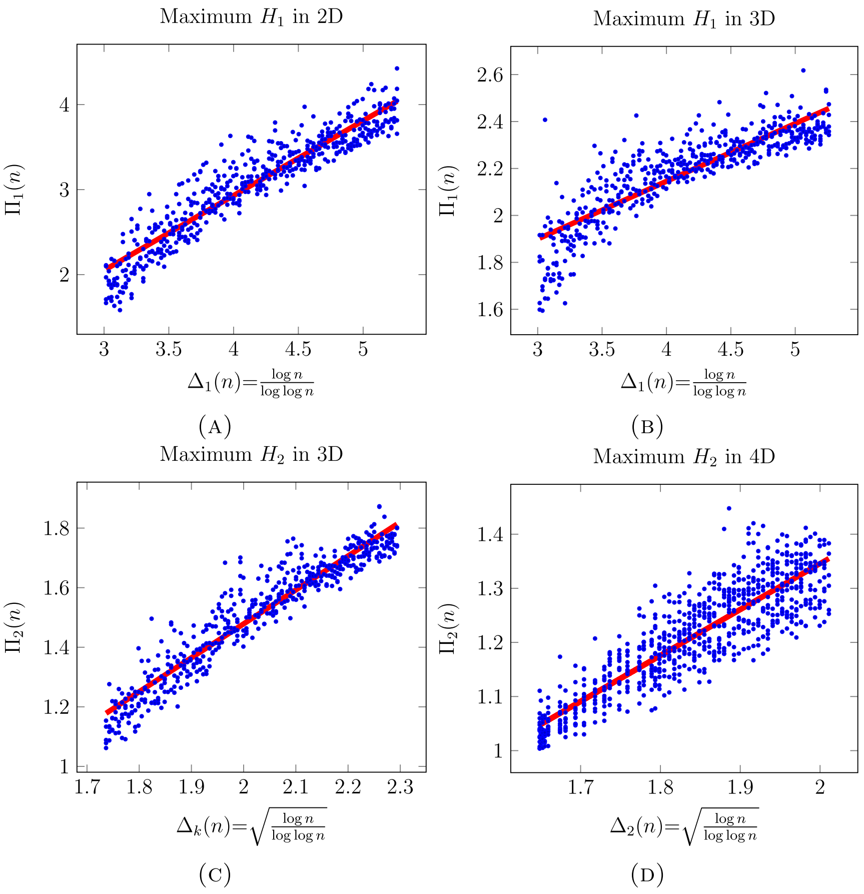
$$\mathbb{P}\left(A_k \le \frac{\Pi_k^{\mathcal{C}}(n)}{\Delta_k(n)} \le B_k\right) \le \mathbb{P}\left(A_k/2 \le \frac{\Pi_k^{\mathcal{R}}(n)}{\Delta_k(n)} \le 2B_k\right).$$

Since the left-hand side converges to 1, so does the right-hand side, which completes the proof. □

7. Numerical experiments. In this section, we present numerical simulations demonstrating the behavior of $\Pi_k(n)$ for the Čech complex in dimensions $d = 2, 3$ and 4. The experiments were run by generating a Poisson process with rate n in the unit cube of the appropriate dimension. To generate randomness, we used the standard implementation of the Mersenne Twister [53]. The persistence diagram computation was done using the PHAT library [5].

For each sample, the Čech complex is built until the point of coverage (or very near coverage), since past coverage the complex is contractible and there are no changes in homology. In dimensions 2 and 3, instead of the Čech filrtration, we use the α-shape filtration [28] which is based on the Delaunay triangulation. To compute the triangulations, we used the CGAL library [47]. The key benefit of this construction is that the simplicial complex is of a smaller size, for example, in 2 dimensions the size of the Delaunay triangulation is at most quadratic in the number of points. The persistence diagram are the same since for any parameter r, the α-complex and Čech complex are homotopy equivalent (see [27]), giving rise to isomorphic homology groups.

