

# A Homologically Persistent Skeleton is a fast and robust descriptor for a sparse cloud of interest points and saliency features in noisy 2D images

Vitaliy Kurlin

vitaliy.kurlin@gmail.com, <http://kurlin.org>,

Microsoft Research Cambridge, 21 Station Road, Cambridge CB1 2FB, UK and  
Department of Mathematical Sciences, Durham University, Durham DH1 3LE, UK

**Abstract.** 2D images often contain irregular salient features and interest points with non-integer coordinates. Our skeletonization problem for such a noisy sparse cloud is to summarize the topology of a given 2D cloud across all scales in the form of a graph, which can be used for combining local features into a more powerful object-wide descriptor.

We extend a classical Minimum Spanning Tree of a cloud to a Homologically Persistent Skeleton, which is scale-and-rotation invariant and depends only on the cloud without extra parameters. This graph

- (1) is computable in time  $O(n \log n)$  for any  $n$  points in the plane;
- (2) has the minimum total length among all graphs that span a 2D cloud at any scale and also have most persistent 1-dimensional cycles;
- (3) is geometrically stable for noisy samples around planar graphs.

**Keywords:** skeleton, Delaunay triangulation, persistent homology

## 1 Introduction: problem and overview

Pixel-based 2D images often contain *salient features* represented as points with non-integer coordinates. The resulting unstructured set is an example of a point *cloud*  $C$ , formally a finite metric space with pairwise distances between points.

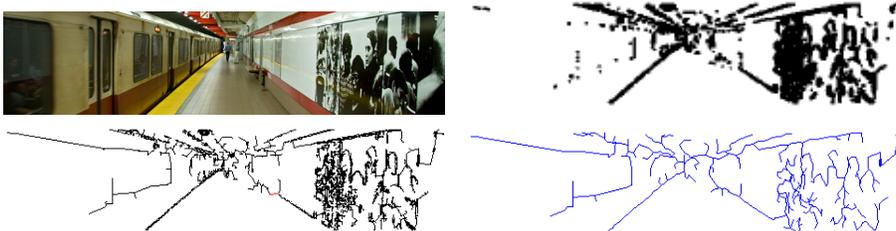
The important problem in low level vision is to extract a meaningful structure from a given irregular cloud  $C$ . The traditional approach is to select a scale parameter, say a radius or the number of neighbors, and build a neighborhood graph. However, a real image may not have a single suitable scale parameter and we need to combine features found at multiple scales. This paper solves the skeletonization problem in its hardest form without any input parameters.

**Parameterless skeletonization for sparse clouds.** Given only an unstructured cloud  $C \subset \mathbb{R}^2$  of points with any real coordinates, find a quickly computable structure that provably represents the topology of  $C$  at all scales.

Our solution is a ‘homological’ extension of a classical Minimum Spanning Tree  $\text{MST}(C)$  of a cloud  $C$  to a *Homologically Persistent Skeleton*  $\text{HoPeS}(C)$  that describes 1-dimensional cycles hidden in  $C$  over all possible scales  $\alpha$ .

In section 2 we explain motivations for building  $\text{HoPeS}(C)$  and give a high level description of our contributions. In section 3 we compare our method with related work. In sections 4–5 we prove that  $\text{HoPeS}(C)$  or its subgraphs are

- **computable** in time  $O(n \log n)$  for a cloud  $C \subset \mathbb{R}^2$  of  $n$  points (Lemma 3)
- **invariant** up to rotations and uniform scale transformations (Lemma 4)
- **optimal** among all graphs capturing cycles of  $C$  at any scale (Theorem 5)
- **stable** under perturbations of samples  $C$  of graphs  $G \subset \mathbb{R}^2$  (Corollary 8).



**Fig. 1.** Top: a cloud  $C$  of feature points. Bottom:  $\text{HoPeS}'(C)$  and its simplification.

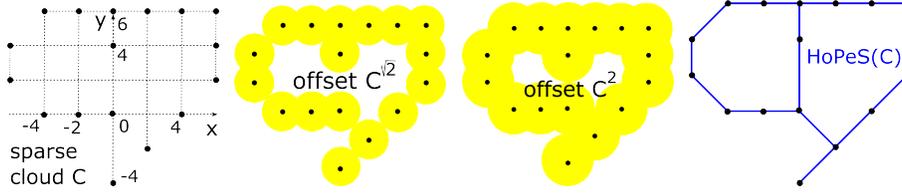
Fig. 1 shows the cloud  $C$  of  $n = 7830$  feature points obtained by thresholding a real image in the top row, see details in section 6. The cloud  $C$  is the *only input* for producing the derived skeleton  $\text{HoPeS}'(C)$  in the bottom row, where we kept only the most persistent cycle. The last picture of Fig. 1 is a simplified version of  $\text{HoPeS}'(C)$  after removing short branches, see Definition 6. So  $\text{HoPeS}'(C)$  provides a best ‘guess’ about the global topology of  $C$  in time  $O(n \log n)$ .

## 2 Our contributions and motivations of $\text{HoPeS}(C)$

Our parameterless skeletonization is based on *persistent homology*, which is the flagship method of Topological Data Analysis [10]. The key idea is to summarize topological features of data over all possible scales. A topological invariant that persists over a long interval of the scale is a true feature of the data, while noisy features have a short life span (a low persistence). The resulting persistent invariants are provably stable under noise, see Theorem 17 in Appendix A.

Fig. 2 shows a cloud  $C$  on the integer lattice for simplicity, though our constructions work for any real coordinates. For any set  $C \subset \mathbb{R}^2$  and  $\alpha > 0$ , the  $\alpha$ -offset  $C^\alpha$  consists of all points in  $\mathbb{R}^2$  that are at most  $\alpha$  away from  $C$ . Here  $\alpha$  is the scale parameter (radius or width) of the  $\alpha$ -offset  $C^\alpha \subset \mathbb{R}^2$  around  $C$ .

We may gradually shrink a disk within itself to its center by making the radius smaller. We can not deform a circle to its center, because a smaller circle would be outside the original circle. So a circle is topologically non-trivial, while any closed loop in a disk is contractible. Spaces connected by such continuous deformations have the same *homotopy type*. We now formalize our problem.



**Fig. 2.** A cloud  $C$ ,  $\alpha$ -offsets  $C^\alpha$  and Homologically Persistent Skeleton  $\text{HoPeS}(C)$

**Multi-scale topological skeletonization:** given a cloud  $C \subset \mathbb{R}^2$ , find a graph whose vertices are all points of  $C$  and whose suitable subgraphs have the homotopy type of the  $\alpha$ -offset  $C^\alpha$  for any  $\alpha$ . A Homologically Persistent Skeleton  $\text{HoPeS}(C)$  is an optimal and stable skeleton satisfying the above requirements.

A cloud  $C$  is an  $\varepsilon$ -sample of ( $\varepsilon$ -close to) a graph  $G \subset \mathbb{R}^2$  if  $G \subset C^\varepsilon$  and  $C \subset G^\varepsilon$ . So any point of  $C$  is at most  $\varepsilon$  away from a point of  $G$  and any point of  $G$  is at most  $\varepsilon$  away from a point of  $C$ . The maximum possible value of  $\varepsilon$  is the upper bound of noise (the *Hausdorff* distance between  $G$  and its sample  $C$ ).

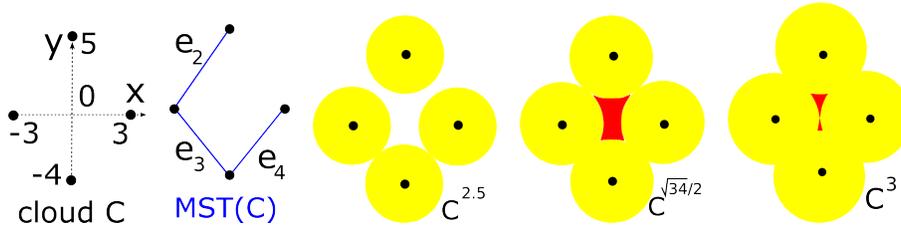
**Here is a high-level description** of our contributions to skeletonization.

- Definition 2 introduces a Homologically Persistent Skeleton  $\text{HoPeS}(C)$  of a cloud  $C \subset \mathbb{R}^2$  summarizing the persistence of 1-dimensional cycles in all  $C^\alpha$ .
- Lemma 3 proves that, for a cloud  $C \subset \mathbb{R}^2$  of any  $n$  points,  $\text{HoPeS}(C)$  has the size  $O(n)$  and is computed in time  $O(n \log n)$  without any extra parameters.
- Lemma 4 shows that  $\text{HoPeS}(C)$  is a scale-and-rotation invariant of  $C \subset \mathbb{R}^2$ .
- Theorem 5 proves that the reduced graph  $\text{HoPeS}(C; \alpha)$  at any scale  $\alpha > 0$  has the minimum length among all graphs that have the homotopy type of  $C^\alpha$ .
- Theorem 7 guarantees that for any  $\varepsilon$ -sample of a simple enough graph  $G \subset \mathbb{R}^2$ ,  $\text{HoPeS}'(C)$  is a correct topological reconstruction of  $G$  in the  $2\varepsilon$ -offset  $G^{2\varepsilon}$ .
- Corollary 8 implies that the derived subgraph  $\text{HoPeS}'(C)$  is stable for any  $\delta$ -perturbation of a cloud  $C$  that was  $\varepsilon$ -sampled around a planar graph  $G$ .

**The novelty of this paper** is not the fast algorithm for 1-dimensional persistence, but the new fundamental concept of a Homologically Persistent Skeleton  $\text{HoPeS}(C)$  that depends only a cloud  $C \subset \mathbb{R}^2$  and solves the skeletonization problem without extra parameters and with guarantees in Theorems 5 and 7.

A graph without cycles is a *forest*. A connected forest is a *tree*. For a cloud  $C \subset \mathbb{R}^2$ , a *Minimum Spanning Tree*  $\text{MST}(C)$  is a tree that has the vertex set  $C$  and the minimum total length of edges, see Fig. 3. The *reduced forest*  $\text{MST}(C; \alpha)$  is obtained from  $\text{MST}(C)$  by removing all open edges longer than  $2\alpha$ .

A connected graph  $G$  *spans* a cloud  $C$  if  $C$  is the vertex set of  $G$ . A graph  $G$  *spans* a possibly disconnected  $\alpha$ -offset  $C^\alpha$  if  $G$  has vertices at all points of the cloud  $C$  and any vertices of  $G$  are in the same connected component of  $G$  if and only if these vertices are in the same connected component of the  $\alpha$ -offset  $C^\alpha$ .



**Fig. 3.** Cloud  $C$ , minimum spanning tree  $\text{MST}(C)$  and  $\alpha$ -offsets  $C^{2.5}$ ,  $C^{\sqrt{34}/2}$ ,  $C^3$ .

