A Higher-Dimensional Homologically Persistent Skeleton

Sara Kališnik Verovšek*, Vitaliy Kurlin†, Davorin Lešnik‡

Abstract

A data set is often given as a point cloud, i.e. a non-empty finite metric space. An important problem is to detect the topological shape of data — for example, to approximate a point cloud by a low-dimensional non-linear subspace such as a graph or a simplicial complex. Classical clustering methods and principal component analysis work very well when data points split into well-separated groups or lie near linear subspaces.

Methods from topological data analysis detect more complicated patterns such as holes and voids that persist for a long time in a 1-parameter family of shapes associated to a point cloud. These features were recently visualized in the form of a 1-dimensional homologically persistent skeleton, which optimally extends a minimal spanning tree of a point cloud to a graph with cycles. We generalize this skeleton to higher dimensions and prove its optimality among all complexes that preserve topological features of data at any scale.

*Department of Mathematics, Brown University, sara_kalisnik_verovsek@brown.edu

†Materials Innovation Factory, University of Liverpool, vkurlin@liverpool.ac.uk

‡Department of Mathematics, University of Ljubljana, davorin.lesnik@fmf.uni-lj.si (this author was partially supported by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF under Award No. FA9550-14-1-0096)

Contents

1 Introduction

1.1 Motivations and Data Skeletonization Problem

Real data is often unstructured and comes in a form of a non-empty finite metric space, called a *point cloud*. Such a point cloud can consist of points in 2D images or of high-dimensional vector descriptors of a molecule. A typical problem is to study interesting groups or clusters within data sets.

However, real data rarely splits into well-separated clusters, though it often has an intrinsic low-dimensional structure. For example, a cloud of mean-centered and normalized 3×3 patches in natural grayscale images has its 50% densest points distributed near a 2-dimensional Klein bottle in a 7-dimensional space [5]. This example motivates the following problem.

Data Skeletonization Problem. Given a point cloud $\mathcal{C}$ in a metric space M, find a low-dimensional complex $\mathcal{S} \subseteq M$, that topologically approximates $\mathcal{C}$ in a way that the inclusions of certain subcomplexes of $\mathcal{S}$ into the *offsets* of $\mathcal{C}$ (unions of balls with a fixed radius and centers at points of $\mathcal{C}$) induce homology isomorphisms up to a given dimension.

The problem stated above is harder than describing the topological shape of a point cloud. Indeed, for a noisy random sample $\mathcal{C}$ of a circle, we aim not only to detect a circular shape $\mathcal{C}$, but also to approximate an unknown circle by a 1-dimensional graph $\mathcal{S}$ that should have exactly one cycle and be close to $\mathcal{C}$.

The proposed solution of the 1-dimensional case in [15] was introducing a homologically persistent skeleton (HoPeS) whose cycles are in a 1-1 correspondence with all 1-dimensional persistent homology classes of given data. The current paper extends the construction and optimality of HoPeS to higher dimensions.

1.2 Review of Closely Related Past Work

A metric graph reconstruction is related to the data skeletonization problem above. The output is an abstract metric graph or a higher-dimensional complex, which should be topologically similar to an input point cloud $\mathcal{C}$, but not embedded into the same space as $\mathcal{C}$, which makes the problem easier.

The classical Reeb graph is such an abstract graph defined for a function $f\colon \mathcal{Q} \to \mathbb{R}$, where $\mathcal{Q}$ is a simplicial complex built on the points of a given point cloud $\mathcal{C}$. For example, $\mathcal{Q}$ can be the Vietoris-Rips complex $\mathrm{VR}(\mathcal{C};\alpha)$ whose simplices are spanned by any set of points whose pairwise distances are at most 2α. Using the Vietoris-Rips complex at a fixed scale parameter, X. Ge et al. [21] proved that under certain conditions the Reeb graph has the expected homotopy type. Their experiments on real data concluded that 'there may be spurious loops in the Reeb graph no matter how we choose the parameter to decide the scale' [21, Section 3.3].

F. Chazal et al. [10] defined a new abstract α-Reeb graph G of a metric space X at a user-defined scale α. If X is ϵ-close to an unknown graph with edges of length at least 8ϵ, the output G is $34(\beta_1(G)+1)\epsilon$-close to the input X, where $\beta_1(G)$ is the first Betti number of G [10, Theorem 3.10]. The similarity between metric spaces was measured by the Gromov-Hausdorff distance. The algorithm runs at $O(n\log n)$ for n points in X.

Another classical approach is to use Forman's discrete Morse theory for a cell complex with a discrete gradient field when one builds a smaller homotopy equivalent complex whose number of critical cells is minimized by the algorithm in [20]. T. Dey et al. [19] built a higher-dimensional Graph Induced Complex GIC depending on a scale α and a user-defined graph that spans a cloud $\mathcal{C}$. If $\mathcal{C}$ is an ϵ-sample of a good manifold, GIC has the same homology H_1 as the Vietoris-Rips complex on $\mathcal{C}$ at scales $\alpha \geq 4\epsilon$.

A 1-dimensional homologically persistent skeleton [15] is based on a classical minimal spanning tree (MST) of a point cloud. Higher-dimensional MSTs (also called *minimal spanning acycles*) are currently a popular topic in the applied topology community, see Hiraoka and Shirai [12].

The most recent work by P. Skraba et al. [17] studies higher-dimensional MSTs from a probabilistic point of view in the case of *distinctly* weighted complexes, which helps to simplify algorithms and proofs. In practice, simplices often have equal weights, which is a generic non-singular case. For example, in the filtrations of Čech, Vietoris-Rips and α-complexes any obtuse triangle and its longest edge have the same weight equal to the half-length of the longest edge. We could allow ourselves arbitrarily small perturbations to make them distinctly weighted, but that is actually counter-productive, since the homologically persistent d-skeleton $\mathtt{HoPeS}^{(d)}$ would become the entire d-skeleton of the complex (not efficient). The more complicated proofs in the paper for non-distinctly weighted complexes are relevant — it is what makes $\texttt{HoPeS}^{(d)}$ reasonably small and thus efficient.

Among the results by P. Skraba et al. [17] the one closest to ours is [17, Theorem 3.23], which establishes a bijection between the set of weights of d-simplices outside of a minimal spanning acycle and the set of birth times in the d-dimensional persistence diagram. All further constructions and proofs in our paper substantially extend the ideas behind the 1-dimensional homologically persistent skeleton introduced in [15].

1.3 Contributions to Data Skeletonization

Definition 4.8 introduces a d-dimensional homologically persistent skeleton $\texttt{HoPeS}^{(d)}(\mathcal{C}_w)$ associated to a point cloud $\mathcal{C}$ or, more generally, to a weighted complex $\mathcal{C}_w$ built on $\mathcal{C}$. In comparison with the past methods, $\texttt{HoPeS}^{(d)}(\mathcal{C}_w)$ does not require an extra scale parameter and solves the Data Skeletonization Problem from Subsection 1.1 in the following sense. For any scale parameter α, a certain subcomplex of the full skeleton $\texttt{HoPeS}^{(d)}(\mathcal{C}_w)$ has the minimal total weight among all (in a suitable sense spanning) subcomplexes that have the homology up to dimension d of a given weighted complex $\mathcal{C}_{w\leq\alpha}$ at the same scale α (Theorem 4.12).

The key ingredient in the construction of $\texttt{HoPeS}^{(d)}(\mathcal{C}_w)$ is a d-dimensional minimal spanning tree whose properties are explored in Theorem 3.7. For completeness, we give a step-by-step algorithm for these trees (Algorithm 3.2), which is similar to algorithms by Kruskal [14] and P. Skraba et al. [17, Algorithm 1].

The original construction of a 1-dimensional homologically persistent skeleton in [15] did not explicitly define the death times of critical edges when they have equal weights. Example 4.4 shows that extra care is needed when assigning death times in those cases. The current paper carefully introduces the death times of critical faces in Definition 4.5. The implementation by Kurlin [15] used a duality between persistence in dimensions 0 and 1, so the death times of critical edges were still correctly computed as birth times of connected components in graphs dual to α-complexes in the plane.

2 Preliminaries

In this section we briefly go over some basic notions and prove basic statements that we will use later in the paper. We start by settling the notation.

2.1 Notation and the Euler Characteristic

- Number sets are denoted by $\mathbb{N}$ (natural numbers), $\mathbb{Z}$ (integers), $\mathbb{Q}$ (rationals), and $\mathbb{R}$ (reals). We treat zero as a natural number (so $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$). We denote the set of extended real numbers by $\overline{\mathbb{R}} = \{-\infty\} \cup \mathbb{R} \cup \{\infty\}$.

- Subsets of number sets, obtained by comparison with a certain number, are denoted by the suitable order sign and that number in the index. For example, $\mathbb{N}_{<42}$ denotes the set $\{n \in \mathbb{N} \mid n < 42\} = \{0, 1, \ldots, 41\}$ of all natural numbers smaller than 42, and $\mathbb{R}_{\geq 0}$ denotes the set $\{x \in \mathbb{R} \mid x \geq 0\}$ of non-negative real numbers.

- Intervals between two numbers are denoted by these two numbers in brackets and in the index. Round, or open, brackets ( ) denote the absence of the boundary in the set, and square, or closed, brackets [ ] its presence; for example $\mathbb{N}_{[5,10)} = \{n \in \mathbb{N} \mid 5 \leq n < 10\} = \{5, 6, 7, 8, 9\}$.

- In this paper we work exclusively with finite simplicial complexes. That is, whenever we refer to a 'complex' (or a 'subcomplex'), we mean a finite simplicial one. By a 'k-complex' (or a 'k-subcomplex') we mean a complex of dimension k or smaller. If $\mathcal{Q}$ is a complex, we denote its k-skeleton by $\mathcal{Q}^{(k)}$.

  Formally, we represent any (sub)complex as the set of its simplices ('faces') and any face as the set of its vertices. We will not need orientation for the results in this paper, so this suffices; had we wanted to take orientation into account, we would represent a face as a tuple.

  Example of this usage: suppose $\mathcal{Q}$ is a complex, $\mathcal{S} \subseteq \mathcal{Q}$ and $F \in \mathcal{Q}$. This means that $\mathcal{S}$ is a subcomplex of $\mathcal{Q}$, F is a face of $\mathcal{Q}$ and $\mathcal{S} \cup \{F\}$ is the subcomplex of $\mathcal{Q}$, obtained by adding the face F to the subcomplex $\mathcal{S}$.

- When we want to refer to the number of k-dimensional faces of a complex $\mathcal{Q}$ in a formula, we write $(\#k\text{-faces in } \mathcal{Q})$.

- For complexes $\mathcal{S} \subseteq \mathcal{Q}$ we use $\mathcal{S} \hookrightarrow \mathcal{Q}$ for the inclusion map. If we have further subcomplexes $\mathcal{S}' \subseteq \mathcal{S}$, $\mathcal{S}' \subseteq \mathcal{S}'' \subseteq \mathcal{Q}$, we use $(\mathcal{S}, \mathcal{S}') \hookrightarrow (\mathcal{Q}, \mathcal{S}'')$ to denote the inclusion of a pair.

- Given $k \in \mathbb{Z}$, a unital commutative ring R and a complex $\mathcal{Q}$,

  - $C_k(\mathcal{Q}; R)$ stands for the R-module of simplicial k-chains with coefficients in R,
  - $Z_k(\mathcal{Q}; R)$ stands for the submodule of k-cycles,
  - $B_k(\mathcal{Q}; R)$ stands for the submodule of k-boundaries,
  - $H_k(\mathcal{Q}; R)$ stands for the simplicial k-homology of $\mathcal{Q}$ with coefficients R.

  It is convenient to allow the dimension k to be any integer, since we sometimes subtract from it (also, the definition of the 0-homology does not have to be treated as a special case). Of course, there are no faces of negative dimension, so $C_k(\mathcal{Q}; R)$, $Z_k(\mathcal{Q}; R)$, $B_k(\mathcal{Q}; R)$ and $H_k(\mathcal{Q}; R)$ are all trivial modules whenever $k < 0$.

  The boundary maps between chains are denoted by

  $$\partial_k : C_k(\mathcal{Q}; R) \to C_{k-1}(\mathcal{Q}; R).$$

  Given a subcomplex $\mathcal{S} \subseteq \mathcal{Q}$, these induce boundary maps, defined on the relative homology,

  $$\partial_k : H_k(\mathcal{Q}, \mathcal{S}; R) \to H_{k-1}(\mathcal{S}; R).$$

  Unless otherwise stated all homologies that we consider in this paper are assumed to be over a given field $\mathbb{F}$, i.e. $H_k(\mathcal{Q})$ stands for $H_k(\mathcal{Q}; \mathbb{F})$. Hence $H_k(\mathcal{Q})$ is a vector space for any $k \in \mathbb{N}$ and any complex $\mathcal{Q}$; in particular it is free (posseses a basis) and has a well-defined dimension. Since we only consider cases when $\mathcal{Q}$ is a finite complex, the dimension $\beta_k(\mathcal{Q}) \coloneqq \dim H_k(\mathcal{Q})$ (the k-th *Betti number* of $\mathcal{Q}$) is a natural number, and there exists an isomorphism $H_k(\mathcal{Q}) \cong \mathbb{F}^{\beta_k(\mathcal{Q})}$.