The results are shown in Figure 5. The number of points was varied from 100 to 1,000,000 (in higher dimensions, this was reduced due to computational complex-



(A)

(B)

(C)

(D)

FIG. 5. *Plots of maximum persistence for the Čech filtration, against the proper scaling term* $\Delta_k(n)$. *We tested different dimensions for the homology and for the ambient space.* (A) H_1 *in* $\mathbb{R}^2$, (B) H_1 *in* $\mathbb{R}^3$, (C) H_2 *in* $\mathbb{R}^3$ (D) H_2 *in* $\mathbb{R}^4$. *Each point is the result of a different trial, and the red line represents the best linear fit. For* (A), (B) *and* (C) *the range of points is* $n = 10^2$ *to* 10^6. *For* (D), *the range is roughly* $n = 10^2$ *to* 10^4. *The reduced range is a consequence of computational considerations—the number of simplices grows quickly as the dimension increases.*

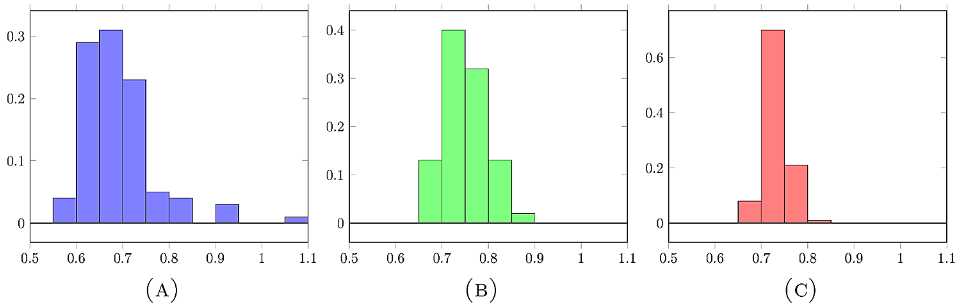


FIG. 6. *Histograms of empirical* $\Pi_1(n)$ *in 2D normalized by* $\frac{\log n}{\log\log n}$ *for* (A) 400 *points* (B) 2000 *points* (C) 2×10^6 *points*.

ity). We tested the behavior of $\Pi_k(n)$ for a few values of k, and d (the ambient dimension). For $d = 2$, the only interesting case is $k = 1$, namely H_1 (A). The resulting plot shows the maximal persistence $\Pi_1(n)$ against $\Delta_1(n) = \log n / \log\log n$. For each value of n, we repeated the experiment several times. Here, we also plot the best linear fit with the constant 0.88. We also show the results for H_1 when $d = 3$ (B), H_2 when $d = 3$ (C), and H_2 when $d = 4$ (D). We note that we performed a the same tests for the Rips filtration and the results were the same (but with different slopes).

There are two particularities in these plots—the first is that the spread is large for any one value of n. While it follows the straight line well, it does not seem to converge to a single value. However, the resulting distributions do seem to converge, albeit slowly, as can be seen in Figure 6. The histograms (A), (B) and (C) present the resulting ratio for 400, 2000 and 2,000,000 points, respectively. While there is a deviation, the distribution does become more concentrated around its peak.

The second issue is is that at smaller n, the maximum value drops off faster than linearly. This can be seen particularly in of Figure 5(B). This phenomenon could be explained by saying that n is simply not large enough for the limiting behavior to apply. Nevertheless, we tried to investigate this issue further, by considering the Čech complex on the flat torus ($\mathbb{T}^2$) by identifying the edges of the unit square. This part was computed using the periodic triangulations provided in CGAL [47]. We generated points in the unit square and then computed the maximal persistence using the Euclidean metric (e.g., the standard case) and using the metric on the flat torus. This was repeated 100 times for each value of n. We computed the mean and standard deviation for each value and show the results in Figure 7. The red line shows the mean for the unit square. The red shaded region showing the interval of the mean +/− the standard deviation. The blue line (and the blue region) are the mean (and standard deviation) for the maximal persistence on the flat torus. The purple region is region where the blue and red regions overlap. As can be seen, for most n the maximal persistence is identical, indicated that the longest lived cycles