Points  $p, q \in C$  are in the same *single-edge cluster* of  $C$  if  $d(p, q) \leq 2\alpha$ . Lemma 1 says that  $\text{MST}(C)$  is a universal optimal object that describes the 0-dimensional topology (all single-edge clusters) of  $C$  across all scales  $\alpha$ .

**Lemma 1** *For a cloud  $C$  and any scale  $\alpha \geq 0$ , the reduced forest  $\text{MST}(C; \alpha)$  has the minimum total length of edges among all graphs that span  $C^\alpha$  at the same scale  $\alpha$ . Hence all connected components of the reduced forest  $\text{MST}(C; \alpha)$  are in a 1-1 correspondence with all single-edge clusters of the cloud  $C$ .*

Lemma 1 and all later results are proved in Appendix B. Theorem 5 extends the optimality of  $\text{MST}(C)$  in Lemma 1 for clusters (dimension 0 approximation of  $C$ ) to the skeleton  $\text{HoPeS}(C)$  for cycles (dimension 1 approximation of  $C$ ).

### 3 Comparison with related past skeletonization work

Our approach may look similar to the well-known scale-space theory [15] that suggests how to find a suitable scale. However, we do not choose any scale, we find topological features with longest life spans, which may not overlap. For instance, if one feature lives over the scale interval  $1 \leq \alpha \leq 2$  and another over  $3 \leq \alpha \leq 4$ , then both features can not be captured at any fixed scale  $\alpha$ . We can capture both features only by analyzing their life spans among all features.

The classical scale selection relies on analyzing data at discrete scales, usually proportional to powers of 2. The persistent homology works over the continuous scale so that all critical scales are found only from a given cloud, not by manually selecting a step size for incrementing the scale. Though we wouldn't say that persistent homology is 'perpendicular' to scale-space theory, our method is at least 'diagonal' to a scale selection, see diagonal gaps in Definition 6.

To the best of our knowledge, all known skeletonization algorithms for clouds need extra parameters such as a scale  $\alpha$  or a noise bound  $\varepsilon$ . Hence all these algorithms can not run on our minimal input, which is only a cloud  $C$ . Since a manual choice of parameters can be unfair, the experimental comparison with the past work seems impossible and we can compare only theoretical aspects.

N. Cornea et al. [6] stated the following requirements for skeletonization.

- *Topology*: a skeleton found from a noisy sample  $C$  is homeomorphic to (or has the homotopy type of) the original shape as in Theorem 7 from section 5.
- *Centering*: if a shape is well-sampled, a skeleton geometrically approximates the original shape in a small offset, see the  $2\varepsilon$ -offset guarantee in Theorem 7.
- *Efficiency*: a near linear time in the number  $n$  of points as in our Lemma 3.

Our skeleton  $\text{HoPeS}(C)$  satisfies the extra conditions: independence of extra parameters, rotation-and-scale invariance and stability under bounded noise.

R. Singh et al. [16] approximated a skeleton of a shape by a subgraph of a Delaunay triangulation based on 2nd order Voronoi regions. The algorithm has 3 threshold parameters:  $K$  for the minimum number of edges in a cycle and  $\delta_{min}, \delta_{max}$  for inserting and merging Voronoi regions. M. Aanjaneya et al. [1] solved a related problem approximating a metric on a large input graph by a metric on a small output graph. So the input is a graph, not a cloud of points.

Starting from a noisy sample of an unknown graph  $G$  with a scale parameter, X. Ge et al. [11] produced the Reeb graph with the same number of loops as the graph  $G$ . This output is an abstract graph of a simplicial complex on a cloud  $C$  and is not intrinsically embedded into any space even if  $C \subset \mathbb{R}^2$ . [11, section 3.3] reported ‘spurious branches or loops in the Reeb graph constructed no matter how we choose a radius or a number of neighbours to decide the scale’.

F. Chazal and J. Sun [4] estimated a distance between an unknown graph  $X$  and a new  $\alpha$ -Reeb graph that was obtained from a noisy sample of  $X$ . Their algorithm has the same fast time  $O(n \log n)$ , but is not directly comparable with ours, because the  $\alpha$ -Reeb graph needs a manually chosen scale  $\alpha$ .

T. Dey et al. [8] built a complex depending on a user-defined graph that spans a cloud  $C$  of  $n$  points. This Graph Induced Complex GIC has the same homology  $H_1$  as the Rips complex of a cloud  $C$  at a suitable scale  $\alpha$ . The 2D skeleton of GIC needed for computing  $H_1$  has the size  $O(n^3)$  in a worst case.

For image segmentation,  $\alpha$ -offsets were similarly used in [14] (with 2 extra parameters) and in [13] (without parameters). These segmentations miss all branches of a graph given by a sample, so our skeletonization problem is harder.

Papers	[11], 2011	[1], 2012	[8], 2013	[4], 2014	this paper
Extra input	radius $r$	radius $r$ , noise $\varepsilon$	graph spanning $C$	scale $\alpha$	no parameters
Complexity	$O(n \log n)$	$O(n^2)$	at least $O(n^3)$	$O(n \log n)$	$O(n \log n)$

**Table 1.** Comparison of similar skeletonization methods for unstructured clouds.

The discussion in [2, section 13] proposed to select features by persistence [10] and led us to the new concept of  $\text{HoPeS}(C)$  in Definition 2. The key advantage of our approach over the past work is the *absence of any user-defined parameters*.

- $\text{HoPeS}(C)$  of a cloud  $C$  has no extra input parameters (such as  $\varepsilon$  or  $\alpha$ ) that are needed in all past skeletonization algorithms for an unstructured cloud  $C$ .

- For a cloud  $C \subset \mathbb{R}^2$  of any  $n$  points, the skeleton  $\text{HoPeS}(C)$  with  $O(n)$  edges can be found in time  $O(n \log n)$ , which is comparable only with [4], [11], [16].
- $\text{HoPeS}(C)$  is the *first universal structure* on a cloud  $C$  that summarizes all cycles of  $C^\alpha$  and has a subgraph  $\text{HoPeS}'(C)$  stable under perturbations of  $C$ .
- $\text{HoPeS}'(C)$  approximates a graph  $G$  in the  $2\varepsilon$ -offset  $G^{2\varepsilon}$  from any  $\varepsilon$ -sample of  $G$ , which has an analogue only in [4] for a metric reconstruction problem.
- Theorem 5 gives guarantees only in simple terms of a graph  $G \subset \mathbb{R}^2$  and its noisy  $\varepsilon$ -sample  $C \subset \mathbb{R}^2$ , while [11, Theorem 3.1] needs a complex  $K$  with a homotopy equivalence  $h : K \rightarrow G$  that  $\varepsilon$ -approximates the metrics of  $K$  and  $G$ .

## 4 Homologically Persistent Skeleton and its optimality

Here we give a rather intuitive introduction into homology theory using only  $\alpha$ -offsets  $C^\alpha$  as typical spaces. All rigorous definitions are in Appendix A.

The *0-dimensional homology*  $H_0$  counts connected components. Formally,  $H_0(C^\alpha)$  is the group (or vector space of linear combinations with coefficients in  $\mathbb{Z}_2 = \{0, 1\}$ ) generated by the components of  $C^\alpha$ . For instance, the offset  $C^{2.5}$  in Fig. 3 has 2 components. Hence  $H_0(C^{2.5}) = \mathbb{Z}_2 \oplus \mathbb{Z}_2$  has rank (*dimension*) 2.

The *1-dimensional homology*  $H_1$  of  $C^\alpha \subset \mathbb{R}^2$  similarly counts *holes* in  $C^\alpha$  (bounded regions in the complement  $\mathbb{R}^2 - C^\alpha$ ). For example, the offset  $C^{\sqrt{34}/2}$  in Fig. 3 has 1 red hole, so  $H_1(C^{\sqrt{34}/2}) = \mathbb{Z}_2$ . This hole splits into 2 holes at  $\alpha = 3$ , hence  $H_1(C^3) = \mathbb{Z}_2 \oplus \mathbb{Z}_2$ . The smaller of the 2 holes disappears when  $\alpha = \frac{25}{8}$  is the circumradius of the triangle on vertices  $(\pm 3, 0)$  and  $(0, -4)$ , so  $H_1(C^{25/8}) = \mathbb{Z}_2$ . The remaining hole dies when  $\alpha = \frac{17}{5}$  is the circumradius of the triangle on vertices  $(\pm 3, 0)$  and  $(0, 5)$ , hence  $H_1(C^{17/5}) = 0$  is trivial.

All  $\alpha$ -offsets form an ascending *filtration* (a nested sequence of spaces)  $C = C^0 \subset \dots \subset C^\alpha \subset \dots \subset C^{+\infty} = \mathbb{R}^2$ . These inclusions induce linear maps in  $H_1$ :

$$C^{2.5} \subset C^{\sqrt{34}/2} \subset C^3 \subset C^{25/8} \subset C^{17/5} \text{ induce } 0 \rightarrow \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \rightarrow 0.$$

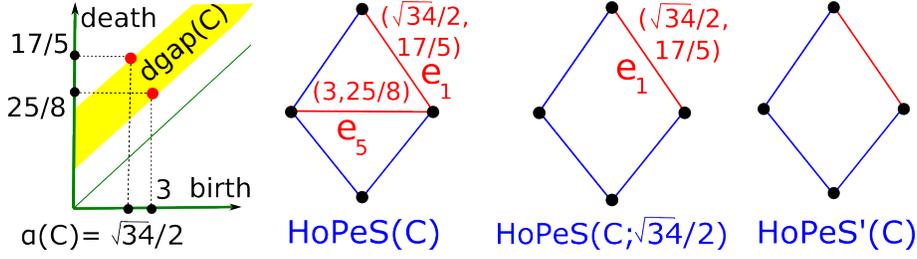
The sequence of the linear maps in  $H_1$  above splits into 2 simpler sequences:

hole 1 lives over the interval  $\frac{\sqrt{34}}{2} \leq \alpha < \frac{17}{5}$ , namely  $0 \rightarrow \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \rightarrow 0$ ,  
hole 2 lives over the short interval  $3 \leq \alpha < \frac{25}{8}$ , namely  $0 \rightarrow 0 \rightarrow \mathbb{Z}_2 \rightarrow 0 \rightarrow 0$ .

At  $\alpha = 3$  when the initial hole splits into 2 smaller holes, we assume that one of the holes ‘inherits’ (continues the life of) the previous hole, while another hole is ‘newborn’ at the splitting moment. The standard convention is to give preference to a longer living hole. So the life spans (the *barcode*) of the filtration  $\{C^\alpha\}$  are  $[\frac{\sqrt{34}}{2}, \frac{17}{5})$  and  $[3, \frac{25}{8})$ . We plot the endpoints of these bars as red dots with coordinates (birth, death) in the *persistence diagram*  $\text{PD}\{C^\alpha\}$ , see Fig. 4.