  We freely use the fact that homology is a functor. For a map $f : \mathcal{Q}' \to \mathcal{Q}''$ we use $H_k(f)$ to denote the induced map $H_k(\mathcal{Q}') \to H_k(\mathcal{Q}'')$. (It is common in literature to use the notation f_* for this purpose, but we find it useful to include the dimension in the notation.)

We recall a couple of classical results in topology.

Proposition 2.1 *[18, Chapter 4, Section 3, Corollary 15] Let $\mathcal{Q}$ be a finite simplicial complex. The alternating sums*

$$\sum_{k\in\mathbb{N}}(-1)^k(\#k\text{-faces in } \mathcal{Q}) \qquad \text{and} \qquad \sum_{k\in\mathbb{N}}(-1)^k\beta_k(\mathcal{Q})$$

are well defined (all terms with $k > \dim\mathcal{Q}$ are zero, so they are effectively finite sums) and equal, regardless of the choice of the field $\mathbb{F}$. The number they are equal to is the Euler characteristic *of $\mathcal{Q}$, and is denoted by $\chi(\mathcal{Q})$.*

Corollary 2.2 *[6, Section 3] Let $\mathcal{Q}$ be a finite simplicial complex, $\mathcal{S}$ a subcomplex, $k\in\mathbb{N}$ and F a k-face in $\mathcal{Q}$ which is not in $\mathcal{S}$. Then either*

- $\beta_{k-1}(\mathcal{S}\cup\{F\}) = \beta_{k-1}(\mathcal{S}) - 1$ *("F kills a dimension in H_{k-1}") or*
- $\beta_k(\mathcal{S}\cup\{F\}) = \beta_k(\mathcal{S}) + 1$ *("F adds a dimension to H_k"),*

while in each case all other Betti numbers are the same for $\mathcal{S}$ and $\mathcal{S}\cup\{F\}$.

2.2 Fitting and Spanning Trees and Forests

In order to generalize a 1-dimensional homologically persistent skeleton based on a Minimal Spanning Tree to an arbitrary dimension, we need higher-dimensional analogues of spanning forests and trees. We also define the notion of 'fittingness' of a subcomplex.

Definition 2.3 Let $k\in\mathbb{N}$. Let $\mathcal{Q}$ be a simplicial complex and $\mathcal{S}$ a k-subcomplex of $\mathcal{Q}$.

- $\mathcal{S}$ is *k-spanning* (in $\mathcal{Q}$) when $\mathcal{S}^{(k-1)} = \mathcal{Q}^{(k-1)}$, i.e. the $(k-1)$-skeleton of $\mathcal{S}$ is the entire $(k-1)$-skeleton of $\mathcal{Q}$.
- $\mathcal{S}$ is a *k-forest* (in $\mathcal{Q}$) when $H_k(\mathcal{S}) = 0$.
- $\mathcal{S}$ is a *k-tree* (in $\mathcal{Q}$) when it is a k-forest and $H_{k-1}(\mathcal{S}\to\bullet)$ is an isomorphism.[1]

[1] Here $\bullet$ denotes a singleton, so there is a unique map $\mathcal{S}\to\bullet$. If $k\neq 1$, the condition for $\mathcal{S}$ being a k-tree simplifies to $H_k(\mathcal{S}) = H_{k-1}(\mathcal{S}) = 0$. For $k = 1$, the induced map $H_{k-1}(\mathcal{S}\to\bullet)$ is an isomorphism if and only if $\mathcal{S}$ has exactly one connected component.

- $\mathcal{S}$ is *k-fitting* (in $\mathcal{Q}$) when $H_i(\mathcal{S} \hookrightarrow \mathcal{Q})$ is an isomorphism for all $i \in \mathbb{N}_{\leq k}$.

For the sake of simplicity, we shorten 'k-spanning k-forest' to a 'spanning k-forest' (or to 'spanning forest', when k is understood). We proceed similarly with trees.

Note that every subcomplex, including $\varnothing$, is 0-spanning, since the (-1)-skeleton is empty. Also, $\varnothing$ is the only 0-forest and the only 0-tree.

Example 2.4 Let T be the set of all non-empty subsets of a set with four elements, i.e. a geometric realization of T is a tetrahedron. Then T is a spanning 3-tree of itself. Figure 1 depicts two spanning 2-trees of T.

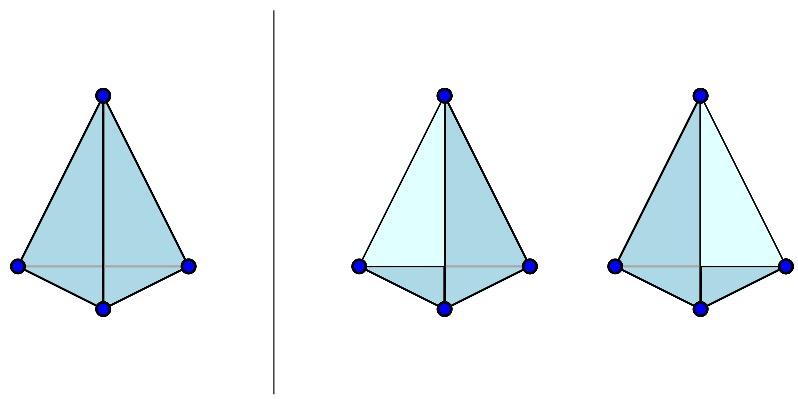


Figure 1: Geometric realization of a tetrahedron T (left) and two of its spanning 2-trees (right).

Remark 2.5 The concepts in Definition 2.3 were inspired by [2, 7], although we tweaked them a bit, to better serve our purposes. In particular, the definition of a k-forest in [7] was given in an 'absolute' sense, as linear independence of the columns of the boundary map ∂_k between $\mathbb{Z}$-chains. This is equivalent to $H_k(\mathcal{S}; \mathbb{R}) = 0$ (or more generally, $H_k(\mathcal{S}; \mathbb{F}) = 0$ if $\mathbb{F}$ is a field of characteristic 0). However, we purposefully define forests (and trees) in a 'relative' sense (depending on the choice of the field $\mathbb{F}$), as this allows us to prove the results of the paper in greater generality.

Remark 2.6 What we call a spanning k-tree some other authors [12, 17] call a k-spanning acycle. This definition originated in Kalai's work [13].

He considered k-dimensional simplicial complexes, which contain the entire $(k-1)$-skeleton and for them defined ‘simplicial spanning trees’.

The following lemma establishes basic properties of spanning subcomplexes that we use throughout the paper.

Lemma 2.7 *Let $\mathcal{Q}$ be a finite simplicial complex and $\mathcal{S}$ a k-spanning k-subcomplex of $\mathcal{Q}$ for some $k \in \mathbb{N}$.*

1. *The map $H_i(\mathcal{S} \hookrightarrow \mathcal{Q})$ is an isomorphism for all $i \in \mathbb{N}_{\le k-2}$ (i.e. $\mathcal{S}$ is $(k-2)$-fitting in $\mathcal{Q}$) and a surjection for $i = k-1$.*

2. *The formula*

$$\Big(\#k\text{-faces in } \mathcal{Q}\Big) + \beta_{k-1}(\mathcal{Q}^{(k)}) - \beta_k(\mathcal{Q}^{(k)}) = \Big(\#k\text{-faces in } \mathcal{S}\Big) + \beta_{k-1}(\mathcal{S}) - \beta_k(\mathcal{S})$$

   holds.

3. *If $\beta_{k-1}(\mathcal{S}) > \beta_{k-1}(\mathcal{Q})$, there exists a k-face F in $\mathcal{Q} \setminus \mathcal{S}$ such that*

$$\beta_k(\mathcal{S} \cup \{F\}) = \beta_k(\mathcal{S}) \quad \textit{and} \quad \beta_{k-1}(\mathcal{S} \cup \{F\}) = \beta_{k-1}(\mathcal{S}) - 1.$$

4. *If $\mathcal{S}$ is $(k-1)$-fitting in $\mathcal{Q}$, a k-subcomplex $F \subseteq \mathcal{S}$ exists, which is $(k-1)$-fitting k-spanning k-forest in $\mathcal{S}$ (and consequently also in $\mathcal{Q}$).*

5. *Suppose $\mathcal{S}$ is $(k-1)$-fitting in $\mathcal{Q}$ and $F \subseteq \mathcal{S}$ is a $(k-1)$-fitting k-spanning k-forest in $\mathcal{S}$ (equivalently, in $\mathcal{Q}$). Then the diagram*

$$\begin{array}{ccc}
 & H_k\left((\mathcal{S},\varnothing) \hookrightarrow (\mathcal{S},F)\right) & \\
H_k(\mathcal{S}) & \longrightarrow & H_k(\mathcal{S},F) \\
{\scriptstyle H_k(\mathcal{S} \hookrightarrow \mathcal{Q})}\Big\downarrow & & \Big\downarrow{\scriptstyle H_k((\mathcal{S},F) \hookrightarrow (\mathcal{Q},F))} \\
H_k(\mathcal{Q}) & \longrightarrow & H_k(\mathcal{Q},F) \\
 & H_k\left((\mathcal{Q},\varnothing) \hookrightarrow (\mathcal{Q},F)\right) &
\end{array}$$

   commutes and the horizontal arrows are isomorphisms. Hence the left arrow is an isomorphism if and only if the right one is.

Proof.

1. Follows from the fact that simplicial homology in dimension i depends only on i- and $(i+1)$-dimensional faces, with i-faces providing the generators and $(i+1)$-faces the relations.

2. Since $\mathcal{S}$ is k-spanning, it has the same number of faces up to dimension $k-1$ and (per the previous item) the same homologies up to dimension $k-2$. Thus

$$(-1)^k\Big(\#k\text{-faces in } \mathcal{Q} - \#k\text{-faces in } \mathcal{S}\Big) = \chi(\mathcal{Q}^{(k)}) - \chi(\mathcal{S}) =$$
$$= (-1)^k\beta_k(\mathcal{Q}^{(k)}) + (-1)^{k-1}\beta_{k-1}(\mathcal{Q}^{(k)}) - (-1)^k\beta_k(\mathcal{S}) - (-1)^{k-1}\beta_{k-1}(\mathcal{S}).$$

   After rearranging the result follows.

3. Let $\{S_1, S_2, \ldots, S_m\}$ be the set of k-faces in $\mathcal{S}$. Consider families of k-faces in $\mathcal{Q} \setminus \mathcal{S}$ which, when added to $\mathcal{S}$, reduce the $(k-1)$-homology (regardless of whether the k-th Betti number of the expanded subcomplex changes). By assumption $\beta_{k-1}(\mathcal{S}) > \beta_{k-1}(\mathcal{Q})$, at least one such family exists, namely the set of *all* k-faces in $\mathcal{Q} \setminus \mathcal{S}$. Let $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$ be one of such families which contains the minimal possible number of k-faces (of course $n \geq 1$). Minimality of $\mathcal{F}$ implies that the image under ∂_k does not change when adding only $n-1$ faces, that is

$$\partial_k\Big(\langle S_1, S_2, \ldots, S_m\rangle\Big) = \partial_k\Big(\langle S_1, S_2, \ldots, S_m, F_1, F_2, \ldots, F_{n-1}\rangle\Big) =: B$$

   (here $\langle\ \rangle$ denotes the linear span). Since adding $\mathcal{F}$ to $\mathcal{S}$ reduces $(k-1)$-homology, a linear combination

$$s := \sum_{i=1}^{m} c_i S_i + \sum_{j=1}^{n} d_j F_j$$

   exists such that $\partial_k(s) \notin B$. Consequently $\partial_k(F_n) \notin B$, so just adding F_n to $\mathcal{S}$ reduces homology in dimension $(k-1)$ (implying that $n = 1$). It follows from Corollary 2.2 that $\mathcal{S} \cup \{F_n\}$ remains a k-forest while $\beta_{k-1}(\mathcal{S} \cup \{F_n\}) = \beta_{k-1}(\mathcal{S}) - 1$.