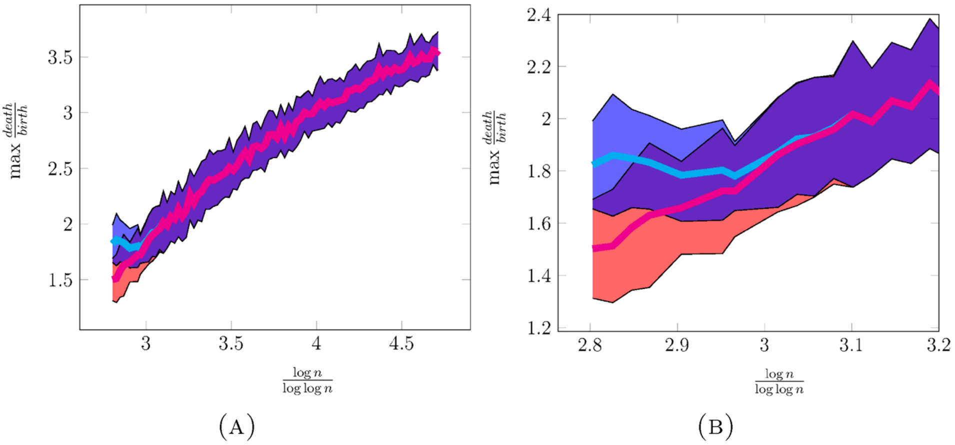


FIG. 7. *The effect of boundaries is larger for a small number of points. The plot shows the mean maximum persistence for H_1 as a function of* $\log n/\log\log n$ *with the shaded region showing interval corresponding to* $+/-$ *the standard deviation. The red line* (*and the red shaded region*) *shows the maximum persistence in the unit square, while the blue line shows the maximum persistence for the same point set in the flat torus. The purple region shows that for most values of* n, *the value of maximal persistence is the same in both cases. This is illustrated by an equal mean as well as the overlapping shaded regions* (*shown as purple*). *In* (A), *we see the plot up to several thousand points, while in* (B) *we show a close-up for small values of* n, *where the results differ.*

did not occur near the boundary. The difference is only visible for small values of n (where there are only a few points). At low values of n, the results on the torus demonstrate a more linear behavior. This provides strong evidence that the nonlinearity is due to boundary effects.

For the case of the flat torus, there are two essential one-dimensional homology classes (cycles with infinite persistence) corresponding to the generators of the torus. For the above results, we ignore the essential classes.

8. Conclusion. In this paper, we examined the maximum persistence of cycles in the persistent homology of either the random Čech or Rips complexes, generated by a homogeneous Poisson process in the unit cube. We showed that with a high probability we have $\Pi_k(n) \sim (\frac{\log n}{\log\log n})^{1/k}$. This paper proves that upper and lower bounds exist, and it remains future work to prove stronger limiting theorems such as a law of large numbers or a central limit theorem for $\Pi_k(n)$.

We note that while we focused on the Poisson process on the cube for simplicity, similar results can be proved with minor adjustments for nonhomogeneous Poisson processes as well, and for many compact spaces other than the cube (e.g., compact Riemannian manifolds). The scale of the maximum persistence should be the same ($\Delta_k(n)$), but the exact constants will be different. An important observation in this case is that $\Pi_k(n)$ should be defined as the maximum persistence among all the "small" cycles, that is, ignoring the cycles that belong to the homology of the underlying space. Recall that these small cycles are considered the noise in various TDA applications. Thus, revealing their distribution would be an important first step in performing noise filtering or reduction. At this point, we would like to offer the following insight related to the "signal to noise ratio" (SNR), in this kind of topological inference problems.