This diagram is a summary of life spans of holes (1-dimensional homology classes) of  $C^\alpha$  across all scales  $\alpha$ . The key result of persistent homology is Stability Theorem 17 [5] roughly saying that any small perturbation of the cloud  $C$  gives rise to a similar small perturbation of the diagram  $\text{PD}\{C^\alpha\}$  in the plane.

If a hole of  $C^\alpha$  is born, then this hole becomes enclosed by a cycle through points of  $C$ . The last longest edge in this enclosing cycle is added at the *birth* time  $\alpha$  of the hole and is *critical* for the hole in question. Hole 1 born at  $\alpha = \frac{\sqrt{34}}{2}$  has the critical edge  $e_1$ , see Fig. 4. Hole 2 born at  $\alpha = 3$  has the critical edge  $e_5$ .



**Fig. 4.** Diagram  $\text{PD}\{C^\alpha\}$  for the cloud  $C$  in Fig. 3 and skeletons from Definitions 2, 6.

For any filtration  $\{C^\alpha\}$ , each red dot in  $\text{PD}\{C^\alpha\}$  has a corresponding critical edge  $e$  (between points of  $C$ ) with the label  $(\text{birth}(e), \text{death}(e))$ . Our Definition 2 transforms the diagram  $\text{PD}\{C^\alpha\}$  of disconnected points into a universal structure on the data cloud  $C$  summarizing the persistence of holes in  $\{C^\alpha\}$  for all  $\alpha$ .

**Definition 2** For a cloud  $C$ , a Homologically Persistent Skeleton  $\text{HoPeS}(C)$  is the union of  $\text{MST}(C)$  and all critical edges with their labels  $(\text{birth}, \text{death})$ , see Fig. 4. The reduced skeleton  $\text{HoPeS}(C; \alpha)$  is obtained from  $\text{HoPeS}(C)$  by removing all edges longer than  $2\alpha$  and all critical edges  $e$  with  $\text{death}(e) \leq \alpha$ .

If  $\alpha = 0$ , then  $\text{HoPeS}(C; 0) = C$  is the given cloud. By Definition 2 a critical edge  $e$  belongs to the reduced skeleton  $\text{HoPeS}(C; \alpha)$  if and only if  $\text{birth}(e) \leq \alpha < \text{death}(e)$ . So a critical edge  $e$  is added to  $\text{HoPeS}(C; \alpha)$  at  $\alpha = \text{birth}(e)$  and is later removed at the larger scale  $\alpha = \text{death}(e)$ . The cloud  $C$  in Fig. 3 has  $\text{HoPeS}(C; \frac{\sqrt{34}}{2}) = \text{MST}(C) \cup e_1$ , but  $\text{HoPeS}(C; 3)$  is the full skeleton  $\text{HoPeS}(C)$ .

The filtration  $\{\text{HoPeS}(C; \alpha)\}$  may not be monotone with respect to the scale  $\alpha$ . But if  $\text{HoPeS}(C; \alpha)$  has become connected, it will stay connected for all larger  $\alpha$ . Indeed, removing a critical edge destroys only a cycle, not connectivity.

Similarly to  $\text{MST}(C)$ , a Homologically Persistent Skeleton  $\text{HoPeS}(C)$  is unique in a general position when the distances between all points of  $C$  are different.

**Lemma 3** For any cloud  $C \subset \mathbb{R}^2$  of  $n$  points, a Homologically Persistent Skeleton  $\text{HoPeS}(C)$  has the size  $O(n)$  and is computable in time  $O(n \log n)$ .

Lemma 4 below help visualize the 1-dimensional persistence diagram  $\text{PD}\{C^\alpha\}$  directly on the cloud  $C$ . Lemma 4 justifies that  $\text{HoPeS}(C)$  is suitable for Computer Vision applications where a scale-and-rotation invariance is important.

**Lemma 4** *For a cloud  $C \subset \mathbb{R}^2$ , the 1-dimensional persistence diagram  $\text{PD}\{C^\alpha\}$  of the filtration of  $\alpha$ -offsets  $C^\alpha$  can be reconstructed from a Homologically Persistent Skeleton  $\text{HoPeS}(C)$ . The topological structure of  $\text{HoPeS}(C)$  is invariant under any affine transformation whose  $2 \times 2$  matrix has equal eigenvalues.*

Our first main Theorem 5 says that  $\text{HoPeS}(C)$  is an optimal graph that extends  $\text{MST}(C)$  and captures the persistence of all holes in the filtration  $\{C^\alpha\}$ .

**Theorem 5** *For any cloud  $C \subset \mathbb{R}^2$  and any  $\alpha > 0$ , the graph  $\text{HoPeS}(C; \alpha)$  has the minimum total length of edges over all graphs  $G \subset C^\alpha$  that span the  $\alpha$ -offset  $C^\alpha$  and induce an isomorphism in 1-dimensional homology  $H_1(G) \rightarrow H_1(C^\alpha)$ .*

A graph  $G$  spans  $C^\alpha$  if  $G \subset C^\alpha$  induces an isomorphism  $H_0(G) \cong H_0(C^\alpha)$ . An isomorphism  $H_1(G) \cong H_1(C^\alpha)$  means that the graph  $G$  has the homotopy type of the  $\alpha$ -offset  $C^\alpha \subset \mathbb{R}^2$ . Hence our Homologically Persistent Skeleton  $G = \text{HoPeS}(C)$  solves the multi-scale skeletonization problem stated in sections 1–2.

## 5 The reconstruction theorem and stability of $\text{HoPeS}(C)$

A Homologically Persistent Skeleton  $\text{HoPeS}(C)$  contains all 1-dimensional cycles in the offsets  $C^\alpha$  across the full range of  $\alpha$ . It is natural to select cycles with highest persistence to get a smaller subgraph  $\text{HoPeS}'(C) \subset \text{HoPeS}(C)$ . So we select not a scale as in scale-space theory, but a widest diagonal gap in the persistence diagram  $\text{PD}\{C^\alpha\}$ . This widest gap makes sense for finite sets  $C$  and for any compact set  $S \subset \mathbb{R}^2$  that is a finite union of closed topological disks.

**Definition 6** *For a compact set  $S \subset \mathbb{R}^2$  and the ascending filtration of offsets  $S^\alpha$ , a diagonal gap in the persistence diagram  $\text{PD}\{S^\alpha\}$  is a largest (by inclusion) strip  $\{0 \leq a < y - x < b\}$  that has no points from the diagram, see Fig. 3.*

*The widest diagonal gap  $\text{dgap}(S)$  has the largest width  $|\text{dgap}(S)| = b - a$ . Let the subdiagram  $\text{PD}'\{S^\alpha\} \subset \text{PD}\{S^\alpha\}$  have only the points above  $\text{dgap}(S)$ . The critical scale  $\alpha(S)$  is the maximum birth over all  $(\text{birth}, \text{death}) \in \text{PD}'\{S^\alpha\}$ .*

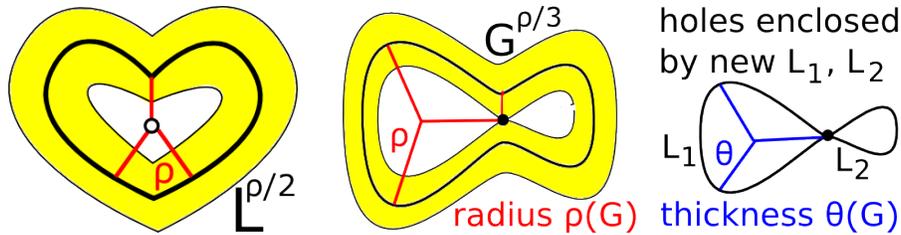
*For a cloud  $C = S$ , the derived skeleton  $\text{HoPeS}'(C)$  is obtained from  $\text{HoPeS}(C)$  by removing (1) all edges longer than  $2\alpha(C)$ , and (2) all critical edges either with death  $\leq \alpha(C)$  or with  $(\text{birth}, \text{death})$  below the widest diagonal gap  $\text{dgap}(C)$ .*

In Definition 6 if there are different gaps with the same width, we say that the gap with largest values along the vertical death axis has the largest width. The cloud  $C$  in Fig. 3 has the widest gap  $\text{dgap}(C)$  between the points  $(\frac{\sqrt{34}}{2}, \frac{17}{5})$  and  $(3, \frac{25}{8})$  in  $\text{PD}\{C(\alpha)\}$ , so the critical scale is  $\alpha(C) = \frac{\sqrt{34}}{2}$ , see Fig. 4.

Condition (1) above guarantees that  $\text{HoPeS}'(C) \subset \text{HoPeS}(C; \alpha(C))$ , because all long critical edges  $e$  with  $\text{birth}(e) > \alpha(C)$  are removed, see Lemmas 24 and 27 in Appendix B. Condition (2) filters out cycles with early deaths and low persistence, but  $\text{HoPeS}'(C) \neq \text{HoPeS}(C; \alpha(C))$ . Instead of selecting a fixed scale as in scale-space theory, we select cycles by their persistence across all scales  $\alpha$ .

We define concepts needed for Theorem 7. A non-self-intersecting cycle  $L$  in a graph  $G \subset \mathbb{R}^2$  is *basic* if  $L$  encloses a bounded region of  $\mathbb{R}^2 - G$ . When  $\alpha$  is increasing, the hole enclosed by the  $\alpha$ -offset  $L^\alpha$  is born at  $\alpha = 0$  and dies at the scale  $\alpha = \rho(L)$  that is called the *radius* of the cycle  $L$ . So the initial hole enclosed by  $L$  has the life span  $[0, \rho(L))$ . The heart-shaped hole in the first picture of Fig. 5 completely dies at  $\alpha = \rho(L)$ , which holds for any convex hole.

In general, when  $\alpha$  is increasing new holes can be born in  $G^\alpha$ , let them be enclosed by  $L_1, \dots, L_k$  at their birth times. The *thickness*  $\theta(G) = \max_{j=1, \dots, k} \rho(L_j)$  is the maximum persistence of these smaller holes born during the evolution of offsets  $G^\alpha$ . If no such holes appear, then  $\theta = 0$ , otherwise  $\theta > 0$ , see Fig. 5.



**Fig. 5.** The ‘heart’ graph has thickness  $\theta = 0$ . The ‘figure-eight’ graph has  $\theta > 0$ .

Theorem 7 says that  $\text{HoPeS}'(C)$  is a close approximation to a graph  $G$  from any its  $\varepsilon$ -sample  $C$ . The homotopy type of a connected graph  $G$  is determined by its  $H_1(G)$ . Namely,  $G$  continuously deforms to a wedge of  $\dim H_1(G)$  loops.