4. Let $\{S_1, S_2, \ldots, S_m\}$ be the set of k-faces in $\mathcal{S}$. Since $\mathcal{S}$ is a k-complex, we have $H_k(\mathcal{S}) \cong Z_k(\mathcal{S})$ (every equivalence class is a singleton). Let

$n \coloneqq \beta_k(\mathcal{S})$ be the dimension of the vector space of k-cycles of $\mathcal{S}$. Choose a basis $b_1, \ldots, b_n$ of $Z_k(\mathcal{S})$ and expand these basis elements as

$$b_i = \sum_{j=1}^{m} c_{ij} S_j.$$

Consider the system of linear equations

$$\sum_{j=1}^{m} c_{ij} x_j = 0.$$

Since a basis is linearly independent, this is a system of n independent linear equations with m variables, where $n \le m$ (since $Z_k(\mathcal{S}) \subseteq C_k(\mathcal{S})$). Thus the system can be solved for n leading variables in the sense that we express them with the remaining $m - n$ ones. Without loss of generality assume that the first n variables are the leading ones. This means that the system can be equivalently written as

$$x_i + \sum_{j=n+1}^{m} \widetilde{c}_{ij} x_j = 0.$$

Define $\widetilde{b}_i \coloneqq S_i + \sum_{j=n+1}^{m} \widetilde{c}_{ij} S_j$; then $\left\{ \widetilde{b}_i \,\middle|\, i \in \mathbb{N}_{[1,n]} \right\}$ is also a basis for $Z_k(\mathcal{S})$.

Define $F \coloneqq \mathcal{S} \setminus \left\{ S_i \,\middle|\, i \in \mathbb{N}_{[1,n]} \right\}$. Clearly F is k-spanning (therefore $(k-2)$-fitting) in $\mathcal{S}$ and $\mathcal{Q}$. Let $z = \sum_{j=n+1}^{m} d_j S_j$ be an arbitrary k-cycle of F. The boundary map has the same definition for F and $\mathcal{S}$, so z is also a cycle in $\mathcal{S}$. Expand it as

$$z = \sum_{i=1}^{n} e_i \widetilde{b}_i.$$

Since z does not include any S_j for $j \le n$, necessarily all e_is are zero, and then $z = 0$. We conclude that F is a k-forest.

Adding n faces to F to recover $\mathcal{S}$ increases the dimension of k-homology by n. Since a change of a k-face either modifies the dimension of k-homology by one or of $(k-1)$-homology by one (Corollary 2.2), the $(k-1)$-homology of F remains the same as of $\mathcal{S}$. That is, F is $(k-1)$-fitting in $\mathcal{S}$ and $\mathcal{Q}$.

5. The long exact sequence of a pair is natural, so the following diagram commutes.

$$\begin{array}{ccccccccc}
\overbrace{H_k(F)}^{0} & \xrightarrow{0} & H_k(\mathcal{S}) & \xrightarrow{\cong} & H_k(\mathcal{S},F) & \xrightarrow{0} & H_{k-1}(F) & \xrightarrow{\cong} & H_{k-1}(\mathcal{S}) \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
\underbrace{H_k(F)}_{0} & \xrightarrow[0]{} & H_k(\mathcal{Q}) & \xrightarrow[\cong]{} & H_k(\mathcal{Q},F) & \xrightarrow[0]{} & H_{k-1}(F) & \xrightarrow[\cong]{} & H_{k-1}(\mathcal{Q})
\end{array}$$

Since $H_k(F) = 0$, the outgoing maps are 0. Since F is $(k-1)$-fitting, the maps $H_{k-1}(F \hookrightarrow \mathcal{S})$ and $H_{k-1}(F \hookrightarrow \mathcal{Q})$ are isomorphisms, so the preceding boundary maps are 0. Thus the maps $H_k((\mathcal{S},\varnothing) \hookrightarrow (\mathcal{S},F))$ and $H_k((\mathcal{Q},\varnothing) \hookrightarrow (\mathcal{Q},F))$ are isomorphisms.

■

Proposition 2.8 *Let $k,n \in \mathbb{N}$ and let Δ_n be a standard n-simplex. The following statements hold.*

1. *There exists a spanning k-tree in Δ_n.*

2. *The number of k-faces in any spanning k-tree in Δ_n is $\binom{n}{k}$ if $k \geq 1$, and 0 if $k = 0$.*

3. *Let F be a spanning k-forest in Δ_n. Then F is a k-tree if and only if it is a maximal k-forest in the sense that for every k-face $E \in \Delta_n \setminus F$ we have $H_k(F \cup \{E\}) \neq 0$.*

Proof.

1. This follows if we apply Lemma 2.7(4) for $\Delta_n^{(k)} \subseteq \Delta_n$, but we can be much more explicit.

   If $k = 0$, then $\varnothing$ is a spanning 0-tree. If $k \geq 1$, choose a vertex v in Δ_n. Define T to consist of the $(k-1)$-skeleton of Δ_n, as well as of those k-faces of $\mathcal{S}$ which contain v. Then T is k-spanning by definition, and there exists an obvious deformation retraction of T onto v. This deformation retraction induces homology isomorphisms in all dimensions, so T is necessarily a tree.

2. The only spanning 0-tree is $\varnothing$, so the statement holds for $k = 0$. Assume $k \geq 1$. Let T be any spanning k-tree in Δ_n and let x be the number of k-faces of T. Counting the number of faces, we obtain

$$\chi(T) = \Big(\sum_{i \in \mathbb{N}_{\leq k-1}} (-1)^i \binom{n+1}{i+1} \Big) + (-1)^k x = -\Big((-1)^k \binom{n}{k} - 1 \Big) + (-1)^k x.$$

On the other hand, since T is k-spanning, it has the same homology up to dimension $k-2$ as the standard simplex Δ_n, and thus the same homology up to dimension $k-2$ as a point. Since T is a k-tree, this holds also for the dimensions $k-1$ and k. Hence

$$\chi(T) = \sum_{i \in \mathbb{N}_{\leq k}} \beta_i(T) = 1.$$

Equating the two versions of the Euler characteristic (as in Proposition 2.1), we obtain $x = \binom{n}{k}$.

3. Clearly the statement holds for the only 0-forest $F = \varnothing$. Assume hereafter that $k \geq 1$.

(⇒)

Suppose F is a k-tree. By Corollary 2.2, adding E to F either decreases β_{k-1} by 1 or increases β_k by 1. The former is impossible: if $k \geq 2$, then $H_{k-1}(F)$ is already trivial, and if $k = 1$ (therefore $\beta_{k-1}(F) = 1$), adding a face cannot decrease the number of connected components to zero.

Hence $\beta_k(F \cup \{E\}) = 1$, so $F \cup \{E\}$ is not a k-forest.

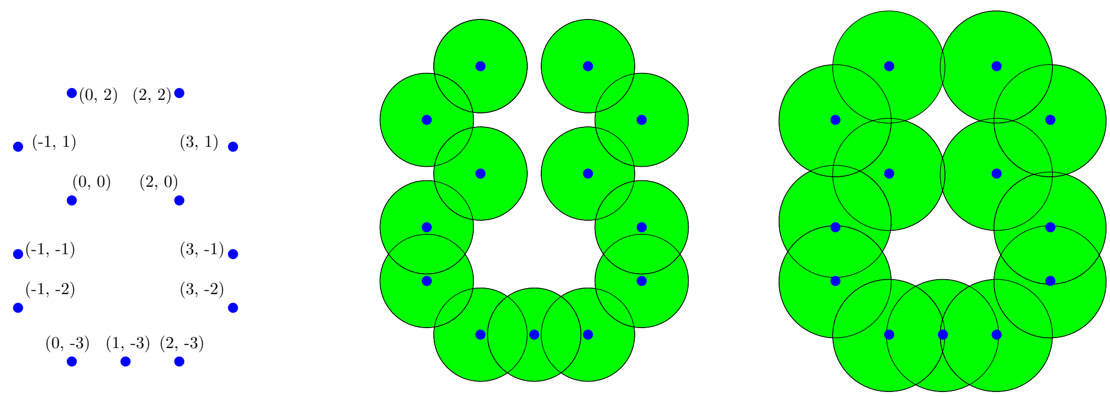
(⇐)

Apply basic graph theory if $k = 1$ (1-forests and 1-trees are just the usual forests and trees). Suppose $k \geq 2$ and assume that the spanning k-forest F is not a k-tree, so $\beta_{k-1}(F) > 0 = \beta_{k-1}(\Delta_n)$. Use Lemma 2.7(3) to find a k-face $E \in \Delta_n \smallsetminus F$ with $\beta_k(F \cup \{E\}) = \beta_k(F) = 0$, contradicting the assumption.

■

2.3 Filtrations on a Point Cloud

In practice, point clouds are often obtained by sampling from a particular shape, which we want to reconstruct. However, from the point of view of a topologist, point clouds themselves do not have an interesting shape — the dimension of 0-homology is the number of points in the point cloud and the higher-dimensional homology groups are all trivial. The idea is to assume that the point cloud is a subspace of a larger metric space (let us denote its metric by D), typically some Euclidean space $\mathbb{R}^N$, in which each point can be thickened to a ball of some specified radius α. The union of these balls is called the α-*offset* of $\mathcal{C}$ and is denoted by $\mathcal{C}(\alpha)$, see Figure 2.



cloud $\mathcal{C}$ 0.95-offset set of $\mathcal{C}$ 1.05-offset set of $\mathcal{C}$

Figure 2: Point cloud $\mathcal{C}$ and two example offsets of $\mathcal{C}$. The 1.05-offset has non-trivial first homology.

The nerve of $\mathcal{C}(\alpha)$ is called the *Čech complex* $\check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha)$ of $\mathcal{C}$ at α. The nerve lemma [1] says that the homotopy type of $\check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha)$ is the same as the homotopy type of $\mathcal{C}(\alpha)$. Hence, $\check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha)$ is a potentially good approximation to the shape, from which we sampled the point cloud.

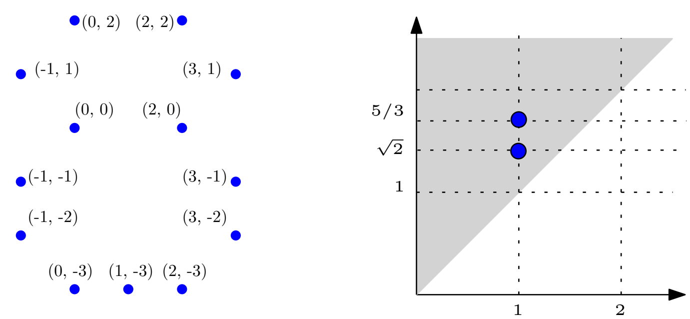
For any $\alpha < \alpha'$, we have the inclusion $\check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha) \subseteq \check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha')$. That is, the collection $\left(\check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha)\right)_{\alpha\in\mathbb{R}}$ is a *filtration*.

Čech filtration is not ideal for computation, as it requires storing all high-dimensional simplices in a computer memory. On the other hand, the filtration of *Vietoris-Rips complexes* is completely determined by the

1-dimensional skeleton. For any scale $\alpha \in \mathbb{R}$, the complex $\mathrm{VR}(\mathcal{C};\alpha)$ has a k-dimensional simplex on points $v_0,\dots,v_k \in \mathcal{C}$ whenever all pairwise distances $D(v_i,v_j) \le 2\alpha$ for all $0 \le i < j \le k$.

2.4 Persistent Homology of a Filtration

For excellent introductions to persistent homology, see [11, 4, 3, 9]. The usual homology is defined for a single complex, but the key idea of persistence is to consider an entire filtration of complexes $\big(Q(\mathcal{C};\alpha)\big)_{\alpha\in\mathbb{R}}$, rather that just a single stage $Q(\mathcal{C};\alpha)$ at a specific scale parameter α. The reason for this is that it is hard (or even impossible) to choose a single parameter value in a way that assures that $Q(\mathcal{C};\alpha)$ is a good approximation to the shape we sampled the point cloud from. Also, choosing a single parameter value is highly unstable.



cloud $\mathcal{C}$ $\quad$ $\mathrm{PD}_1((\check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha))_\alpha)$

Figure 3: Point cloud $\mathcal{C}$ and its persistence diagram in dimension 1 (with homology coefficients $\mathbb{R}$), obtained via a filtration of Čech complexes on $\mathcal{C}$. Each point in the diagram represents a cycle present over a range of parameters α.

Persistent homology in dimension k tracks changes in the k-homology $H_k\left(Q(\mathcal{C};\alpha)\right)$ over a range of scales α. This information can be summarized by a *persistence diagram* $\mathrm{PD}_k\Big(\big(Q(\mathcal{C};\alpha)\big)_{\alpha\in\mathbb{R}}\Big)$. A dot (p,q) in a persistence diagram represents an interval $\mathbb{R}_{[p,q)}$ corresponding to a topological feature, a

k-dimensional void, which appears at p and disappears at q. These barcodes play the same role as a histogram would in summarizing the shape of data, with long intervals corresponding to strong topological signals and short ones to noise.

In Figure 3 the persistence diagram $\mathrm{PD}_1\Big(\big(\check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha)\big)_{\alpha\in\mathbb{R}}\Big)$ consists of 2 dots. The dot $(1,\sqrt{2})$ says that a 1-dimensional cycle enclosing the smaller hole (the upper bounded component of $\mathbb{R}^2 \setminus \check{\mathrm{C}}\mathrm{ech}(\mathcal{C};\alpha)$) was born at $\alpha = 1$ and died at $\alpha = \sqrt{2}$ when this hole was filled. Similarly, the dot $(1, \frac{5}{3})$ says that the larger hole persisted from the same birth time $\alpha = 1$ until the later death time $\alpha = \frac{5}{3}$.