Suppose that the samples are generated from a distribution on a compact manifold $\mathcal{M}$, and our interest is in recovering its homology $H_k(\mathcal{M})$. The cycles that belong to the homology of $\mathcal{M}$ will show up in the Čech complex at some radius, and we can denote by $\Pi_k^{\mathcal{M}}(n)$ the *minimal* persistence of these cycles (in the Čech filtration). One question we might ask is—how do the signal and the noise compare? In other words—what can we say about $\Pi_k^{\mathcal{M}}(n)/\Pi_k(n)$?

The analysis we have so far already offers a preliminary answer to this question. For every cycle γ that belongs to the homology of $\mathcal{M}$, we know two things: (a) γ_{death} is approximately constant (depending on the geometry of $\mathcal{M}$), and (b) $\gamma_{\mathrm{birth}} \le C(\frac{\log n}{n})^{1/d}$ (since there are no more changes in homology past coverage, see Theorem 4.9 in [9]). Therefore, we can conclude that

$$\Pi_k^{\mathcal{M}}(n) \ge C'\left(\frac{n}{\log n}\right)^{1/d}.$$

Combining this bound with our bound for $\Pi_k(n)$, we have, for example, that for any $\epsilon > 0$,

$$\frac{\Pi_k^{\mathcal{M}}(n)}{\Pi_k(n)} \ge n^{1/d-\epsilon}.$$

To get a better estimate for this ratio, we will need to refine our results for $\Pi_k(n)$, as well as to make more precise statements about the birth times of cycles that belong to $\mathcal{M}$ (instead of using a crude upper bound).

To conclude, we believe that the results in this paper are a promising lead in the direction of noise filtering for topological inference, and will be very useful for future analysis of probabilistic models in TDA.

Acknowledgments. All three authors acknowledge the Institute for Mathematics and its Applications, where some of this work was completed. The authors would like to thank Larry Guth for helpful conversations about isoperimetry, and to Robert Adler and Sayan Mukherjee for useful comments and fruitful discussions.

REFERENCES

[1] Adler, R. J., Bobrowski, O. and Weinberger, S. (2014). Crackle: The homology of noise. *Discrete Comput. Geom.* **52** 680–704. MR3279544

[2] Appel, M. J. B. and Russo, R. P. (2002). The connectivity of a graph on uniform points on $[0,1]^d$. *Statist. Probab. Lett.* **60** 351–357. MR1947174

[3] Aronshtam, L. and Linial, N. (2015). When does the top homology of a random simplicial complex vanish? *Random Structures Algorithms* **46** 26–35. MR3291292

[4] BALOGH, J., GONZÁLEZ-AGUILAR, H. and SALAZAR, G. (2013). Large convex holes in random point sets. *Comput*. *Geom*. **46** 725–733. MR3030663

[5] BAUER, U., KERBER, M. and REININGHAUS, J. (2014). PHAT (Persistent Homology Algorithm Toolbox). Online; accessed, 7-May-2015.

[6] BOBROWSKI, O. and ADLER, R. J. (2014). Distance functions, critical points, and the topology of random čech complexes. *Homology*, *Homotopy Appl*. **16** 311–344. MR3280987

[7] BOBROWSKI, O. and BORMAN, M. S. (2012). Euler integration of Gaussian random fields and persistent homology. *J*. *Topol*. *Anal*. **4** 49–70. MR2914873

[8] BOBROWSKI, O. and KAHLE, M. (2016). Topology of random geometric complexes: A survey. Topology in Statistical Inference, the Proceedings of Symposia in Applied Mathematics. To appear.

[9] BOBROWSKI, O. and MUKHERJEE, S. (2015). The topology of probability distributions on manifolds. *Probab*. *Theory Related Fields* **161** 651–686. MR3334278

[10] BOBROWSKI, O., MUKHERJEE, S. and TAYLOR, J. E. (2017). Topological consistency via kernel estimation. *Bernoulli* **23** 288–328. MR3556774

[11] BOBROWSKI, O. and WEINBERGER, S. (2015). On the vanishing of homology in random Čech complexes. Preprint. Available at arXiv:1507.06945.