**Theorem 7** *Let  $C$  be any  $\varepsilon$ -sample of a connected graph  $G \subset \mathbb{R}^2$  with a thickness  $\theta(G) \geq 0$  and  $m \geq 1$  basic cycles having ordered radii  $\rho_1 \leq \dots \leq \rho_m$ . If  $\rho_1 > 7\varepsilon + \theta(G) + \max_{i=1, \dots, m-1} \{\rho_{i+1} - \rho_i\}$ , then the critical scale  $\alpha(C) \leq \varepsilon$ , and the derived skeleton  $\text{HoPeS}'(C)$  is  $2\varepsilon$ -close to  $G$  and has the homotopy type of  $G$ .*

The inequality above means that the cycles of the graph  $G$  have ‘comparable’ sizes, i.e. the smallest radius  $\rho_1$  is larger by a good margin than any gap  $\rho_{i+1} - \rho_i$  between the ordered radii. Hence the diagonal gap  $\{\theta(G) < \text{death} - \text{birth} < \rho_1\}$  in the diagram  $\text{PD}\{G^\alpha\}$  of the graph  $G$  will remain wide enough to be automatically recognized in the perturbed diagram  $\text{PD}\{C^\alpha\}$  for any  $\varepsilon$ -sample  $C$  of  $G$ .

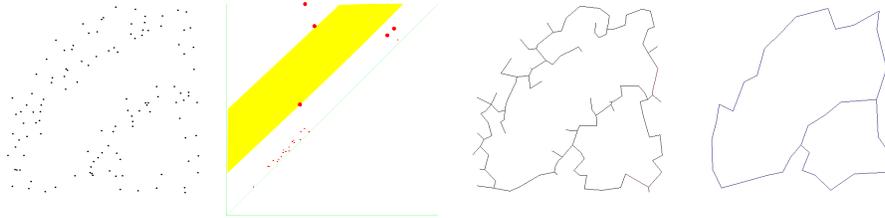
Theorem 7 is stronger than any estimate of homology from noisy samples. In addition we build on a sample  $C$  an actual skeleton  $\text{HoPeS}'(C)$  that is  $2\varepsilon$ -close to an unknown graph  $G$ . Theorem 7 extends simpler [13, Theorem 32], which works only for a much smaller class of graphs  $G \subset \mathbb{R}^2$  with thickness  $\theta = 0$ .

**Corollary 8** *In the conditions of Theorem 7 if another cloud  $\tilde{C}$  is  $\delta$ -close to  $C$ , then the perturbed derived skeleton  $\text{HoPeS}'(\tilde{C})$  is  $(2\delta + 4\varepsilon)$ -close to  $\text{HoPeS}'(C)$ .*

We can't expect that  $\text{HoPeS}'(C)$  is locally stable for any cloud  $C$ , because a minimum spanning tree  $\text{MST}(C)$  is sensitive to perturbations of  $C$ . However, Corollary 8 guarantees the overall stability of the derived skeleton (within a small offset) in the most practical case for noisy sample of graphs. Proofs were checked during talks at Oxford (UK), IST (Austria), TU Wien, TU Graz.

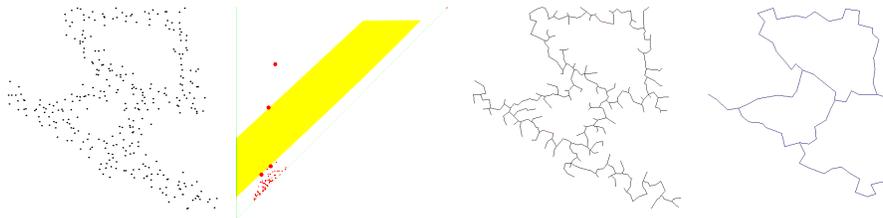
## 6 Algorithm, experiments and practical applications

Lemma 12 in Appendix A justifies that complicated  $\alpha$ -offsets  $C^\alpha$  can be replaced by simpler  $\alpha$ -complexes  $C(\alpha)$ , which filter a Delaunay triangulation  $\text{Del}(C)$ . Starting from a cloud  $C \subset \mathbb{R}^2$  of  $n$  points, we build  $\text{Del}(C)$  in time  $O(n \log n)$  with  $O(n)$  space. Regions in the complement  $\mathbb{R}^2 - C(\alpha)$  are dual to their boundaries. This duality [3] reduces 1-dimensional persistence of cycles in the filtration  $\{C(\alpha)\}$  to 0-dimensional persistence of connected components in  $\mathbb{R}^2 - C(\alpha)$ .



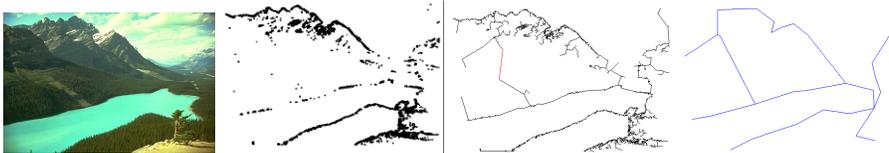
**Fig. 6.** A sample  $C$  of O45, diagram  $\text{PD}\{C(\alpha)\}$ ,  $\text{HoPeS}'(C)$  and its simplification.

The 0-dimensional persistence is computed in time  $O(nA^{-1}(n))$  using a union-find structure [10], where  $A^{-1}(n)$  is the slow growing inverse Ackermann function. We extend this algorithm by recording a critical edge along which two regions of  $\mathbb{R}^2 - C(\alpha)$  merge when  $\alpha$  is decreasing. See details in Appendix C.



**Fig. 7.** A sample  $C$  of D33, diagram  $\text{PD}\{C(\alpha)\}$ ,  $\text{HoPeS}'(C)$  and its simplification.

Fig. 6 shows 121 random points sampled from a real image of hieroglyph O45. The second picture of Fig. 6 is the diagram  $PD\{C(\alpha)\}$  with a widest diagonal gap clearly separating the noise near the diagonal from 2 red points corresponding to 2 cycles in the derived graph  $HoPeS'(C)$ . Theorem 7 gives the lower bound  $\alpha(C)$  for the unknown noise level  $\varepsilon$ . We use this intrinsic critical scale  $\alpha(C)$  for pruning short branches and collapsing short edges to get a simplified version of  $HoPeS'(C)$  in the last picture of Fig. 6, see details in Appendix C. Fig. 7 has similar results for 321 points sampled from another hieroglyph D33.



**Fig. 8.** Image BSD176035, cloud  $C$  of 3603 points,  $HoPeS'(C)$  and its simplification.

In Fig. 8 we selected feature points from a challenging image by simply comparing the color of each pixel with the average in  $5 \times 5$  neighborhood. The threshold for the normal deviation in the 3-dimensional RGB space was 65 as in Fig. 1. Fig. 9 shows similar results for the normal deviation 100 of the color.



**Fig. 9.** Image BSD42049, cloud  $C$  of 1763 points,  $HoPeS'(C)$  and its simplification.

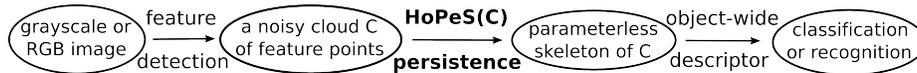
Table 2 has the running time in milliseconds for the database BSD500, where all images have  $481 \times 321$  pixels and we used 2 thresholds in each images, see details in Appendix D. We ran our not yet optimized code on a small laptop with 1.33GHz RAM 2GB to show that the algorithm is fast for embedded systems.

Images from BSD500	42049	42049	176035	176035	175083	175083	134049	134049
Time for a cloud $C$ , ms	1022	1336	1017	1038	1066	1015	1051	1165
Points in the cloud $C$	2664	3604	3603	4249	3928	4950	4396	6767
Time for $HoPeS'(C)$ , ms	969	1789	1898	2629	2143	3602	3780	6259

**Table 2.** Time for extracting  $C$  from images in BSD500 and computing  $HoPeS'(C)$ .

We have demonstrated the following practical applications of  $HoPeS(C)$ .

- Robust recognition of low quality scans in Fig. 6, 7, 15, 16, 17, 18. Such visual markers [7] can replace shop barcodes that are not readable by humans.
- A fast topological summary of images, see Fig. 1, 8, 9, 19, 20, 21, 22.



**Fig. 10.** Pipeline for building an object-wide descriptor from noisy local features

We are open to collaboration on extending  $\text{HoPeS}(C)$  to higher dimensions and using  $\text{HoPeS}(C)$  for clouds  $C$  of interest points such as SIFT or SURF. We thank all reviewers in advance for their comments and helpful suggestions.

## References

1. M. Aanjaneya, F. Chazal, D. Chen, M. Glisse, L. Guibas, D. Morozov. Metric graph reconstruction from noisy data. *IJCGA*, v. 22 (2012), 305–325.
2. D. Attali, J.-D. Boissonnat, and H. Edelsbrunner. Stability and computation of medial axes - a state-of-the-art report. In *Math. Foundations of Visualization, Computer Graphics, and Massive Data Exploration*, 109-125. Springer, 2009.
3. D. Attali, M. Glisse, S. Hornus, F. Lazarus, and D. Morozov. Persistence-sensitive simplification of functions on surfaces in linear time. *Proceedings of TopoInVis'09*.
4. F. Chazal, J. Sun. Gromov-Hausdorff approximation of filament structure using Reeb-type graph. *Proceedings of SoCG 2014*, to appear in Discrete Comp. Geometry.
5. D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, v. 37 (2007), p. 103–130.
6. N. Cornea, D. Silver, and P. Min. Curve-Skeleton Properties, Applications, and Algorithms *IEEE Trans. Visualization Comp. Graphics*, v. 13 (2007), 530–548.
7. E. Costanza and J. Huang. Designable visual markers. *Proceedings of SIGCHI 2009: Special Interest Group on Computer-Human Interaction*, p. 1879-1888.
8. T. Dey, F. Fan, and Y. Wang. Graph induced complex on data points. *Proceedings of SoCG 2013: Symposium on Computational Geometry*, p. 107–116.
9. H. Edelsbrunner. The union of balls and its dual shape. *Discrete Computational Geometry*, v. 13 (1995), p. 415–440.
10. H. Edelsbrunner, J. Harer. *Computational topology. An introduction*. AMS, 2010.
11. X. Ge, I. Safa, M. Belkin, and Y. Wang. Data skeletonization via Reeb graphs. *Proceedings of NIPS 2011*, p. 837–845.
12. V. Kurlin. A fast and robust algorithm to count topologically persistent holes in noisy clouds. *Proceedings of CVPR 2014*, available at <http://kurlin.org>.
13. V. Kurlin. Auto-completion of contours in sketches, maps and sparse 2D images based on persistence. *Proceedings of CTIC 2014*, available at <http://kurlin.org>.
14. D. Letscher, J. Fritts. Image segmentation using topological persistence. *Proceedings of CAIP 2007: Computer Analysis of Images and Patterns*, 587–595
15. T. Lindeberg. *Scale-Space Theory in Computer Vision*, Kluwer Publishers, 1994.
16. R. Singh, V. Cherkassky, N. Papanikolopoulos. Self-organizing maps for the skeletonization of sparse shapes. *Tran. Neural Networks*, v. 11 (2000), 241–248.