2.5 Weighted Simplices

Given a filtration, we can assign to any face in it its *weight* as the parameter value when it appears in the filtration. The union of all stages in the filtration, together with the weights of all simplices, is thus a weighted complex (a higher-dimensional analogue of weighted graphs).

For both Čech and Vietoris-Rips filtrations of a point cloud $\mathcal{C}$, the simplicial complex for parameter values $\alpha \geq \max_{v_i,v_j\in\mathcal{C}} \frac{D(v_i,v_j)}{2}$ is a full simplex on $|\mathcal{C}|$ vertices. Thus we can think of the whole filtration as being encoded by a weighted *simplex*. For this reason most of the results stated in this paper are in terms of weighted simplices. If a certain filtration does not terminate with a full simplex, we can always complete the simplicial complex at the last step to a full simplex by adding the missing faces and assigning them weight bigger than that of all faces in the original filtration.

The main reason to work with a single weighted simplex is to have a simpler notion of a minimal spanning d-tree at the last stage of the filtration. Otherwise 'minimal spanning d-tree' should be replaced with 'minimal $(d-1)$-fitting d-spanning d-forest' and arguments would get more complicated.

We give a formal definition of a weighted simplex. As mentioned in the subsection on notation, we will not need orientation, so we can encode faces with sets, rather than tuples.

Definition 2.9 Given a set $\mathcal{C}$, let $\mathscr{P}_+(\mathcal{C})$ denote the set of non-empty subsets of $\mathcal{C}$.

- A *weighting* on a set $\mathcal{C}$ is a map $w\colon \mathscr{P}_+(\mathcal{C}) \to \mathbb{R}_{\geq 0}$ which is monotone in the sense that if $\varnothing \neq A \subseteq B \subseteq \mathcal{C}$, then $w(A) \leq w(B)$. For any $A \in \mathscr{P}_+(\mathcal{C})$ the value $w(A)$ is the *weight* of A (relative to the weighting w).

- A *weighted simplex* is a pair $(\mathcal{C}, w)$, where $\mathcal{C}$ is a non-empty finite set and w a weighting on it. We denote $\mathcal{C}_w \coloneqq (\mathcal{C}, w)$ for short.

- For a weighted simplex $\mathcal{C}_w$ and any family of subsets $\mathscr{S} \subseteq \mathscr{P}_+(\mathcal{C})$ we denote its *total weight* by $\operatorname{tw}(\mathscr{S}) \coloneqq \sum_{A\in\mathscr{S}} w(A)$.

Monotonicity of weighting implies that

$$\mathcal{C}_{w\leq\alpha} \coloneqq \left\{A \in \mathscr{P}_+(\mathcal{C}) \,\middle|\, w(A) \leq \alpha\right\}$$

is a subcomplex for any $\alpha \in \mathbb{R}$, and $(\mathcal{C}_{w\leq\alpha})_{\alpha\in\mathbb{R}}$ is a filtration. Note that the image of a weighting is a finite subset of $\mathbb{R}_{[0,w(\mathcal{C})]}$, and we have $\mathcal{C}_{w\leq\alpha} = \mathscr{P}_+(\mathcal{C})$ for all $\alpha \in \mathbb{R}_{\geq w(\mathcal{C})}$.

Conversely, a filtration $\big(Q(\mathcal{C};\alpha)\big)_{\alpha\in\mathbb{R}}$ induces a weighted *complex* with the weighting

$$w(A) = \sup\{\alpha \in \mathbb{R} \mid A \notin Q(\mathcal{C};\alpha)\} = \inf\{\alpha \in \mathbb{R} \mid A \in Q(\mathcal{C};\alpha)\},$$

and we get a weighted *simplex* whenever each non-empty subset of $\mathcal{C}$ appears in the filtration at a specific time $\alpha \in \mathbb{R}_{\geq 0}$.

In the specific case of Čech filtration, the weighting is given by

$$w(A) \coloneqq \inf\left\{\alpha \in \mathbb{R}_{\geq 0} \,\middle|\, \exists\, x \in X \,.\, \forall\, a \in A \,.\, D(a,x) \leq \alpha\right\},$$

and in the case of Vietoris-Rips filtration by

$$w(A) \coloneqq \frac{1}{2} \cdot \sup_{a,b\in A} D(a,b)$$

for $A \in \mathscr{P}_+(\mathcal{C})$.

3 Minimal Spanning d-Tree

The first step in constructing a 1-dimensional homologically persistent skeleton in [15] was to take a classical (1-dimensional) Minimal Spanning Tree of a given point cloud. With this idea in mind, we generalize the concept of a minimal spanning tree to higher dimensions. Hereafter fix a weighted simplex $\mathcal{C}_w$ and a dimension $d \in \mathbb{N}$.

Definition 3.1 (Minimal Spanning Tree) A *minimal spanning d-tree* (or simply *minimal spanning tree* when d is understood) of $\mathcal{C}_w$ is a spanning d-tree of $\mathcal{C}_w$ with minimal total weight. We use $\mathtt{MST}^{(d)}(\mathcal{C}_w)$ to denote any chosen minimal spanning tree, and shorten this to $\mathtt{MST}^{(d)}$ when $\mathcal{C}_w$ is understood from the context. For any $\alpha \in \mathbb{R}$ we define

$$\mathtt{MST}^{(d)}_\alpha(\mathcal{C}_w) := \left\{ A \in \mathtt{MST}^{(d)}(\mathcal{C}_w) \;\middle|\; w(A) \leq \alpha \right\}$$

and shorten this to $\mathtt{MST}^{(d)}_\alpha$ when there is no ambiguity.

By Proposition 2.8(1) a spanning d-tree of $\mathcal{C}_w$ exists, and so a minimal spanning d-tree exists also. In general there may be many minimal spanning trees; for example, any two edges form a minimal spanning 1-tree in an equilateral triangle.

In the next subsection we give an explicit construction for a minimal spanning tree and then prove that all minimal spanning trees are obtained this way. This allows us to later prove optimality of minimal spanning trees at all scales (Theorem 3.7).

3.1 Construction of Minimal Spanning d-Trees

The idea to obtain a minimal spanning tree is to go through the image of w and inductively construct a $(d-1)$-fitting d-spanning d-forest $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ in $\mathcal{C}_{w\leq\alpha}$, with minimal total weight among such, for every $\alpha \in \mathbb{R}$.

Let $w_1 < w_2 < \ldots < w_n$ be all elements of $\mathrm{im}(w)$ and set additionally $w_0 = -\infty$, $w_{n+1} = \infty$. Declare $\widetilde{\mathtt{MST}}^{(d)}_\alpha := \varnothing$ for all $\overline{\mathbb{R}}_{[-\infty,w_1)}$.

Take $k \in \mathbb{N}_{[1,n]}$ and assume that $\widetilde{\mathtt{MST}}^{(d)}_\gamma$ has been defined for all $\gamma < w_k$. We define $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ for $\alpha \in \mathbb{R}_{[w_k,w_{k+1})}$ to consist of the subcomplex we had at the previous stage, but with as many faces of weight w_k added as possible while still keeping the subcomplex a forest.

Explicitly, let $F_1, F_2, \dots, F_m$ be all d-faces of weight w_k.[2] Define $\mathcal{S}_0$ to be the union of $\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}$ with the set of all faces in $\mathcal{C}_w$ which have the weight w_k and dimension at most $d-1$. Note that $\mathcal{S}_0$ is a spanning d-forest in $\mathcal{C}_{w \le \alpha}$.

Suppose inductively that we have defined a spanning d-forest $\mathcal{S}_{i-1}$, where $i \in \mathbb{N}_{[1,m]}$. If $\mathcal{S}_{i-1} \cup \{F_i\}$ is still a d-forest, define $\mathcal{S}_i \coloneqq \mathcal{S}_{i-1} \cup \{F_i\}$, otherwise define $\mathcal{S}_i \coloneqq \mathcal{S}_{i-1}$. In the end, set $\widetilde{\mathtt{MST}}^{(d)}_{\alpha} \coloneqq \mathcal{S}_m$ which is a spanning d-forest by construction.

Here is the summary of this procedure, written as an explicit algorithm.

Algorithm 3.2 Construction of a minimal spanning d-tree

$w_0 \coloneqq -\infty$
$w_1, w_2, \dots, w_n \coloneqq$ elements of $\mathrm{im}(w)$, in order
$w_{n+1} \coloneqq \infty$
$\widetilde{\mathtt{MST}}^{(d)}_{\alpha} \coloneqq \varnothing$ for all $\alpha \in \overline{\mathbb{R}}_{[-\infty, w_1)}$
for $k = 1$ **to** n **do**
    $F_1, F_2, \dots, F_m \coloneqq d$-faces of weight w_k in $\mathcal{C}_w$
    $\mathcal{S}_0 \coloneqq \widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}} \cup \left\{ A \in \mathcal{C}^{(d-1)}_w \;\middle|\; w(A) = w_k \right\}$
    for $i = 1$ **to** m **do**
        if $\beta_d(\mathcal{S}_{i-1} \cup \{F_i\}) = 0$ **then**
            $\mathcal{S}_i \coloneqq \mathcal{S}_{i-1} \cup \{F_i\}$
        else
            $\mathcal{S}_i \coloneqq \mathcal{S}_{i-1}$
        end if
    end for
    $\widetilde{\mathtt{MST}}^{(d)}_{\alpha} \coloneqq \mathcal{S}_m$ for all $\alpha \in \overline{\mathbb{R}}_{[w_k, w_{k+1})}$
end for

Example 3.3 Let $\mathcal{C}$ be a point cloud consisting of four vertices with pairwise distances as specified on the left-hand side of Figure 4. The right-hand side of Figure 4 depicts a minimal spanning 2-tree at different scales α. The weighting is induced by the Čech filtration on $\mathcal{C}$.

[2] The order of these faces can be chosen arbitrarily. It is because of this freedom that there are in general many minimal spanning trees.

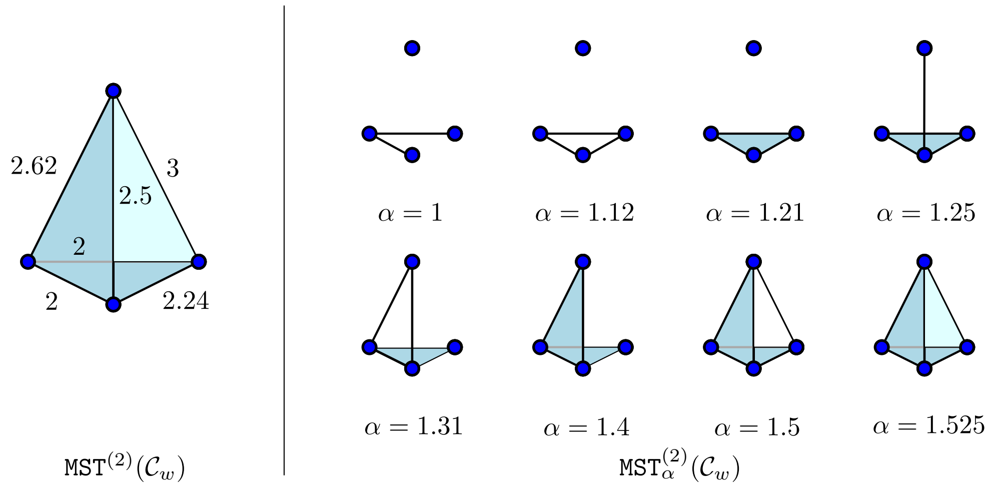


Figure 4: Geometric realizations of $\mathtt{MST}^{(2)}(\mathcal{C}_w)$ and its reduced forms with respect to Čech filtration of a point cloud $\mathcal{C}$ with four vertices.

3.2 Optimality of Minimal Spanning d-Trees

In this subsection we prove that the (final stage of the) d-forest constructed earlier is indeed a minimal spanning d-tree, that any minimal spanning tree can be obtained this way, and finally, that reduced versions of minimal spanning trees are optimal in the sense that they have minimal total weight among all $(d-1)$-fitting d-spanning d-forests in $\mathcal{C}_{w\leq\alpha}$ (that is, they are optimal at every scale, not just at the final one, as per definition).

Lemma 3.4 *For every $\alpha \in \mathbb{R}$ the subcomplex $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ is a $(d-1)$-fitting d-spanning d-forest in $\mathcal{C}_{w\leq\alpha}$, and moreover has minimal total weight among all $(d-1)$-fitting d-spanning d-forests in $\mathcal{C}_{w\leq\alpha}$.*

Proof. $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ is a d-spanning d-forest by construction. As for the rest, it suffices to prove this for $\alpha \in \operatorname{im}(w) = \{w_k \mid k \in \mathbb{N}_{[1,n]}\}$. We prove it by induction on k. Certainly, this holds for $k = 0$ (as before, we use the notation $w_0 = -\infty$, $w_{n+1} = \infty$).