[12] BOLLOBÁS, B. and RIORDAN, O. M. (2003). Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks* 1–34. Wiley, Weinheim. MR2016117

[13] BORSUK, K. (1948). On the imbedding of systems of compacta in simplicial complexes. *Fund*. *Math*. **35** 217–234. MR0028019

[14] BUBENIK, P. and KIM, P. T. (2007). A statistical approach to persistent homology. *Homology*, *Homotopy Appl*. **9** 337–362. MR2366953

[15] BUCHET, M., CHAZAL, F., OUDOT, S. Y. and SHEEHY, D. R. (2015). Efficient and robust persistent homology for measures. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms* 168–180. SIAM, Philadelphia, PA. MR3451037

[16] CARLSSON, G. (2009). Topology and data. *Bull*. *Amer*. *Math*. *Soc*. (*N*.*S*.) **46** 255–308. MR2476414

[17] CARLSSON, G., ISHKHANOV, T., DE SILVA, V. and ZOMORODIAN, A. (2008). On the local behavior of spaces of natural images. *Int*. *J*. *Comput*. *Vis*. **76** 1–12.

[18] CHAZAL, F., DE SILVA, V. and OUDOT, S. (2014). Persistence stability for geometric complexes. *Geom*. *Dedicata* **173** 193–214. MR3275299

[19] CHAZAL, F., FASY, B. T., LECCI, F., RINALDO, A. and WASSERMAN, L. (2014). Stochastic convergence of persistence landscapes and silhouettes. In *Computational Geometry* (*SoCG*'14) 474–483. ACM, New York. MR3382329

[20] CHAZAL, F., GLISSE, M., LABRUÈRE, C. and MICHEL, B. (2015). Convergence rates for persistence diagram estimation in topological data analysis. *J*. *Mach*. *Learn*. *Res*. **16** 3603–3635. MR3450548

[21] CHAZAL, F., GUIBAS, L. J., OUDOT, S. Y. and SKRABA, P. (2011). Scalar field analysis over point cloud data. *Discrete Comput*. *Geom*. **46** 743–775. MR2846177

[22] CHAZAL, F., GUIBAS, L. J., OUDOT, S. Y. and SKRABA, P. (2013). Persistence-based clustering in Riemannian manifolds. *J*. *ACM* **60** Art. 41, 38. MR3144911

[23] CHAZAL, F., TERESE FASY, B., LECCI, F., RINALDO, A., SINGH, A. and WASSERMAN, L. (2013). On the bootstrap for persistence diagrams and landscapes. Preprint. Available at arXiv:1311.0376.

[24] CLARKSON, K. L. (2006). Nearest-neighbor searching and metric space dimensions. *Nearest-neighbor Methods for Learning and Vision*: *Theory and Practice* 15–59.

[25] DE SILVA, V. and GHRIST, R. (2007). Coverage in sensor networks via persistent homology. *Algebr*. *Geom*. *Topol*. **7** 339–358. MR2308949

[26] DEY, T. K., FAN, F. and WANG, Y. (2015). Graph induced complex on point data. *Comput*. *Geom*. **48** 575–588. MR3350802

[27] EDELSBRUNNER, H. (1993). The union of balls and its dual shape. In *Proceedings of the Ninth Annual Symposium on Computational Geometry* 218–231. ACM, New York.

[28] EDELSBRUNNER, H. (1995). The union of balls and its dual shape. *Discrete Comput. Geom.* **13** 415–440. MR1318786

[29] ERDŐS, P. and RÉNYI, A. (1959). On random graphs. *Publ. Math. Debrecen* **6** 290–297.