## Appendix A: $\alpha$ -complexes $C(\alpha)$ and persistent homology

To avoid any confusion, we continue numbering claims, figures as in the paper.

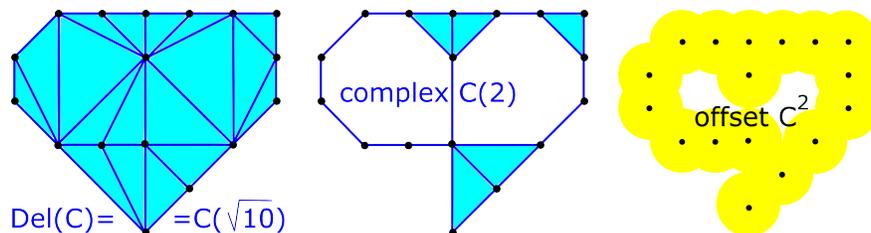
**Definition 9** A plane graph is a subset  $G \subset \mathbb{R}^2$  consisting of finitely many vertices and non-intersecting edges (continuous arcs) joining vertices.

We study topological structures such as graphs and 2-dimensional complexes. Topological invariants are preserved under continuous deformations including any projective transformations. So topological invariants are insensitive to geometry and help recognize objects that are viewed from different angles.

If every bounded region in the complement  $\mathbb{R}^2 - G$  of a plane graph is a triangle, then the graph  $G$  defines a *triangulation* on its vertices. A Delaunay triangulation  $\text{Del}(C)$  with vertices in a cloud  $C \subset \mathbb{R}^2$  consists of ‘nice’ triangles.

**Definition 10** For a cloud  $C = \{p_1, \dots, p_n\} \subset \mathbb{R}^2$  of  $n$  distinct points, a Delaunay triangulation  $\text{Del}(C)$  has all triangles with vertices  $p_i, p_j, p_k \in C$  whose circumcircle doesn’t enclose any other points of the cloud  $C$ , see Fig. 11.

A Delaunay triangulation  $\text{Del}(C)$  of a cloud  $C \subset \mathbb{R}^2$  is not unique if  $C$  contains 4 points on the same circle. The boundary edges of  $\text{Del}(C)$  form the convex hull( $C$ ) of  $C$ . It is well-known that, for any finite cloud  $C \subset \mathbb{R}^2$  of  $n$  points, a Delaunay triangulation  $\text{Del}(C)$  has  $O(n)$  edges and triangles, and can be found in time  $O(n \log n)$ .



**Fig. 11.** Delaunay triangulation  $\text{Del}(C)$ ,  $C(2)$  and  $C^2$  for the cloud  $C$  in Fig. 2.

To study a cloud  $C$  at different scales, we shall define subcomplexes containing the elements of  $\text{Del}(C)$  whose sizes are smaller than a fixed scale  $\alpha$ . We introduce Voronoi cells, which are ‘natural’ neighborhoods of points  $p \in C$ .

For a point  $p_i \in C$ , the *Voronoi cell* consists of all points  $q \in \mathbb{R}^2$  that are closer to  $p_i$  than to all other points of  $C$ , so  $V(p_i) = \{q \in \mathbb{R}^2 : d(p_i, q) \leq d(p_j, q) \forall j \neq i\}$ . Then a *Delaunay triangulation*  $\text{Del}(C)$  contains all triangles with vertices  $p, q, r \in C$  such that  $V(p) \cap V(q) \cap V(r) \neq \emptyset$ .

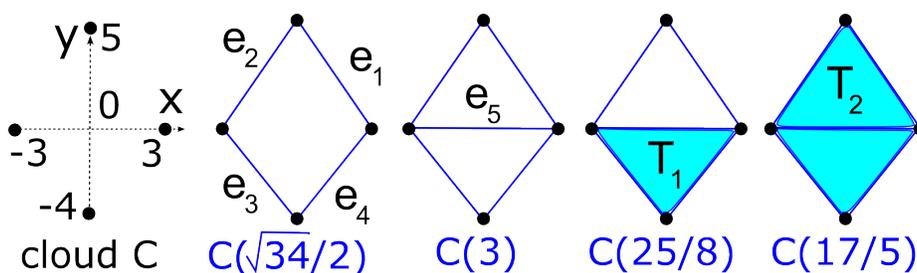
The  $\alpha$ -complex  $C(\alpha)$  is defined below as a subtriangulation of  $\text{Del}(C)$  at a scale  $\alpha$ . Denote by  $B(p; \alpha) \subset \mathbb{R}^2$  the closed disk with a center  $p$  and radius  $\alpha$ .

**Definition 11** For a finite cloud  $C \subset \mathbb{R}^2$ , the  $\alpha$ -complex  $C(\alpha) \subset \mathbb{R}^2$  contains all edges between points  $p, q \in C$  such that  $V(p) \cap B(p; \alpha)$  meets  $V(q) \cap B(q; \alpha)$ , see [10, section III.4]. Similarly, the  $\alpha$ -complex  $C(\alpha)$  contains all triangles with vertices  $p, q, r$  such that  $V(p) \cap B(p; \alpha) \cap V(q) \cap B(q; \alpha) \cap V(r) \cap B(r; \alpha) \neq \emptyset$ .

If  $\alpha > 0$  is small,  $C(\alpha)$  consists of all isolated points of  $C$ . If  $\alpha$  is large enough, then  $C(\alpha) = \text{Del}(C)$ . All  $\alpha$ -complexes form a sequence of nested complexes, called a *filtration*  $C = C(0) \subset \dots \subset C(\alpha) \subset \dots \subset C(+\infty) = \text{Del}(C)$ . So  $\text{Del}(C)$  is built on points of  $C$  by adding edges and triangles at these critical values:

- edge between points  $p_i, p_j$  is added at  $\alpha = \frac{1}{2}d(p_i, p_j)$ ;
- an *acute* triangle (that has all angles  $< \frac{\pi}{2}$ ) is added at the critical value  $\alpha$  equal to the circumradius of the triangle;
- a non-acute triangle is added to  $C(\alpha)$  at the scale  $\alpha$  that is equal to the half-length of the largest side in the triangle.

The ascending filtration of  $\alpha$ -complexes  $C(\alpha)$  is illustrated for the cloud  $C$  in Fig. 12. The lengths are  $|e_3| = |e_4| = 5 < |e_1| = |e_2| = \sqrt{34} < |e_5| = 6$ . The triangles  $T_1, T_2$  enter  $C(\alpha)$  when  $\alpha$  equals their circumradii  $\frac{25}{8}, \frac{17}{5}$ .



**Fig. 12.** The cloud  $C$  of 4 points,  $C(\alpha)$  for  $\alpha = \frac{\sqrt{34}}{2}, 3, \frac{25}{8}, \frac{17}{5}$ .

The complex  $C(\alpha)$  is a subset of the offset  $C^\alpha$ , because any edge of  $C(\alpha)$  has a half-length  $\frac{1}{2}d(p_i, p_j) \leq \alpha$ . Nerve Lemma 12 below says that the  $\alpha$ -offset  $C^\alpha$  can be gradually made thinner until it becomes the complex  $C(\alpha)$ . This deformation is easy to see for  $C^2$  in Fig. 2 and  $C(2)$  in Fig. 11.

**Lemma 12** [9, Theorem 3.2(i)] For any cloud  $C \subset \mathbb{R}^2$  and any scale  $\alpha > 0$ , the  $\alpha$ -offset  $C^\alpha$  is homotopy equivalent to the  $\alpha$ -complex  $C(\alpha)$ . Namely, the identity map  $\text{id} : C^\alpha \rightarrow C^\alpha$  is included into a continuous family of deformations  $f_t : C^\alpha \rightarrow C^\alpha$ ,  $t \in [0, 1]$ , where  $f_0 = \text{id}$  and  $f_1$  projects  $C^\alpha$  to  $C(\alpha)$ .

The  $\alpha$ -complex  $C(\alpha)$  is an example of a *simplicial* complex  $S \subset \mathbb{R}^2$  consisting of finitely many edges and triangles that are naturally joined. For instance, the intersection of any two triangles should be their common edge, or their common vertex or empty. A *cycle* in a complex  $S$  is any cycle formed by edges of  $S$ .

The homology group  $H_1(S)$  is a vector space whose dimension is the number of ‘independent’ cycles in  $S$ . The complex  $C(3)$  in Fig. 12 has 2 independent cycles  $e_1 \cup e_2 \cup e_5$  and  $e_3 \cup e_4 \cup e_5$ . The third cycle  $e_1 \cup e_2 \cup e_3 \cup e_4$  can be considered as a ‘sum’ of the above cycles modulo 2, which is formalized below.

**Definition 13** *Cycles of a complex  $S$  can be algebraically written as linear combinations of edges with coefficients 0 or 1 in the group  $\mathbb{Z}_2 = \{0, 1\}$ . The vector space  $C_1$  consists of all these linear combinations. The boundaries of all triangles in  $S$  (as cycles of 3 edges) generate the subspace  $B_1 \subset C_1$ . The quotient  $C_1/B_1$  is the 1-dimensional homology group  $H_1(S)$  with the coefficients  $\mathbb{Z}_2 = \{0, 1\}$ .*

The complex  $C(3)$  in Fig. 12 has  $C_1 = \mathbb{Z}_2 \oplus \mathbb{Z}_2$  generated by  $e_1 + e_2 + e_5$  and  $e_3 + e_4 + e_5$  whose sum modulo 2 is the third cycle  $e_1 + e_2 + e_3 + e_4$ , so  $H_1(C(3)) = \mathbb{Z}_2 \oplus \mathbb{Z}_2$ . In the complex  $C(\frac{25}{8})$  the triangle  $T_1$  has the boundary  $\partial T_1 = e_3 + e_4 + e_5$  and kills this cycle in the quotient  $C_1/B_1$ , hence  $H_1(C(\frac{25}{8})) = \mathbb{Z}_2$ . Similarly,  $T_2 \subset C(\frac{17}{5})$  kills  $\partial T_2 = e_1 + e_2 + e_5$ , so  $H_1(C(\frac{17}{5})) = 0$ .