Take $k \in \mathbb{N}_{[1,n]}$ and assume $\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}$ is a minimal $(d-1)$-fitting d-spanning d-forest. For fittingness it suffices to check that $\widetilde{\mathtt{MST}}^{(d)}_{w_k}$ is $(d-1)$-fitting in $\mathcal{C}^{(d)}_{w\leq w_k}$. By Lemma 2.7(1), $\widetilde{\mathtt{MST}}^{(d)}_{w_k}$ is at least $(d-2)$-fitting and the map

$H_{d-1}\left(\widetilde{\mathtt{MST}}^{(d)}_{w_k} \hookrightarrow \mathcal{C}^{(d)}_{w\le w_k}\right)$ is surjective. To prove it is bijective, it suffices to verify that the dimensions of the domain and the codomain match.

Using Lemma 2.7(2) for w_k and w_{k-1} yields

$$\left(\#d\text{-faces in } \mathcal{C}_{w\le w_k}\right) + \beta_{d-1}(\mathcal{C}^{(d)}_{w\le w_k}) - \beta_d(\mathcal{C}^{(d)}_{w\le w_k}) = \\ = \left(\#d\text{-faces in } \widetilde{\mathtt{MST}}^{(d)}_{w_k}\right) + \beta_{d-1}(\widetilde{\mathtt{MST}}^{(d)}_{w_k}) - \beta_d(\widetilde{\mathtt{MST}}^{(d)}_{w_k})$$

and

$$\left(\#d\text{-faces in } \mathcal{C}_{w\le w_{k-1}}\right) + \beta_{d-1}(\mathcal{C}^{(d)}_{w\le w_{k-1}}) - \beta_d(\mathcal{C}^{(d)}_{w\le w_{k-1}}) = \\ = \left(\#d\text{-faces in } \widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}\right) + \beta_{d-1}(\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}) - \beta_d(\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}).$$

We know that $\widetilde{\mathtt{MST}}^{(d)}_{w_k}$ and $\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}$ are d-forests, and the induction hypothesis tells us $\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}$ is $(d-1)$-fitting, so the above equalities reduce to

$$\left(\#d\text{-faces in } \mathcal{C}_{w\le w_k}\right) + \beta_{d-1}(\mathcal{C}^{(d)}_{w\le w_k}) - \beta_d(\mathcal{C}^{(d)}_{w\le w_k}) = \left(\#d\text{-faces in } \widetilde{\mathtt{MST}}^{(d)}_{w_k}\right) + \beta_{d-1}(\widetilde{\mathtt{MST}}^{(d)}_{w_k}),$$
$$\left(\#d\text{-faces in } \mathcal{C}_{w\le w_{k-1}}\right) - \beta_d(\mathcal{C}^{(d)}_{w\le w_{k-1}}) = \left(\#d\text{-faces in } \widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}\right).$$

Subtract these two equalities and rearrange the result to get

$$\beta_{d-1}(\mathcal{C}^{(d)}_{w\le w_k}) - \beta_{d-1}(\widetilde{\mathtt{MST}}^{(d)}_{w_k}) =$$

$$= \underbrace{\left(\beta_d(\mathcal{C}^{(d)}_{w\le w_k}) - \beta_d(\mathcal{C}^{(d)}_{w\le w_{k-1}})\right)}_{\#d\text{-faces with weight } w_k \text{ which increase } \beta_d} + \\ + \underbrace{\left(\#d\text{-faces in } \widetilde{\mathtt{MST}}^{(d)}_{w_k} - \#d\text{-faces in } \widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}\right)}_{\#d\text{-faces with weight } w_k \text{ which do not increase } \beta_d} - \\ - \underbrace{\left(\#d\text{-faces in } \mathcal{C}_{w\le w_k} - \#d\text{-faces in } \mathcal{C}_{w\le w_{k-1}}\right)}_{\#d\text{-faces with weight } w_k}$$

which is zero, proving the desired equality of dimensions.

We now prove minimality inductively on k. Clearly, the statement holds for $k = 0$.

Let $\mathcal{S}$ be a $(d-1)$-fitting d-spanning d-forest in $\mathcal{C}_{w\leq w_k}$. Define

$$\mathcal{S}' := \{F \in \mathcal{S} \mid w(F) < w_k\}.$$

Then $\mathcal{S}'$ is a d-spanning d-forest in $\mathcal{C}_{w\leq w_{k-1}}$; in particular, $H_{d-1}(\mathcal{S}' \hookrightarrow \mathcal{C}_{w\leq w_{k-1}})$ is surjective. Denote $m := \beta_{d-1}(\mathcal{S}') - \beta_{d-1}(\mathcal{C}_{w\leq w_{k-1}})$. Using Lemma 2.7(3) m times, we get d-faces $F_1, \dots, F_m \in \mathcal{C}_{w\leq w_{k-1}} \setminus \mathcal{S}'$, such that $\mathcal{S}'' := \mathcal{S}' \cup \{F_1, \dots, F_m\}$ is a $(d-1)$-fitting d-spanning d-forest in $\mathcal{C}_{w\leq w_{k-1}}$.

By the induction hypothesis the total weight of $\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}$ is at most the total weight of $\mathcal{S}''$. Let $a \in \mathbb{N}$ be the number of faces in $\mathcal{C}_w$ of dimension at most $d-1$ with weight w_k and let $b \in \mathbb{N}$ be the number of d-faces in $\mathcal{S}$ of weight w_k. Then

$$\operatorname{tw}(\mathcal{S}) = \operatorname{tw}(\mathcal{S}') + (a+b)\cdot w_k = \operatorname{tw}(\mathcal{S}'') - \sum_{i=1}^{m} w(F_i) + (a+b)\cdot w_k \geq$$

$$\geq \operatorname{tw}(\mathcal{S}'') + (a+b-m)\cdot w_k \geq \operatorname{tw}(\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}) + (a+b-m)\cdot w_k = \operatorname{tw}(\widetilde{\mathtt{MST}}^{(d)}_{w_k}),$$

where we still need to justify the final equality. That is, we need to check that we add $a+b-m$ faces when going from $\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}$ to $\widetilde{\mathtt{MST}}^{(d)}_{w_k}$. Since $\widetilde{\mathtt{MST}}^{(d)}_{\alpha}$ is d-spanning at all times, this reduces to checking that $\widetilde{\mathtt{MST}}^{(d)}_{w_k}$ has $b-m$ more d-dimensional faces than $\widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}$.

Refer again to Lemma 2.7(2) to get

$$\begin{aligned}\left(\#d\text{-faces in } \mathcal{C}_{w\leq w_k}\right) + \beta_{d-1}(\mathcal{C}^{(d)}_{w\leq w_k}) - \beta_d(\mathcal{C}^{(d)}_{w\leq w_k}) &= \\ = \left(\#d\text{-faces in } \mathcal{S}\right) + \beta_{d-1}(\mathcal{S}) - \beta_d(\mathcal{S}),\end{aligned}$$

$$\begin{aligned}\left(\#d\text{-faces in } \mathcal{C}_{w\leq w_{k-1}}\right) + \beta_{d-1}(\mathcal{C}^{(d)}_{w\leq w_{k-1}}) - \beta_d(\mathcal{C}^{(d)}_{w\leq w_{k-1}}) &= \\ = \left(\#d\text{-faces in } \mathcal{S}'\right) + \beta_{d-1}(\mathcal{S}') - \beta_d(\mathcal{S}').\end{aligned}$$

This reduces to

$$\left(\#d\text{-faces in } \mathcal{C}_{w\leq w_k}\right) - \beta_d(\mathcal{C}^{(d)}_{w\leq w_k}) = \left(\#d\text{-faces in } \mathcal{S}\right),$$

$$\left(\#d\text{-faces in } \mathcal{C}_{w\leq w_{k-1}}\right) + \beta_{d-1}(\mathcal{C}^{(d)}_{w\leq w_{k-1}}) - \beta_d(\mathcal{C}^{(d)}_{w\leq w_{k-1}}) = \left(\#d\text{-faces in } \mathcal{S}'\right) + \beta_{d-1}(\mathcal{S}').$$

Hence

$$m = \beta_{d-1}(\mathcal{S}') - \beta_{d-1}(\mathcal{C}^{(d)}_{w\leq w_{k-1}}) =$$

$$= b + \Big(\beta_d(\mathcal{C}^{(d)}_{w\le w_k}) - \beta_d(\mathcal{C}^{(d)}_{w\le w_{k-1}})\Big) - \Big(\#d\text{-faces in } \mathcal{C}_{w\le w_k} - \#d\text{-faces in } \mathcal{C}_{w\le w_{k-1}}\Big),$$

so

$$b - m = \Big(\#d\text{-faces in } \mathcal{C}_{w\le w_k} - \#d\text{-faces in } \mathcal{C}_{w\le w_{k-1}}\Big) - \Big(\beta_d(\mathcal{C}^{(d)}_{w\le w_k}) - \beta_d(\mathcal{C}^{(d)}_{w\le w_{k-1}})\Big)$$

which by the calculation for $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ in the fittingness part of the proof above equals

$$\Big(\#d\text{-faces in } \widetilde{\mathtt{MST}}^{(d)}_{w_k}\Big) - \Big(\#d\text{-faces in } \widetilde{\mathtt{MST}}^{(d)}_{w_{k-1}}\Big).$$

■

We claim that minimal spanning trees (as given by Definition 3.1) are precisely the complexes, obtained in Algorithm 3.2, at their final stage.

Lemma 3.5 (Correctness of Algorithm 3.2) *Let $\mathcal{C}_w$ be a weighted simplex and $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ as given by Algorithm 3.2.*

1. *For $\alpha \in \mathbb{R}_{\ge w(\mathcal{C})}$ the complex $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ is a minimal spanning d-tree of $\mathcal{C}_w$. Denoting $\mathtt{MST}^{(d)} := \widetilde{\mathtt{MST}}^{(d)}_{w(\mathcal{C})}$, we have $\mathtt{MST}^{(d)}_\alpha = \widetilde{\mathtt{MST}}^{(d)}_\alpha$ for all $\alpha \in \mathbb{R}$.*

2. *Every minimal spanning d-tree of $\mathcal{C}_w$ is of the form $\widetilde{\mathtt{MST}}^{(d)}_{w(\mathcal{C})}$, obtained via Algorithm 3.2.*

Proof.

1. Use Lemma 3.4 for $\alpha \ge w(\mathcal{C})$ while noting that in this case $\mathcal{C}_{w\le\alpha}$ is the whole simplex, so has the homology of a point.

2. Let $\mathtt{MST}^{(d)}$ be any minimal spanning tree. We get $\mathtt{MST}^{(d)}_\alpha = \widetilde{\mathtt{MST}}^{(d)}_\alpha$ for all $\alpha \in \mathbb{R}$ if we choose the order of d-faces at any weight w_k to start with the d-faces in $\mathtt{MST}^{(d)}$, followed by those not in $\mathtt{MST}^{(d)}$. It is clear from Algorithm 3.2 that $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ includes all d-faces of weight w_k in $\mathtt{MST}^{(d)}_\alpha$. To get the converse, note that $\widetilde{\mathtt{MST}}^{(d)}_\alpha$ and $\mathtt{MST}^{(d)}_\alpha$ (both of which are $(d-1)$-fitting d-spanning d-forests) have the same number of d-faces at every stage by Lemma 2.7(2).

■

Remark 3.6 We conclude that Algorithm 3.2 yields a minimal spanning tree. The general idea of the algorithm was to take the necessary amount of d-faces in the tree (the exact number is given by Proposition 2.8(2)) while choosing first among lighter faces, so the greedy algorithm works (as one would anticipate from matroid theory).

Theorem 3.7 (Optimality of Minimal Spanning d-Trees) *For every minimal spanning tree* $\mathtt{MST}^{(d)}$ *of a weighted simplex* $\mathcal{C}_w$ *and every* $\alpha \in \mathbb{R}$ *the subcomplex* $\mathtt{MST}^{(d)}_\alpha$ *is a* $(d-1)$*-fitting d-spanning d-forest in* $\mathcal{C}_{w\leq\alpha}$*, and moreover has minimal total weight among all* $(d-1)$*-fitting d-spanning d-forests in* $\mathcal{C}_{w\leq\alpha}$.

Proof. By Lemma 3.5(2) and Lemma 3.4. ∎

4 Homologically Persistent d-Skeleton

We proved in Theorem 3.7 that homology of the minimal spanning d-tree matches the homology of a weighted simplex up to dimension $d-1$ for all parameter values. The purpose of the homologically persistent skeleton is to add and remove d-faces, called critical d-faces, in a way that ensures an isomorphism of homology groups in dimension d as well.

4.1 Critical Faces of a Weighted Simplex

Fix a minimal spanning d-tree $\mathtt{MST}^{(d)}$ of a weighted simplex $\mathcal{C}_w$.

Definition 4.1 A d-face K of $\mathcal{C}_w$ is *critical* when K is not in $\mathtt{MST}^{(d)}$.