[30] ERDŐS, P. and RÉNYI, A. (1961). On the evolution of random graphs. *Bull. Inst. Internat. Statist.* **38** 343–347. MR0148055

[31] FASY, B. T., LECCI, F., RINALDO, A., WASSERMAN, L., BALAKRISHNAN, S. and SINGH, A. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42** 2301–2339. MR3269981

[32] FEDERER, H. and FLEMING, W. H. (1960). Normal and integral currents. *Ann. of Math.* (2) **72** 458–520. MR0123260

[33] GHRIST, R. (2008). Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.* (*N.S.*) **45** 61–75. MR2358377

[34] GUTH, L. (2006). Notes on Gromov's systolic estimate. *Geom. Dedicata* **123** 113–129. MR2299729

[35] HATCHER, A. (2002). *Algebraic Topology*. Cambridge Univ. Press, Cambridge, Cambridge. MR1867354

[36] HUDSON, B., MILLER, G. L., OUDOT, S. Y. and SHEEHY, D. R. (2009). Mesh enhanced persistent homology.

[37] KAHLE, M. (2009). Topology of random clique complexes. *Discrete Math.* **309** 1658–1671. MR2510573

[38] KAHLE, M. (2011). Random geometric complexes. *Discrete Comput. Geom.* **45** 553–573. MR2770552

[39] KAHLE, M. (2014). Sharp vanishing thresholds for cohomology of random flag complexes. *Ann. of Math.* (2) **179** 1085–1107. MR3171759

[40] KAHLE, M. and MECKES, E. (2013). Limit theorems for Betti numbers of random simplicial complexes. *Homology, Homotopy Appl.* **15** 343–374. MR3079211

[41] LINIAL, N. and MESHULAM, R. (2006). Homological connectivity of random 2-complexes. *Combinatorica* **26** 475–487. MR2260850

[42] MUNKRES, J. R. (1984). *Elements of Algebraic Topology*. Addison-Wesley, Menlo Park, CA. MR0755006

[43] OWADA, T. and ADLER, R. J. (2015). Limit theorems for point processes under geometric constraints (and topological crackle). Preprint. Available at arXiv:1503.08416.

[44] PENROSE, M. (2003). *Random Geometric Graphs*. *Oxford Studies in Probability* **5**. Oxford Univ. Press, Oxford. MR1986198

[45] PENROSE, M. D. (1997). The longest edge of the random minimal spanning tree. *Ann. Appl. Probab.* **7** 340–361. MR1442317

[46] PHILLIPS, J. M., WANG, B. and ZHENG, Y. (2013). Geometric inference on kernel density estimates. Preprint. Available at arXiv:1307.7760.

[47] THE CGAL PROJECT (2015). *CGAL User and Reference Manual. CGAL Editorial Board, 4.6 edition.*

[48] SHEEHY, D. R. (2013). Linear-size approximations to the Vietoris-Rips filtration. *Discrete Comput. Geom.* **49** 778–796. MR3068574

[49] SHEEHY, D. R. (2014). The persistent homology of distance functions under random projection. In *Computational Geometry* (*SoCG*'14) 328–334. ACM, New York. MR3382313

[50] YOGESHWARAN, D. and ADLER, R. J. (2015). On the topology of random complexes built over stationary point processes. *Ann. Appl. Probab.* **25** 3338–3380. MR3404638

[51] YOGESHWARAN, D., SUBAG, E. and ADLER, R. J. (2014). Random geometric complexes in the thermodynamic regime. submitted. Preprint. Available at arXiv:1403.1164.

[52] ZOMORODIAN, A. and CARLSSON, G. (2005). Computing persistent homology. *Discrete Comput*. *Geom*. **33** 249–274. MR2121296

[53] Random number generation in c++11. https://isocpp.org/files/papers/n3551.pdf.

O. BOBROWSKI
DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION–ISRAEL INSTITUTE OF TECHNOLOGY
HAIFA, 32000
ISRAEL
E-MAIL: omer@math.duke.edu

M. KAHLE
DEPARTMENT OF MATHEMATICS
THE OHIO STATE UNIVERSITY
231 W. 18TH AVE.
COLUMBUS, OHIO 43210
USA
E-MAIL: mkahle@math.osu.edu

P. SKRABA
ARTIFICIAL INTELLIGENCE LABORATORY
JOZEF STEFAN INSTITUTE
1000 LJUBLJANA
SLOVENIA
E-MAIL: primoz.skraba@ijs.si