A class  $\gamma \in H_1(S)$  may be represented by several *homologically equivalent* cycles  $L \subset S$ . The classes  $e_1 + e_2 + e_5$  and  $e_1 + e_2 + e_3 + e_4$  are equal in  $H_1(C(\frac{25}{8}))$ , see Fig. 12, because their difference modulo 2 is  $\partial T_1 = e_3 + e_4 + e_5$ , which can be continuously shrunk to a point through  $T_1$ .

The inclusions  $C(0) \subset \dots \subset C(\alpha) \subset \dots \subset C(+\infty)$  induce linear maps of corresponding homology groups  $H_1(C(0)) \rightarrow \dots \rightarrow H_1(C(\alpha)) \rightarrow \dots \rightarrow H_1(C(+\infty))$ . The inclusions  $C \subset C(\frac{\sqrt{34}}{2}) \subset C(3) \subset C(\frac{25}{8}) \subset C(\frac{17}{5})$  of complexes in Fig. 12 induce the linear maps  $0 \rightarrow \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \rightarrow 0$ .

The homology class  $e_1 + e_2 + e_3 + e_4$  is born at  $\alpha = \frac{\sqrt{34}}{2}$  and persists through  $\alpha = 3$ , when the class  $e_1 + e_2 + e_5$  is born. At  $\alpha = \frac{25}{8}$  their sum (modulo 2)  $e_3 + e_4 + e_5$  becomes trivial. We keep the older class and let the younger class  $e_1 + e_2 + e_5$  die by the ‘elder’ rule of persistence [10, p. 150]. This merger corresponds to the projection  $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \rightarrow \mathbb{Z}_2$  above. The class  $e_1 + e_2 + e_3 + e_4$  dies at  $\alpha = \frac{17}{5}$  and has the life span (or persistence) death – birth =  $\frac{17}{5} - \frac{\sqrt{34}}{2}$ . We now formally define births and deaths of classes.

**Definition 14** *In a filtration  $\{C(\alpha)\}$  of complexes a class  $\gamma \in H_1(C(\alpha_i))$  is born at  $\alpha_i = \text{birth}(\gamma)$  if  $\gamma$  is not in the image of the map  $H_1(C(\alpha)) \rightarrow H_1(C(\alpha_i))$  for any  $\alpha < \alpha_i$ . The class  $\gamma$  dies at  $\alpha_j = \text{death}(\gamma) > \alpha_i$  when the image of  $\gamma$  under  $H_1(C(\alpha_i)) \rightarrow H_1(C(\alpha_j))$  merges into the image of  $H_1(C(\alpha)) \rightarrow H_1(C(\alpha_j))$  for some  $\alpha < \alpha_i$ .*

If  $H_1(C(\alpha - \varepsilon)) \rightarrow H_1(C(\alpha + \varepsilon))$  is not an isomorphism for any small enough  $\varepsilon > 0$ , then  $\alpha$  is *critical* (a birth or a death). Births and deaths are visualized below as points (birth, death) in the persistence diagram on the plane.

**Definition 15** *For a cloud  $C \subset \mathbb{R}^2$ , let  $\alpha_1, \dots, \alpha_k$  be all values of  $\alpha$  when the homology group  $H_1(C(\alpha))$  changes. Let  $\mu_{ij}$  be the number of independent classes in  $H_1(C(\alpha))$  that are born at  $\alpha_i$  and die at  $\alpha_j$ . The persistence diagram  $\text{PD}\{C(\alpha)\} \subset \mathbb{R}^2$  is the multi-set consisting of all points  $(\alpha_i, \alpha_j)$  with the multiplicity  $\mu_{ij}$  and all diagonal points  $(x, x)$  with the infinite multiplicity.*

Points close to the diagonal are considered as noise, because corresponding homology classes have a low persistence or a short life span from birth to death.  $\text{PD}\{C(\alpha)\}$  in Fig. 4 has only 2 off-diagonal points  $(\frac{\sqrt{34}}{2}, \frac{17}{5})$  and  $(3, \frac{25}{8})$ .

The homology and persistence diagram can be defined for spaces more complicated than a simplicial complex, say for  $\alpha$ -offsets  $G^\alpha$  of a graph  $G \subset \mathbb{R}^2$ . Lemma 12 implies that  $H_1(C^\alpha) = H_1(C(\alpha))$ , so  $\text{PD}\{C^\alpha\} = \text{PD}\{C(\alpha)\}$ .

The dimension of  $H_1(C(\alpha))$  equals the number of points  $(b, d) \in \text{PD}\{C(\alpha)\}$  counted with multiplicities in the ‘infinite rectangle’  $\{\text{birth} \leq \alpha < \text{death}\} = [0, \alpha] \times (\alpha, +\infty)$ . Indeed, a homology class  $\gamma$  that is alive at the scale  $\alpha$  was born earlier at  $\text{birth} \leq \alpha$  and should die later at  $\text{death} > \alpha$ . The complex  $C(\frac{\sqrt{34}}{2})$  in Fig. 12 is a topological circle and has  $H_1 = \mathbb{Z}_2$ , because the diagram  $\text{PD}\{C(\alpha)\}$  in Fig. 4 has one point  $(\frac{\sqrt{34}}{2}, \frac{17}{5})$  satisfying  $\text{birth} \leq \frac{\sqrt{34}}{2} \leq \text{death}$ .

The key advantage of the persistence diagram over classical homology is *stability under noise*. Replacing one point in a cloud by several close points can introduce many small cycles in  $C(\alpha)$  at a fixed scale  $\alpha$ , but these cycles have a low persistence. To state Stability Theorem 17, we define the bottleneck distance between persistence diagrams. Informally, the bottleneck distance between two sets is the minimum perturbation that allows us to match the sets.

**Definition 16** For points  $p = (x_1, y_1)$ ,  $q = (x_2, y_2)$  in  $\mathbb{R}^2$ , we set  $\|p - q\|_\infty = \max\{|x_1 - x_2|, |y_1 - y_2|\}$ . The bottleneck distance between  $\text{PD}$  and  $\widetilde{\text{PD}}$  is  $d_B = \inf_\psi \sup_q \|q - \psi(q)\|_\infty$  over all bijections  $\psi : \text{PD} \rightarrow \widetilde{\text{PD}}$  of infinite multi-sets.

If we consider life intervals  $[\text{birth}, \text{death}] \subset \mathbb{R}$  instead of points  $(\text{birth}, \text{death}) \in \mathbb{R}^2$ , the distance  $\varepsilon$  means that we perturb endpoints of intervals by at most  $\varepsilon$ , so a class may be born a bit earlier/later or may die a bit earlier/later. Celebrated Stability Theorem 17 briefly says that any perturbation of a given cloud by  $\varepsilon$  yields a perturbation of the persistence diagram by at most  $\varepsilon$ .

**Theorem 17** [5]  $d_B(\text{PD}\{G^\alpha\}, \text{PD}\{C(\alpha)\}) \leq \varepsilon$  for any  $\varepsilon$ -sample  $C$  of  $G$ .

Finally, we formally define critical edges needed for Definition 2.

**Definition 18** For the filtration  $\{C(\alpha)\}$  of  $\alpha$ -complexes of a cloud  $C \subset \mathbb{R}^2$ , each point  $(\alpha_i, \alpha_j) \in \text{PD}\{C(\alpha)\}$  corresponds to a homology class  $\gamma \in H_1(C(\alpha))$  for  $\alpha_i \leq \alpha < \alpha_j$ . This class  $\gamma$  was born when a new (last) edge  $e(\gamma)$  entered  $C(\alpha_i)$ . The edge  $e(\gamma)$  between points of  $C$  is called *critical* and has the label  $(\alpha_i, \alpha_j)$ .

If a point  $(\alpha_i, \alpha_j) \in \text{PD}\{C(\alpha)\}$  has a multiplicity  $\mu_{ij} > 1$ , then there are  $\mu_{ij}$  independent classes in  $H_1(C(\alpha))$  living over  $\alpha_i \leq \alpha < \alpha_j$  and also  $\mu_{ij}$  corresponding critical edges with the label  $(\alpha_i, \alpha_j)$ . For the cloud  $C$  in Fig. 12, let  $\text{MST}(C) = e_2 \cup e_3 \cup e_4$ . Then  $e_1$  is the critical edge of the class  $\gamma$  persisting over  $\frac{\sqrt{34}}{2} \leq \alpha < \frac{17}{5}$ , while  $e_5$  is the critical edge of the class persisting over  $3 \leq \alpha < \frac{25}{8}$ . The length of any critical edge is  $|e(\gamma)| = 2\text{birth}(\gamma)$  by Lemma 19.

For any non-acute Delaunay triangle  $T$ , its longest edge  $e$  is not critical, because  $e$  and  $T$  simultaneously enter  $C(\alpha)$  at the scale  $\alpha = \frac{|e|}{2}$ . So  $e$  is homologically equivalent to  $\partial T - e$  and doesn't give birth to a new homology class.

## Appendix B: detailed proofs of all results in the paper

**Proof of Lemma 1.** Let  $e_1, \dots, e_m \subset \text{MST}(C)$  be all edges that are longer than  $\alpha$ , so  $\text{MST}(C) = \text{MST}(C; \alpha) \cup e_1 \cup \dots \cup e_m$ . Assume that there is a graph  $G$  that spans  $C(\alpha)$  and is shorter than  $\text{MST}(C; \alpha)$ . Then  $G \cup e_1 \cup \dots \cup e_m$  spans the cloud  $C$  and is shorter than  $\text{MST}(C)$ , which is a contradiction.  $\square$

Lemma 1 means that  $\text{MST}(C)$  at every  $\alpha > 0$  provides an optimal graph that identifies all clusters of  $C$ . Main Theorem 5 extends Lemma 1 for a skeleton that summarizes all 1-dimensional persistence of  $C$  instead of only clusters.

**Proof of Lemma 3.** The algorithm in Appendix C constructs a Delaunay triangulation  $\text{Del}(C)$  in time  $O(n \log n)$  with  $O(n)$  space. Then we go over all  $O(n)$  Delaunay edges sorted by their length in the decreasing order to build  $\alpha$ -complexes and mark critical edges added to  $\text{HoPeS}(C)$ . Sorting edges and maintaining a union-find structure  $\text{Forest}(\alpha)$  needs only  $O(n \log n)$  time.  $\square$

**Proof of Lemma 4.** By Definition 2 the labels (birth, death) on all critical edges of  $\text{HoPeS}(C)$  are in a 1-1 correspondence with all points in  $\text{PD}\{C(\alpha)\}$ .