In order to obtain isomorphisms on the level of homology in Theorem 4.12, critical faces play a crucial role as generators of homology (at all stages $\alpha \in \mathbb{R}$). However, a critical face might contribute to many nontrivial cycles, so the connection between critical d-faces and generators in $H_d(\mathcal{C}_{w\leq\alpha})$ is not canonical in general. We resolve this issue by using relative homology.

Lemma 4.2 *Let* $\alpha \in \mathbb{R}$ *and let* $\mathcal{S}$ *be a subcomplex of* $\mathcal{C}_{w\leq\alpha}$ *which contains* $\mathtt{MST}^{(d)}_\alpha$*. Let* $K_1, K_2, \ldots, K_m$ *be the critical d-faces in* $\mathcal{S}$*.*

1. *Each K_i represents a relative homology class $[K_i] \in H_d\left(\mathcal{S}, \mathtt{MST}_\alpha^{(d)}\right)$.*

2. *The classes $[K_1], [K_2], \ldots, [K_m]$ generate $H_d\left(\mathcal{S}, \mathtt{MST}_\alpha^{(d)}\right)$.*

3. *If $\mathcal{S}$ is a d-complex, the classes $[K_1], [K_2], \ldots, [K_m]$ form a basis of $H_d\left(\mathcal{S}, \mathtt{MST}_\alpha^{(d)}\right)$.*

Proof.

1. By assumption K_i is in $\mathcal{S}$. The boundary of K_i is in the $(d-1)$-skeleton of $\mathcal{C}_{w\le\alpha}$ and thus also in $\mathtt{MST}_\alpha^{(d)}$, meaning that K_i is a relative d-cycle. Hence $[K_i] \in H_d\left(\mathcal{S}, \mathtt{MST}_\alpha^{(d)}\right)$.

2. Take any $[z] \in H_d\left(\mathcal{S}, \mathtt{MST}_\alpha^{(d)}\right)$. We write $z = \sum_i c_i F_i$, where $c_i \in \mathbb{F}$ and F_is are d-faces of $\mathcal{S}$. Whenever F_i is in the minimal spanning tree, $[F_i] = 0$ in $H_d\left(\mathcal{S}, \mathtt{MST}_\alpha^{(d)}\right)$, which implies that $[z] = \sum_{i, F_i \notin \mathtt{MST}_\alpha^{(d)}} c_i [F_i]$. The class $[z]$ can therefore be expressed as a linear combination of classes represented by critical faces.

3. Suppose we have $\sum_{i=1}^m c_i [K_i] = 0$; then $\left[\sum_{i=1}^m c_i K_i\right] = 0$. This means there exist $v \in C_{d+1}(\mathcal{S})$ and $u \in C_d\left(\mathtt{MST}_\alpha^{(d)}\right)$ such that

$$\partial_{d+1} v = u + \sum_{i=1}^m c_i K_i.$$

But as a d-complex, $\mathcal{S}$ only has 0 as a $(d+1)$-chain, so we get

$$u + \sum_{i=1}^m c_i K_i = 0.$$

Write $u = \sum_{j=1}^k d_j F_j$ where F_js are d-faces in $\mathtt{MST}_\alpha^{(d)}$. Thus

$$\sum_{j=1}^k d_j F_j + \sum_{i=1}^m c_i K_i$$

is the zero chain, and since d-faces form a basis of the space of d-chains, all the coefficients must be zero. In particular, c_is are zero, proving the desired linear independence.

∎

To build the homologically persistent skeleton associated to $\mathcal{C}_w$, we add and remove critical d-faces, to which we assign birth and death times inductively (Subsection 4.2). The following lemma guarantees that for $d > 0$ all homology classes generated by critical faces eventually die.

Lemma 4.3 *For any $\alpha \in \mathbb{R}_{\geq w(\mathcal{C})}$ we have*

$$H_d\left(\mathcal{C}_{w\leq\alpha}, \mathtt{MST}^{(d)}_\alpha\right) \cong H_d\left(\mathcal{C}_w, \mathtt{MST}^{(d)}\right) \cong H_d\left(\mathcal{C}_w\right) \cong \begin{cases} 0 & \text{if } d > 0, \\ \mathbb{F} & \text{if } d = 0. \end{cases}$$

Proof. The first isomorphism is obvious. So is the last one, since $\mathcal{C}_w$ is contractible (it is a simplex). As for the middle one, consider first the case $d \geq 2$. In the long exact sequence of a pair

$$\ldots \to H_d\left(\mathtt{MST}^{(d)}\right) \to H_d\left(\mathcal{C}_w\right) \to H_d\left(\mathcal{C}_w, \mathtt{MST}^{(d)}\right) \to H_{d-1}\left(\mathtt{MST}^{(d)}\right) \to \ldots$$

we have $H_d\left(\mathtt{MST}^{(d)}\right) = H_{d-1}\left(\mathtt{MST}^{(d)}\right) = 0$, so we get the desired isomorphism. If $d = 1$, the map $H_0\left(\mathtt{MST}^{(d)}\right) \to H_0\left(\mathcal{C}_w\right)$, induced by the inclusion, is an isomorphism $\mathbb{F} \cong \mathbb{F}$, so the boundary map $H_1\left(\mathcal{C}_w, \mathtt{MST}^{(d)}\right) \to H_0\left(\mathtt{MST}^{(d)}\right)$ is zero. Hence $H_1\left(\mathcal{C}_w\right) \to H_1\left(\mathcal{C}_w, \mathtt{MST}^{(d)}\right)$ is surjective. It is also injective since $H_1\left(\mathtt{MST}^{(d)}\right) = 0$. If $d = 0$, then $\mathtt{MST}^{(d)} = \varnothing$, so $H_0\left(\mathcal{C}_w\right) \cong H_0\left(\mathcal{C}_w, \mathtt{MST}^{(d)}\right)$. ∎

4.2 Birth and Death of a Critical Face

For each critical d-face we define its *birth time* (or simply *birth*) to be its weight. We wish to define the death time of a critical face as the parameter value at which the homology generator it created dies, however, it can happen that multiple critical faces enter at the same time. In that case assigning death times correctly is critical for Theorem 4.12 to hold.

Example 4.4 Consider the point cloud depicted in Figure 5 for $d = 1$. The only two critical 1-faces, which do not immediately die, are depicted in red (they appear at the parameter value 1). The generators they create die at times $\frac{5}{4}$ and $\frac{\sqrt{13}}{2}$, but since they are born at the same time, the question that arises is which death time to associate to which critical face. It turns out that for Theorem 4.12 to hold, the choice of assignments in Figure 5 is the only valid one. However, that does not mean that we never have any freedom of assigning death times. A minor change in the example (see Figure 6) allows us two possibilities, both valid for Theorem 4.12.

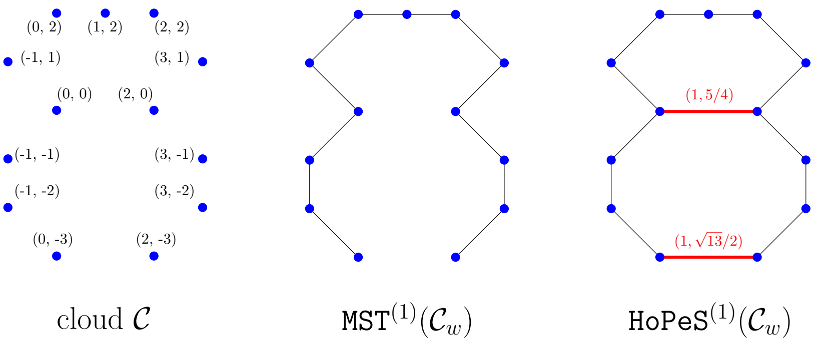


Figure 5: Cloud $\mathcal{C}$ whose simplex $\mathcal{C}_w$ has weights from its Čech complex, its minimal spanning 1-tree and its homologically persistent 1-skeleton.

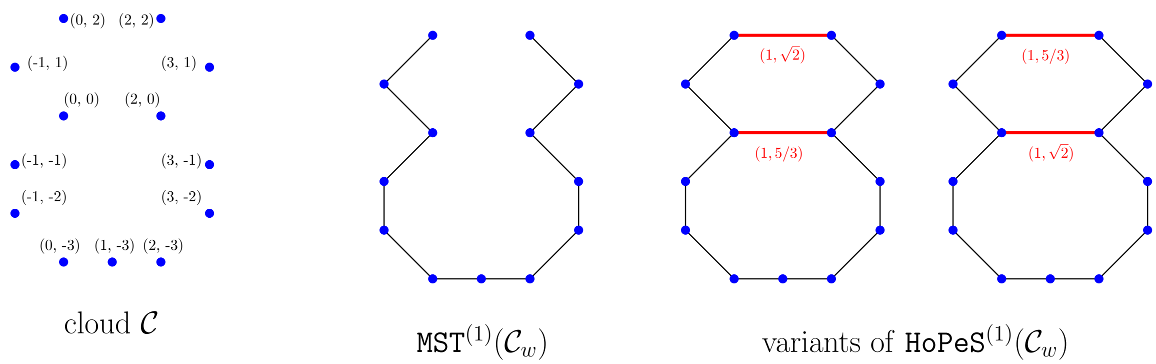


Figure 6: Cloud $\mathcal{C}$ whose simplex $\mathcal{C}_w$ has weights from its Čech complex, its minimal spanning 1-tree and two possible homologically persistent 1-skeleta.

As Example 4.4 shows, we need to know when and to what extent the assignment of death times is determined. We describe an algorithm, which assigns death times to all critical d-faces and determines exactly how much freedom we have for these assignments.

Deaths can only occur at times when a simplex is added to $\mathcal{C}_{w\leq\alpha}$, i.e. at values in the image $\mathrm{im}(w)$. We go through $\mathrm{im}(w)$ with α in increasing order and for each such $\alpha \in \mathrm{im}(w)$ decide which (if any) critical d-faces die at α.

Definition 4.5 (Deaths of Critical Faces) Define $\widetilde{\mathcal{K}}_\alpha \coloneqq \{K_1, K_2, \ldots, K_s\}$ to be the set of critical d-faces born before or at α that have not yet been assigned a death time. By Lemma 4.2 the classes $[K_1], [K_2], \ldots, [K_s]$ form a basis of $H_d\Big((\mathtt{MST}_\alpha^{(d)} \cup \widetilde{\mathcal{K}}_\alpha, \mathtt{MST}_\alpha^{(d)})\Big)$. Denote

$$f \coloneqq H_d\Big((\mathtt{MST}_\alpha^{(d)} \cup \widetilde{\mathcal{K}}_\alpha, \mathtt{MST}_\alpha^{(d)}) \hookrightarrow (\mathcal{C}_{w\le\alpha}, \mathtt{MST}_\alpha^{(d)})\Big)$$

and set $r \coloneqq \dim\ker(f)$. Choose a basis $\{b_1, b_2, \ldots, b_r\}$ of $\ker(f)$. We can expand each basis vector as

$$b_i = \sum_{j=1}^{s} c_{ij}[K_j]$$

with $c_{ij} \in \mathbb{F}$. Consider the system of equations (in the field $\mathbb{F}$)

$$\sum_{j=1}^{s} c_{ij}x_j = 0$$

for $i \in \mathbb{N}_{[1,r]}$. Since basis elements are linearly independent, so are these equations. Thus there are r leading variables, for which the system may be solved, expressing them with the remaining $s-r$ free variables. Let $I \subseteq \mathbb{N}_{[1,s]}$ be the set (possibly empty, if $\ker(f)$ is trivial) of indices of leading variables. For each $i \in I$ we declare that the *death time* of the critical face K_i is α.

Depending on the system of equations, we might have many possible choices, which variables to choose as the leading ones. No further restriction on this choice is necessary for Theorem 4.12(1) (fittingness), but to get the rest of the theorem (optimality), we need to further insist on the *elder rule* (compare with the elder rule for the construction of the persistence diagram [8, page 151]): among all available choices for the set of leading variables, choose the one with the largest total weight. There may be more than one set of possible leading variables with the maximal total weight — this is the amount of freedom we have when choosing death times.

If $d \geq 1$, this process assigns death times to all critical d-faces: if any are still left at $\alpha = w(\mathcal{C})$, they all die at that time since $H_d\Big(\mathcal{C}_{w\le w(\mathcal{C})}, \mathtt{MST}_{w(\mathcal{C})}^{(d)}\Big) = 0$ by Lemma 4.3. However, if $d = 0$, $H_d\Big(\mathcal{C}_{w\le w(\mathcal{C})}, \mathtt{MST}_{w(\mathcal{C})}^{(d)}\Big)$ is 1-dimensional rather than 0-dimensional. As such, we declare the death time of the final critical 0-face to be ∞. This makes sense: critical 0-faces (i.e. vertices) die as the complex becomes more and more connected, but in the end a single connected component endures indefinitely.

Here is the summary of this procedure, given as an explicit algorithm.