If an affine transformation is given by a shift vector and a  $2 \times 2$  matrix  $A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  with both eigenvalues equal to  $\lambda$ , then all disks,  $\alpha$ -offsets  $C^\alpha$  and  $\alpha$ -complexes  $C(\alpha)$  are scaled by the factor  $\lambda$ . So a Homologically Persistent Skeleton  $\text{HoPeS}(C)$  has the same topological (even combinatorial) structure, but all labels (births, deaths) on critical edges are multiplied by  $\lambda$ .  $\square$

**Lemma 19** *Let a homology class  $\gamma \in H_1(C(\alpha))$  be born due to a critical edge  $e(\gamma)$  added to  $C(\alpha)$ . Then the length of the edge  $e(\gamma)$  equals  $2\text{birth}(\gamma)$ .*

*Proof.* The critical edge  $e(\gamma)$  is the last edge joining a cycle  $L \subset C(\alpha)$  giving birth to the homology class  $\gamma$  at the scale  $\alpha = \text{birth}(\gamma)$ . By Definition 11 the edge  $e(\gamma)$  enters  $C(\alpha)$  when the length  $|e(\gamma)| = 2\alpha$ . So  $|e(\gamma)| = 2\text{birth}(\gamma)$ .  $\square$

An open edge  $e$  is *splitting* a connected graph  $G$  if  $G - e$  is disconnected. Otherwise the edge  $e$  is called *non-splitting* and should be in a cycle of  $G$ .

**Lemma 20** *For a cloud  $C$  and any  $\alpha$ , the reduced skeleton  $\text{HoPeS}(C; \alpha)$  is contained in  $C(\alpha)$  at the same scale  $\alpha$ .*

*Proof.* By Definition 2 all edges of the reduced graph  $\text{HoPeS}(C; \alpha)$  have lengths at most  $\alpha$ . By Definition 11 the  $\alpha$ -complex  $C(\alpha)$  contains all edges from  $\text{Del}(C)$  with a length at most  $\alpha$ . Hence  $\text{HoPeS}(C; \alpha) \subset C(\alpha)$  as required.  $\square$

Now the inclusion  $f : \text{HoPeS}(C; \alpha) \rightarrow C(\alpha)$  from Lemma 20 induces the linear map  $f_*$  in 1-dimensional homology. Lemma 21 below analyzes what happens with  $f_*$  when a critical edge  $e$  is added to the  $\alpha$ -complex  $S = C(\alpha)$  and also to the graph  $G = \text{HoPeS}(C; \alpha)$  at the same scale  $\alpha$ .

**Lemma 21 (addition)** *Let an inclusion  $f : G \rightarrow S$  of a graph  $G$  into a simplicial complex  $S$  induce an isomorphism  $f_* : H_1(G) \rightarrow H_1(S)$ . Let us add a critical edge  $e$  to both  $G, S$ , which creates a new homology class  $\gamma \in H_1(S \cup e)$ . Then  $f_*$  extends to an isomorphism  $H_1(G \cup e) \rightarrow H_1(S \cup e)$ .*

*Proof.* Let  $L \subset G \cup e$  be a cycle containing the added edge  $e$ . Then  $H_1(G \cup e) \cong H_1(G) \oplus \langle [L] \rangle$ . Considering  $L$  as a cycle  $f(L) \subset S \cup e$ , we get  $H_1(S \cup e) \cong H_1(S) \oplus \langle [f(L)] \rangle$ . Mapping  $[L]$  to  $[f(L)] \in H_1(S \cup e)$ , we extend  $f_*$  to an isomorphism  $H_1(G) \oplus \langle [L] \rangle \rightarrow H_1(S) \oplus \langle [f(L)] \rangle$ .  $\square$

Lemma 22 analyzes the homology of the reduced skeleton  $G = \text{HoPeS}(C; \alpha)$  and  $S = C(\alpha)$  when a homology class  $\gamma$  dies in the homology  $H_1(S)$ .

**Lemma 22 (deletion)** *Let an inclusion  $f : G \rightarrow S$  of a graph  $G$  into a simplicial complex  $S$  induce an isomorphism  $f_* : H_1(G) \rightarrow H_1(S)$ . Let a homology class  $\gamma \in H_1(S)$  die after adding an open triangle  $T$  to the complex  $S$ . Let  $e$  be the longest open edge of the triangle  $T$ . Then  $f_*$  descends to an isomorphism  $H_1(G - e) \rightarrow H_1(S \cup T)$ .*

*Proof.* Adding the triangle  $T$  to  $S$  kills the homology class  $[\partial T]$  of the boundary  $\partial T$ , so  $H_1(S \cup T) \cong H_1(S) / \langle [\partial T] \rangle$ . Deleting the open edge  $e$  from  $\partial T \subset G$  makes the homology group smaller:  $H_1(G - e) \cong H_1(G) / \langle [\partial T] \rangle$ . Then the isomorphism  $f_*$  descends to the isomorphism  $H_1(G) / \langle [\partial T] \rangle \rightarrow H_1(S) / \langle [\partial T] \rangle$ .  $\square$

**Proposition 23** *For a cloud  $C \subset \mathbb{R}^2$  and  $\alpha > 0$ , the inclusion of the reduced graph  $\text{HoPeS}(C; \alpha) \rightarrow C(\alpha)$  induces an isomorphism of 1-dimensional homology.*

*Proof.* For any small enough  $\alpha > 0$ ,  $\text{HoPeS}(C; \alpha) = C(\alpha)$  is a disconnected cloud  $C$ , so their homology  $H_1$  is trivial. Each time when a homology class is born or dies in  $H_1(C(\alpha))$ , the isomorphism  $H_1(\text{HoPeS}(C; \alpha)) \cong H_1(C(\alpha))$  induced by the inclusion  $\text{HoPeS}(C; \alpha) \subset C(\alpha)$  is preserved by Lemmas 21 and 22.  $\square$

**Lemma 24** *Let  $L \subset C(\alpha)$  be a cycle that represents a homology class  $\gamma \in H_1(C(\alpha))$ . Then any longest edge  $e \subset L$  has a length at least  $2\text{birth}(\gamma)$ .*

*Proof.* Let a longest edge  $e$  of a cycle  $L \subset C(\alpha)$  representing  $\gamma$  have a half-length  $\alpha < \text{birth}(\gamma)$ . Then the cycle  $L$  enters the complex  $C(\alpha)$  earlier than  $\text{birth}(\gamma)$  and can not represent the class  $\gamma$  that starts living at  $\alpha = \text{birth}(\gamma)$ .  $\square$

By Definition 3 a forest  $\text{MST}(C; \alpha)$  on a cloud  $C \subset \mathbb{R}^2$  is obtained from a minimum spanning tree  $\text{MST}(C)$  by removing all edges longer than  $2\alpha$ .

**Proposition 25** *For a fixed scale  $\alpha > 0$ , let a graph  $G \subset C(\alpha)$  span a complex  $C(\alpha) \subset \mathbb{R}^2$  and  $H_1(G) \rightarrow H_1(C(\alpha))$  be the isomorphism induced by the inclusion. Let  $(b_i, d_i)$ ,  $i = 1, \dots, m$ , be all points in  $\text{PD}\{C(\alpha)\}$  counted with multiplicities within the ‘rectangle’  $\{\text{birth} \leq \alpha < \text{death}\}$ . Then the total length of  $G$  is bounded below by the total length of the forest  $\text{MST}(C; \alpha)$  plus  $2 \sum_{i=1}^m b_i$ .*

*Proof.* Let the subgraph  $G_1 \subset G$  consist of all non-splitting edges of  $G$  and  $e_1 \subset G_1$  be a longest open edge. Removing  $e_1$  from  $G$  makes  $H_1(G)$  smaller. Hence there is a cycle  $L_1 \subset G$  containing  $e_1$  and representing a homology class  $\gamma_1 \in H_1(C(\alpha))$  that corresponds to the (say) point  $(b_1, d_1) \in \text{PD}\{C(\alpha)\}$ , so  $\gamma_1$  lives over  $\text{birth}(\gamma_1) = b_1 \leq \alpha < d_1$ . The length satisfies  $|e_1| \geq 2b_1$  by Lemma 24.

Let the subgraph  $G_2 \subset G - e_1$  consist of all non-splitting edges and  $e_2 \subset G_2$  be a longest open edge. We similarly find the corresponding point  $(b_2, d_2)$  and conclude that  $|e_2| \geq b_2$  and so on. Finally, we get  $\sum_{i=1}^m |e_i| \geq 2 \sum_{i=1}^m b_i$ .

After removing  $e_1, \dots, e_m$ , the graph  $G - (e_1 \cup \dots \cup e_m)$  still spans  $C(\alpha)$ , because each time we chose a non-splitting edge. The total length of  $G - (e_1 \cup \dots \cup e_m)$  is not smaller than the total length of  $\text{MST}(C; \alpha)$  by Lemma 4.  $\square$

**Proof of Theorem 5.** For any  $\alpha > 0$ , the reduced graph  $\text{HoPeS}(C; \alpha)$  satisfies the homology condition by Proposition 23. Let classes  $\gamma_1, \dots, \gamma_m$  correspond to all  $m$  points in  $\text{PD}\{C(\alpha)\}$  counted with multiplicities in  $\{\text{birth} \leq \alpha < \text{death}\}$ . Then  $\gamma_1, \dots, \gamma_m$  form a basis of  $H_1(C(\alpha)) \cong H_1(\text{HoPeS}(C; \alpha))$  by Definition 7.

By Lemma 19 the total length of  $\text{HoPeS}(C; \alpha)$  equals the total length of  $\text{MST}(C; \alpha)$  plus  $2 \sum_{i=1}^m \text{birth}(\gamma_i)$ . By Proposition 25 this length is minimal over all graphs  $G \subset C(\alpha)$  that span  $C$  and have the same homology as  $C(\alpha)$ .  $\square$

A geometric approximation by the reduced skeleton  $\text{HoPeS}(C; \alpha)$  at any scale  $\alpha$  below will be used for proving the first approximation part of Theorem 7 below.