Algorithm 4.6 Death times of critical d-faces

death$(K) \coloneqq \infty$ for all $K \in \mathcal{C}_w^{(d)} \setminus \mathtt{MST}^{(d)}$
$w_1, w_2, \ldots, w_n \coloneqq$ elements of $\mathrm{im}(w)$, in order
for $l = 1$ **to** n **do**
  $\widetilde{\mathcal{K}}_\alpha \coloneqq \{K_1, K_2, \ldots, K_s\} \coloneqq \left\{K \in \mathcal{C}_{w\le w_l}^{(d)} \setminus \mathtt{MST}_{w_l}^{(d)} \;\middle|\; \mathrm{death}(K) = \infty\right\}$
  $f \coloneqq H_d\Big((\mathtt{MST}_{w_l}^{(d)} \cup \widetilde{\mathcal{K}}_\alpha, \mathtt{MST}_{w_l}^{(d)}) \hookrightarrow (\mathcal{C}_{w\le w_l}, \mathtt{MST}_{w_l}^{(d)})\Big)$
  $\{b_1, b_2, \ldots, b_r\} \coloneqq$ a choice of a basis of $\ker(f)$
  for $i = 1$ **to** r **do**
    for $j = 1$ **to** s **do**
      $c_{ij} \coloneqq$ coefficient at $[K_j]$ in the expansion of b_i
    end for
  end for
  $I \coloneqq$ a choice of an r-element subset of $\mathbb{N}_{[1,s]}$, such that
  - the system $\left(\sum_{j=1}^s c_{ij} x_j = 0\right)_{i\in\mathbb{N}_{[1,r]}}$ is solvable on variables $\{x_j \mid j \in I\}$,
  - the total weight of $\{K_j \mid j \in I\}$ is maximal among such subsets

  death$(K_j) \coloneqq w_l$ for all $j \in I$
end for

For any critical d-face K define its *lifespan* to be death(K) − birth(K). It is possible for a critical d-face K to have the lifespan 0, if the homology class $[K]$ gets killed by some $(d+1)$-face(s) that have the same weight as K.

Lemma 4.7 *For any $\alpha \in \mathbb{R}$ define*

$$\mathcal{K}_\alpha \coloneqq \{K \text{ critical } d\text{-face} \mid \mathit{birth}(K) \le \alpha < \mathit{death}(K)\}.$$

The classes, represented by faces in $\mathcal{K}_\alpha$, form a basis of $H_d\Big(\mathcal{C}_{w\le\alpha}, \mathtt{MST}_\alpha^{(d)}\Big)$.

Proof. It suffices to check this for $\alpha \in \mathrm{im}(w)$. We know that $[K]$s with birth$(K) \le \alpha$ generate $H_d\Big(\mathcal{C}_{w\le\alpha}, \mathtt{MST}_\alpha^{(d)}\Big)$ by Lemma 4.2. We need to check that $[K]$ represented by a critical face, which is dead at α, can be expressed by those still living at α. We prove this inductively on decreasing death times. Let $\delta \le \alpha$ be the death time of K. By Definition 4.5 we can write

$$[K] = \sum_{j=1}^s c_j [K_j]$$

in $H_d\left(\mathcal{C}_{w\leq\delta}, \mathtt{MST}_\delta^{(d)}\right)$, where $\mathcal{K}_\delta = \{K_1, K_2, \ldots, K_s\}$, i.e. death times of K_j are larger than δ. By applying $H_d\left((\mathcal{C}_{w\leq\delta}, \mathtt{MST}_\delta^{(d)}) \hookrightarrow (\mathcal{C}_{w\leq\alpha}, \mathtt{MST}_\alpha^{(d)})\right)$ we can see that this equation also holds in $H_d\left(\mathcal{C}_{w\leq\alpha}, \mathtt{MST}_\alpha^{(d)}\right)$. By the induction hypothesis all of these $[K_j]$ can be expressed by the still living critical faces and therefore, so can $[K]$.

As for linear independence, redefine $K_1, \ldots, K_s$ to be all the faces in $\mathcal{K}_\alpha$. Assume that $\sum_{j=1}^s c_j[K_j] = 0$ in $H_d\left(\mathcal{C}_{w\leq\alpha}, \mathtt{MST}_\alpha^{(d)}\right)$. This implies that

$$\sum_{j=1}^{s} c_j[K_j] \in \ker H_d\left((\mathtt{MST}_\alpha^{(d)} \cup \widetilde{\mathcal{K}}_\alpha, \mathtt{MST}_\alpha^{(d)}) \hookrightarrow (\mathcal{C}_{w\leq\alpha}, \mathtt{MST}_\alpha^{(d)})\right).$$

By assumption none of $[K_j]$s die at α, so this kernel is trivial, meaning $\sum_{j=1}^s c_j[K_j] = 0$ in $H_d\left(\mathtt{MST}_\alpha^{(d)} \cup \widetilde{\mathcal{K}}_\alpha, \mathtt{MST}_\alpha^{(d)}\right)$. Since $[K_j]$s form a basis of this homology (Lemma 4.2), the coefficients c_j are zero. ∎

4.3 Optimality of a Homologically Persistent d-Skeleton

We continue following the blueprint from [15] where the homologically persistent 1-skeleton was constructed by taking a minimal spanning 1-tree and adding labeled critical edges. However, we find it more convenient to have all simplices in the homologically persistent skeleton to be of the same type, so we shall label *all* faces. Define a *label* to be a pair $(l, r) \in \overline{\mathbb{R}} \times \overline{\mathbb{R}}$ such that $0 \leq l < r$. Call l the *left label* and r the *right label*.

Definition 4.8 (homologically persistent skeleton) Given $d \in \mathbb{N}$ and a weighted simplex $\mathcal{C}_w$, its *homologically persistent d-skeleton* $\mathtt{HoPeS}^{(d)}(\mathcal{C}_w)$ is the (choice of a) minimal spanning d-tree together with all critical d-faces with positive lifespan:

$$\mathtt{HoPeS}^{(d)}(\mathcal{C}_w) := \mathtt{MST}^{(d)} \cup \left\{K \in \mathcal{C}_w^{(d)} \setminus \mathtt{MST}^{(d)} \,\middle|\, \mathrm{birth}(K) < \mathrm{death}(K)\right\}.$$

Each face F in $\mathtt{HoPeS}^{(d)}(\mathcal{C}_w)$ is labeled: if F is in $\mathtt{MST}^{(d)}$, by $\left(w(F), \infty\right)$; otherwise by $\left(\mathrm{birth}(F), \mathrm{death}(F)\right)$. We write simply $\mathtt{HoPeS}^{(d)}$ instead of $\mathtt{HoPeS}^{(d)}(\mathcal{C}_w)$ when there is no ambiguity.

Note that the set of labels $\{(l,r) \in \overline{\mathbb{R}} \times \overline{\mathbb{R}} \mid 0 \leq l < r\}$ can be seen as a form of an interval domain [16]. In particular, we have the *information order* $\sqsubseteq$, given by

$$(l', r') \sqsubseteq (l'', r'') \quad :\Leftrightarrow \quad l' \leq l'' \ \wedge \ r' \geq r''.$$

Labeling of $\mathtt{HoPeS}^{(d)}$ is monotone in the following sense. Let F and G be faces in $\mathtt{HoPeS}^{(d)}$ with labels ℓ_F and ℓ_G respectively. If $F \subseteq G$, then $\ell_F \sqsubseteq \ell_G$.

This means that $\mathtt{HoPeS}^{(d)}$ is a kind of a 'weighted complex' itself — except that instead of the weighting mapping into $\mathbb{R}_{\geq 0}$ with its usual order $\leq$, it maps into the interval domain of labels, equipped with the information order. The consequence is that we can define the reduced version of the homologically persistent skeleton for any $\alpha \in \mathbb{R}$:

$$\mathtt{HoPeS}^{(d)}_\alpha(\mathcal{C}_w) := \Big\{\big(F, (l,r)\big) \in \mathtt{HoPeS}^{(d)}(\mathcal{C}_w) \;\Big|\; l \leq \alpha < r\Big\}.$$

As usual, we shorten $\mathtt{HoPeS}^{(d)}_\alpha(\mathcal{C}_w)$ to $\mathtt{HoPeS}^{(d)}_\alpha$ when there is no ambiguity. Due to monotonicity of labeling, $\mathtt{HoPeS}^{(d)}_\alpha$ is a (labeled) simplicial complex.

Example 4.9 Let $\mathcal{C}$ be a point cloud from Example 3.3. In the example from Figure 7 the complexes $\mathtt{HoPeS}^{(d)}_\alpha$ do not differ all that much from $\mathcal{C}_w$ for most α. This is due to the fact that, pictorially, we are restricted to relatively small, low-dimensional complexes. The true potential of homologically persistent skeleton lies in working with very large, high-dimensional complexes.

Lemma 4.10 *$H_d\Big(\mathtt{HoPeS}^{(d)}_\alpha \hookrightarrow \mathcal{C}_{w\leq\alpha}\Big)$ is an isomorphism for any $\alpha \in \mathbb{R}$.*

Proof. By Lemma 2.7(5) the map $H_d\Big(\mathtt{HoPeS}^{(d)}_\alpha \hookrightarrow \mathcal{C}_{w\leq\alpha}\Big)$ is an isomorphism if and only if the map $H_d\Big((\mathtt{HoPeS}^{(d)}_\alpha, \mathtt{MST}^{(d)}_\alpha) \hookrightarrow (\mathcal{C}_{w\leq\alpha}, \mathtt{MST}^{(d)}_\alpha)\Big)$ is. But that follows immediately from Lemma 4.2(3) and Lemma 4.7. ■

Lemma 4.11 *Take any $\alpha \in \mathbb{R}$ and any d-fitting d-spanning d-subcomplex $\mathcal{S}$ in $\mathcal{C}_{w\leq\alpha}$. By Lemma 2.7(4) $\mathcal{S}$ contains a d-subcomplex which is a $(d-1)$-fitting d-spanning d-forest; let F denote one with minimal total weight.*

1. *The number of d-faces in $\mathcal{S}$ is*

$$\Big(\#d\text{-faces in } \mathcal{C}_{w\leq\alpha}\Big) - \beta_d(\mathcal{C}^{(d)}_{w\leq\alpha}) + \beta_d(\mathcal{C}_{w\leq\alpha}).$$

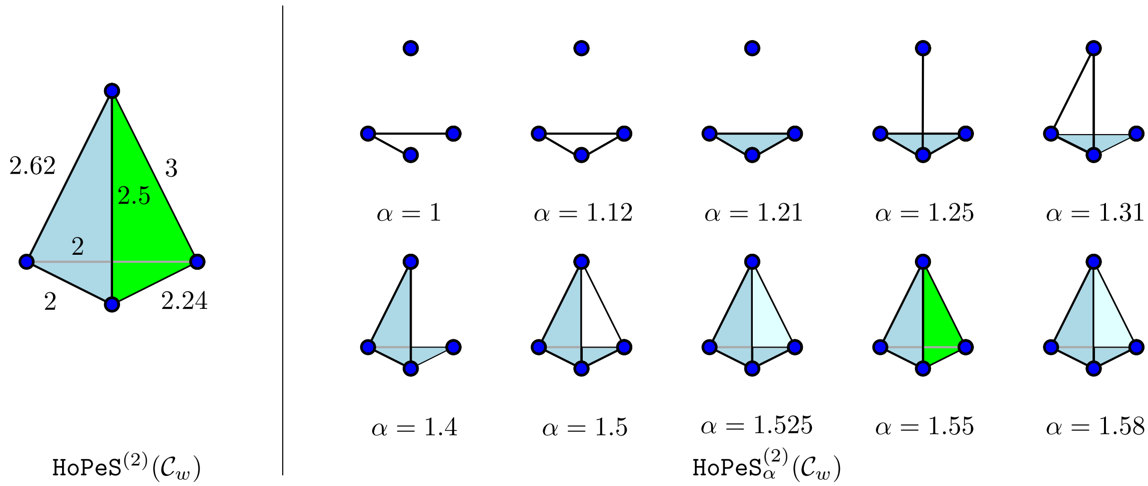


Figure 7: Geometric realizations of $\mathtt{HoPeS}^{(2)}(\mathcal{C}_w)$ and its reduced versions with respect to Čech filtration of a point cloud $\mathcal{C}$ with four vertices. $\mathtt{HoPeS}^{(2)}(\mathcal{C}_w)$ consists of the boundary of the tetrahedron, its only critical face marked by green. The remaining 2-faces are a part of the minimal spanning tree (cf. Figure 4).

The number of d-faces in F is

$$\left(\#d\text{-faces in } \mathcal{C}_{w\le\alpha}\right) - \beta_d(\mathcal{C}^{(d)}_{w\le\alpha}).$$

Consequently, the number of d-faces in $\mathcal{S} \smallsetminus F$ is equal to $\beta_d(\mathcal{C}_{w\le\alpha})$.

2. The diagram

$$\begin{array}{ccc}
 & H_d\left((\mathcal{S},\varnothing) \hookrightarrow (\mathcal{S},F)\right) & \\
H_d(\mathcal{S}) & \longrightarrow & H_d(\mathcal{S},F) \\
\Big\downarrow {\scriptstyle H_d\left(\mathcal{S} \hookrightarrow \mathcal{C}_{w\le\alpha}\right)} & & \Big\downarrow {\scriptstyle H_d\left((\mathcal{S},F) \hookrightarrow (\mathcal{C}_{w\le\alpha},F)\right)} \\
H_d(\mathcal{C}_{w\le\alpha}) & \longrightarrow & H_d(\mathcal{C}_{w\le\alpha},F) \\
 & H_d\left((\mathcal{C}_{w\le\alpha},\varnothing) \hookrightarrow (\mathcal{C}_{w\le\alpha},F)\right) &
\end{array}$$

commutes and all maps in it are isomorphisms.

3. *The diagram in the previous item induces a bijective correspondence between the set of d-faces in $\mathcal{S} \smallsetminus F$ and the set of dots (p,q) in the persistence diagram $\mathrm{PD}_d(\mathcal{C}_w)$ with $p \le \alpha < q$. If a d-face S is associated to the dot (p,q), then $p \le w(S)$.*

Proof.

1. Apply Lemma 2.7(2) for $\mathcal{S}$ and F (as subcomplexes of $\mathcal{C}_{w\le\alpha}$) and take their properties into account.

2. Use Lemma 2.7(5) and the assumption that $\mathcal{S}$ is d-fitting in $\mathcal{C}_{w\le\alpha}$.

3. Denote

$$\begin{aligned} f \;&:=\; H_d\left(\mathcal{S} \hookrightarrow \mathcal{C}_{w\le\alpha}\right) \circ \left(H_d\left((\mathcal{S},\varnothing) \hookrightarrow (\mathcal{S},F)\right)\right)^{-1} = \\ &= \left(H_d\left((\mathcal{C}_{w\le\alpha},\varnothing) \hookrightarrow (\mathcal{C}_{w\le\alpha},F)\right)\right)^{-1} \circ H_d\left((\mathcal{S},F) \hookrightarrow (\mathcal{C}_{w\le\alpha},F)\right); \end{aligned}$$

   this is an isomorphism between $H_d(\mathcal{S},F)$ and $H_d(\mathcal{C}_{w\le\alpha})$ by the previous item. Let $S_1,\ldots,S_m$ be d-faces in $\mathcal{S} \smallsetminus F$. By a similar argument as in Lemma 4.2 the classes $[S_i]$ form a basis of $H_d(\mathcal{S},F)$. Hence $f([S_i])$ form a basis of $H_d(\mathcal{C}_{w\le\alpha})$ and are thus in bijective correspondence with dots (p,q) in $\mathrm{PD}_d(\mathcal{C}_w)$ with $p \le \alpha < q$. Let us denote the dot, associated to S_i, by (p_i,q_i).

   Since F has minimal total weight, S_i has the largest weight among faces (with non-zero coefficients) in the cycle which represents $f([S_i])$. Hence the homology class $f([S_i])$ could not be born after $w(S_i)$.

∎

Theorem 4.12 (Fittingness and Optimality of Reduced d-Skeletons) *The following holds for every weighted simplex $\mathcal{C}_w$, $d \in \mathbb{N}$, and $\alpha \in \mathbb{R}$.*

1. $\mathtt{HoPeS}^{(d)}_\alpha$ *is d-fitting in $\mathcal{C}_{w\le\alpha}$.*

2. *For every critical d-face K in $\mathtt{HoPeS}^{(d)}_\alpha$, the dot (p,q) in the persistence diagram, associated to it via the bijective correspondence from Lemma 4.11(3) (for $\mathcal{S} = \mathtt{HoPeS}^{(d)}_\alpha$ and $F = \mathtt{MST}^{(d)}_\alpha$), is the same as the label of K. In particular $p = w(K)$.*

3. $\mathtt{HoPeS}^{(d)}_\alpha$ *has the minimal total weight among all d-fitting d-spanning subcomplexes $\mathcal{S} \subseteq \mathcal{C}_{w\le\alpha}$.*

Proof.

1. Between Lemmas 2.7(1) and 4.10 we only still need to check that the map $H_{d-1}\left(\mathtt{HoPeS}_{\alpha}^{(d)} \hookrightarrow \mathcal{C}_{w\leq\alpha}\right)$ is injective, or equivalently, that its kernel is trivial.

   Let $K_1, K_2, \dots, K_s$ be critical d-faces living at α and let $K_{s+1}, \dots, K_m$ be the remaining critical d-faces born before or at α. Take such a cycle $z \in Z_{d-1}\left(\mathtt{HoPeS}_{\alpha}^{(d)}\right)$ that $[z] = 0$ in $H_{d-1}\left(\mathcal{C}_{w\leq\alpha}\right)$. This means there exists a chain $v \in C_d\left(\mathcal{C}_{w\leq\alpha}\right)$ with $\partial_d v = z$. Write

$$v = \sum_{i=1}^{m} c_i K_i + u$$

   where $u \in C_d\left(\mathtt{MST}_{\alpha}^{(d)}\right)$. Using Lemma 4.7 and unpacking relative homology we can express each K_i with $i > s$ as

$$K_i = \left(\sum_{l=1}^{s} e_l K_l\right) + u_i + \partial_{d+1} t_i$$

   where $u_i \in C_d\left(\mathtt{MST}_{\alpha}^{(d)}\right)$ and $t_i \in C_{d+1}\left(\mathcal{C}_{w\leq\alpha}\right)$. Hence

$$v = \sum_{i=1}^{s} c_i' K_i + u' + \partial_{d+1} t'$$

   for suitable $c_i' \in \mathbb{F}$, $u' \in C_d\left(\mathtt{MST}_{\alpha}^{(d)}\right)$ and $t' \in C_{d+1}\left(\mathcal{C}_{w\leq\alpha}\right)$. Set

$$v' := \sum_{i=1}^{s} c_i' K_i + u',$$

   so $v' \in C_d\left(\mathtt{HoPeS}_{\alpha}^{(d)}\right)$. Then

$$\partial_d v' = \partial_d v' + \partial_d \partial_{d+1} t' = \partial_d v = z.$$

   We conclude that $[z] = 0$ in $H_{d-1}\left(\mathtt{HoPeS}_{\alpha}^{(d)}\right)$.

2. Let $K_1, \dots, K_m$ be critical d-faces in $\mathtt{HoPeS}_{\alpha}^{(d)}$ and for each K_i let (p_i, q_i) be the dot in the persistence diagram $\mathrm{PD}_d(\mathcal{C}_w)$, associated to it. By the assignment of birth and death times of critical faces, as well as the previous item, we see that a cycle representing a homology class associated to K_i (the birth of which is p_i) is born exactly at the time K_i appeared in the homologically persistent skeleton, i.e. at $w(K_i)$.

3. Let $S_1, \dots, S_m$ be d-faces in $\mathcal{S} \setminus F$ and for each S_i let (p_i, q_i) be the dot in the persistence diagram $\mathrm{PD}_d(\mathcal{C}_w)$, associated to it; we have $p_i \leq w(S_i)$ (Lemma 4.11). Taking into account the previous item, we conclude

$$\mathrm{tw}(\mathcal{S}) = \mathrm{tw}(F) + \sum_{i=1}^{m} w(S_i) \geq \mathrm{tw}(F) + \sum_{i=1}^{m} p_i \geq$$

$$\geq \mathrm{tw}(\mathtt{MST}_\alpha^{(d)}) + \sum_{i=1}^{m} p_i = \mathrm{tw}(\mathtt{MST}_\alpha^{(d)}) + \sum_{i=1}^{m} w(K_i) = \mathrm{tw}(\mathtt{HoPeS}_\alpha^{(d)}).$$

■

5 Conclusion

We introduced a d-dimensional homologically persistent skeleton solving the Skeletonization Problem from Subsection 1.1 in an arbitrary dimension d.

- Given a filtration of complexes on a point cloud $\mathcal{C}$, Theorem 3.7(3) proves the optimality of Minimal Spanning d-Trees of the cloud $\mathcal{C}$.
- Definition 4.8 introduces $\mathtt{HoPeS}^{(d)}$ by adding to a Minimal Spanning d-tree all critical d-faces that represent persistent homology d-cycles of $\mathcal{C}_w$, hence $\mathtt{HoPeS}^{(d)}$ visualizes the persistence directly on data.
- For any scale α by Theorem 4.12 the full skeleton $\mathtt{HoPeS}^{(d)}$ contains a reduced subcomplex $\mathtt{HoPeS}_\alpha^{(d)}$, which has a minimal total weight among all d-subcomplexes containing $\mathcal{C}_{w\leq\alpha}^{(d-1)}$ such that the inclusion into $\mathcal{C}_{w\leq\alpha}$ induces isomorphisms in homology in all degrees up to d.

The independence of the Euler characteristic from homology coefficients has helped to prove all results for homology over an arbitrary field $\mathbb{F}$. Do the results (specifically Theorems 3.7 and 4.12) hold over an arbitrary unital commutative ring R? The answer is no, at least not in the form as they are currently stated. Assume that the theorems hold for R. Note that the proof of Lemma 4.2 works for a general R, so $H_d(\mathcal{C}_{w\leq\alpha}; R) \cong H_d\left(\mathtt{HoPeS}_\alpha^{(d)}; R\right) \cong H_d\left(\mathtt{HoPeS}_\alpha^{(d)}, \mathtt{MST}_\alpha^{(d)}; R\right)$ are free R-modules. That is, the results can only work if the homology over R of every finite simplicial complex in every dimension is free. This of course excludes all the usual non-field homology coefficients, including $\mathbb{Z}$.

We have implemented an algorithm computing the homologically persistent skeleton in Mathematica and look forward to collaborating with practitioners working on real data.

References

[1] A. Björner. "Handbook of Combinatorics (Vol. 2)". In: MIT Press, 1995. Chap. Topological Methods, pp. 1819–1872.

[2] E. D. Bolker. "Simplical Geometry and Transportation Polytopes". In: *Transactions of the American Mathematical Society* 217 (1976), pp. 121–142.

[3] G. Carlsson. "Topological pattern recognition for point cloud data". In: *Acta Numerica* 23 (2013), pp. 289–368.

[4] G. Carlsson. "Topology and Data". In: *Bulletin of the American Mathematical Society* 46 (2009), pp. 255–308.

[5] G. Carlsson et al. "On the local behavior of spaces of natural images". In: *International Journal of Computer Vision* 76 (2008), pp. 1–12.

[6] C. J. A. Delfinado and H. Edelsbrunner. "An incremental algorithm for Betti numbers of simplicial complexes". In: *Proceedings of the ninth annual Symposium on Computational geometry, ACM* (1993), pp. 232–239.

[7] A. M. Duval, C. J. Klivans, and J. L. Martin. "Simplicial and Cellular Trees". In: *ArXiv e-prints* (2015).

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

[9] H. Edelsbrunner, D. Letscher, and A. J. Zomorodian. "Topological persistence and simplification". In: *Discrete and Computational Geometry* 28 (2002), pp. 511–533.

[10] F. Chazal, R. Huang, and J. Sun. "Gromov-Hausdorff Approximation of Filament Structure Using Reeb-type Graph". In: *Discrete Computational Geometry* 53 (2015), pp. 621–649.

[11] R. Ghrist. "Barcodes: The persistent topology of data". In: *Bulletin of the American Mathematical Society* 45 (2008), pp. 61–75.

[12] Y. Hiraoka and T. Shirai. "Minimum spanning acycle and lifetime of persistent homology in the Linial-Meshulam process". In: *to appear in Random Structures and Algorithms* (2015).

[13] G. Kalai. "Enumeration of q-acyclic simplicial complexes". In: *Israel Journal of Mathematics* 45.4 (1983), pp. 337–351.

[14] J. B. Kruskal. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem". In: *Proceedings of the American Mathematical Society* 7.1 (1956), pp. 48–50.

[15] V. Kurlin. "A one-dimensional homologically persistent skeleton of an unstructured point cloud in any metric space". In: *Computer Graphics Forum* 34.5 (2015), pp. 253–262.

[16] D. Scott. *Outline of a Mathematical Theory of Computation*. Tech. rep. PRG02. OUCL, 1970, p. 30.

[17] P. Skraba, G. Thoppe, and D. Yogeshwaran. *Randomly weighted d-complexes: minimal spanning acycles and persistence diagrams*. 2017.

[18] E. H. Spanier. *Algebraic topology*. McGraw-Hill Book Co., 1966.

[19] T. K. Dey, F. Fan, and Y. Wang. "Graph induced complex on data points". In: *Proceedings of SoCG*. 2013, pp. 107–116.

[20] T. Lewiner, H. Lopes, and G. Tavares. "Applications of Forman's discrete Morse theory to topology visualization and mesh compression". In: *IEEE Transactions on Visuzalization and Computer Graphics* 10 (2004), pp. 499–508.

[21] X. Ge et al. "Data skeletonization via Reeb graphs". In: *Proceedings of NIPS*. 2011.