**Proposition 26** *Let  $C$  be any  $\varepsilon$ -sample of a compact set  $S \subset \mathbb{R}^2$ . Then the reduced skeleton  $\text{HoPeS}(C; \alpha)$  is  $(\alpha + \varepsilon)$ -close to  $S$  for any scale  $\alpha > 0$ .*

*Proof.* The cloud  $C$  is within the  $\varepsilon$ -offset  $S^\varepsilon$  of the shape  $S$ . By Definition 2 all edges of  $\text{HoPeS}(C; \alpha)$  have a half-length at most  $\alpha$ . The straight edge between any points  $p, q \in C$  is covered by the disks of the radius  $\alpha$  with the centers  $p, q$ , so  $\text{HoPeS}(C; \alpha) \subset C^\alpha \subset S^{\alpha+\varepsilon}$ . Since  $S \subset C^\varepsilon$  is covered by the  $\varepsilon$ -offset of  $\text{HoPeS}(C; \alpha)$ , the reduced skeleton  $\text{HoPeS}(C; \alpha)$  is indeed  $(\alpha + \varepsilon)$ -close to  $S$ .  $\square$

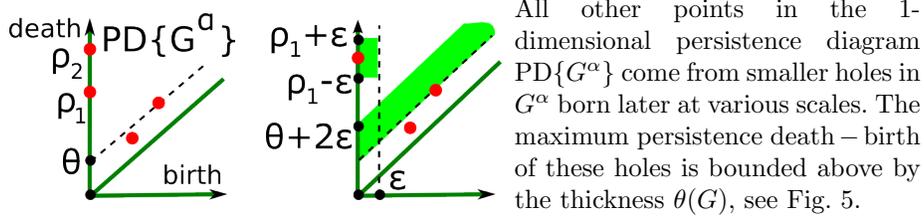
**Lemma 27** *For any cloud  $C \subset \mathbb{R}^2$ , the derived graph  $\text{HoPeS}'(C)$  is a subgraph of  $\text{HoPeS}(C; \alpha(C))$  at the critical scale  $\alpha(C)$  introduced in Definition 6.*

*Proof.* By Definition 2 all edges of  $\text{HoPeS}(C; \alpha(C))$  have lengths at most  $2\alpha(C)$ . Moreover,  $\text{death} > \alpha(C)$  for all critical edges. Definition 6 imposes the extra restriction on critical edges of  $\text{HoPeS}'(C)$ , namely each  $(\text{birth}, \text{death}) \in \text{PD}'\{C(\alpha)\}$  above the widest gap. So  $\text{HoPeS}'(C) \subset \text{HoPeS}(C; \alpha(C))$ .  $\square$

**Lemma 28** *The dimension of  $H_1(\text{HoPeS}'(C))$  equals the number of points  $(b, d)$  counted with multiplicities in the subdiagram  $\text{PD}'\{C(\alpha)\}$  subject to  $d > \alpha(C)$ .*

*Proof.* The dimension of  $H_1(\text{HoPeS}'(C))$  equals the number of homology classes whose  $(\text{birth}, \text{death})$  in  $\text{PD}\{C(\alpha)\}$  is above the widest gap and  $\text{death} > \alpha(C)$ . By Definition 6 any  $\text{birth} \leq \alpha(C)$ , so we count all  $(\text{birth}, \text{death}) \in \text{PD}'\{C(\alpha)\}$  with  $\text{birth} \leq \alpha(C) < \text{death}$ . All these  $m$  points are in a 1-1 correspondence with all critical edges of  $\text{HoPeS}'(C)$ . So the dimension of  $H_1(\text{HoPeS}'(C))$  is  $m$ .  $\square$

**Proof of Theorem 7.** The homology group  $H_1(G)$  is generated by the  $m$  basic cycles  $L_1, \dots, L_m$  that enclose  $m$  holes (bounded regions in the complement  $\mathbb{R}^2 - G$ ). These  $m$  cycles give points  $(0, \rho_i)$  in the vertical axis of  $\text{PD}\{G^\alpha\}$ .



The given condition  $\rho_1 > 7\epsilon + \theta(G) + \max_{i=1, \dots, m-1} \{\rho_{i+1} - \rho_i\}$  guarantees that the widest diagonal gap  $\{\theta(G) < y - x < \rho_1\}$  in  $\text{PD}\{G^\alpha\}$  is wider than any other gaps including the higher gaps of the widths  $\rho_{i+1} - \rho_i$ ,  $i = 1, \dots, m - 1$ .

By Stability Theorem 17 the perturbed diagram  $\text{PD}\{C(\alpha)\}$  of the cloud  $C$  is in the  $\epsilon$ -offset of  $\text{PD}\{G^\alpha\} \subset \cup_{i=1}^m (0, \rho_i) \cup \{y - x < \theta(G)\}$  with respect to the  $L_\infty$  metric on  $\mathbb{R}^2$ . All noisy points near the diagonal in  $\text{PD}\{C(\alpha)\}$  can not be higher than  $\theta(G) + 2\epsilon$  after projecting along the diagonal to the vertical axis.

The remaining points can not be lower than  $\rho_1 - 2\epsilon$  under the same projection. Hence the smaller diagonal strip  $\{\theta(G) + 2\epsilon < y - x < \rho_1 - 2\epsilon\}$  of the vertical width  $\rho_1 - 4\epsilon - \theta(G)$  is still empty in the perturbed diagram  $\text{PD}\{C(\alpha)\}$ .

By Stability Theorem 17 any point  $(0, \rho_i) \in \text{PD}\{G^\alpha\}$ ,  $i \geq 2$ , can not jump lower than the line  $y - x = \rho_i - 2\epsilon$  or higher than  $y - x = \rho_i + \epsilon$ . Then the widest diagonal gap between these perturbed points has a vertical width at most  $\max_{i=1, \dots, m-1} \{\rho_{i+1} - \rho_i\} + 3\epsilon$ . All points near the diagonal have diagonal gaps not wider than  $\theta(G) + 2\epsilon$ . Hence in all cases the 2nd widest gap in  $\text{PD}\{C(\alpha)\}$  has a vertical width smaller than  $\rho_1 - 4\epsilon - \theta(G)$ . Hence the 1st widest diagonal gap  $\text{dgap}(C)$  covers the strip  $\{\theta(G) + 2\epsilon < y - x < \rho_1 - 2\epsilon\} \subset \text{dgap}(G) \subset \text{PD}\{G^\alpha\}$ .

Hence the subdiagram  $\text{PD}'\{C(\alpha)\}$  above the line  $y - x = \rho_1 - 2\epsilon$  contains only perturbations of the original points  $(0, \rho_i)$  in the vertical strip  $\{0 \leq x \leq \epsilon\}$ . By Definition 6 the critical scale  $\alpha(C)$  is the maximum birth over all points in  $\text{PD}'\{C(\alpha)\}$ . These points are at most  $\epsilon$  away from their corresponding points  $(0, \rho_i)$  in the vertical axis, hence the critical scale  $\alpha(C)$  is bounded above by  $\epsilon$ .

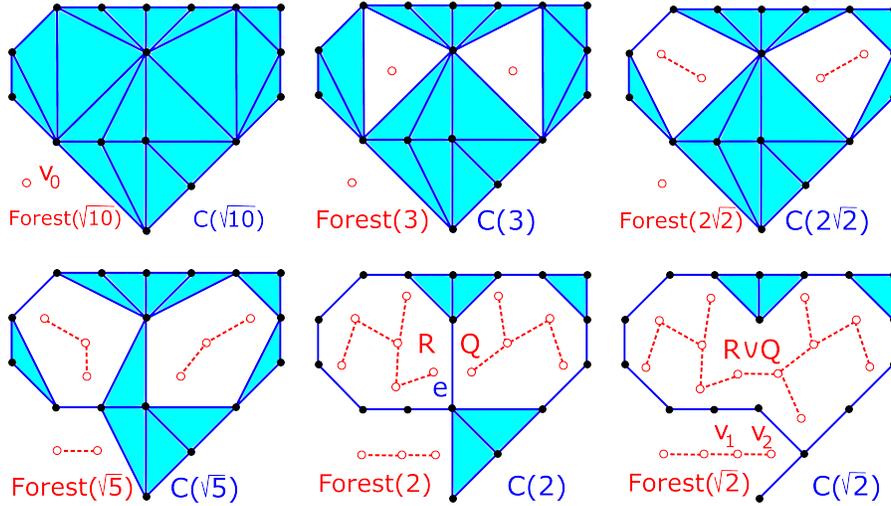
The deaths of all points in  $\text{PD}'\{C(\alpha)\}$  are larger than  $\rho_1 - 2\epsilon > \epsilon \geq \alpha(C)$ . Hence  $\text{HoPeS}'(C)$  contains all critical edges corresponding to the  $m$  points in  $\text{PD}'\{C(\alpha)\}$ , so  $H_1(\text{HoPeS}'(C))$  has the expected dimension  $m$ . The  $2\epsilon$ -closeness of  $\text{HoPeS}'(C)$  and  $G$  follows from Proposition 26 and Lemma 27 for the set  $S = C$  after replacing the critical scale  $\alpha(C)$  by its upper bound  $\epsilon$ .  $\square$

**Proof of Corollary 8.** The condition that the perturbed cloud  $\tilde{C}$  is  $\delta$ -close to the original cloud  $C$ , which is  $\epsilon$ -closed to the graph  $G$ , implies that  $\tilde{C}$  is  $(\delta + \epsilon)$ -close to  $G$ . Theorem 7 for the  $\epsilon$ -sample  $C$  and  $(\delta + \epsilon)$ -sample  $\tilde{C}$  of  $G$  says that  $\text{HoPeS}'(C)$  is  $2\epsilon$ -close to  $G$  and  $\text{HoPeS}'(\tilde{C})$  is  $(2\delta + 2\epsilon)$ -close to  $G$ . Hence  $\text{HoPeS}'(C)$  and  $\text{HoPeS}'(\tilde{C})$  is  $(2\delta + 4\epsilon)$ -close as required.  $\square$

## Appendix C: algorithm for HoPeS( $C$ ) and HoPeS'( $C$ )

The only input is a cloud  $C$  of  $n$  points  $(x_i, y_i)$ ,  $i = 1, \dots, n$  with real coordinates. There are no user-defined parameters. The outputs are the persistence diagram  $\text{PD}\{C(\alpha)\}$ , a Homologically Persistent Skeleton  $\text{HoPeS}(C)$ ,  $\text{HoPeS}'(G)$  and its simplified version, all found in time  $O(n \log n)$  with  $O(n)$  space.

**Forest( $\alpha$ ).** Given a cloud of  $n$  points  $C \subset \mathbb{R}^2$ , we build a Delaunay triangulation  $\text{Del}(C)$ . We shall maintain a union-find structure  $\text{Forest}(\alpha)$  on abstract nodes that are in a 1-1 correspondence with all regions of  $\mathbb{R}^2 - C(\alpha)$ . Initially  $\text{Forest}(\alpha)$  is one node  $v_0$  corresponding to the external region of  $C(+\infty) = \text{Del}(C)$ . If the scale  $\alpha$  is decreasing, then edges and triangles are disappearing from  $\text{Del}(C)$  and we gradually get smaller  $\alpha$ -complexes  $C(\alpha)$ .



**Fig. 13.** Trees of  $\text{Forest}(\alpha)$  represent regions of  $\mathbb{R}^2 - C(\alpha)$ .

**Shrinking  $C(\alpha)$ .** Removing a triangle from  $C(\alpha)$  adds one node  $v$  to  $\text{Forest}(\alpha)$ . When an acute triangle  $T$  is removed at the scale  $\alpha$  equal to the circumradius of  $T$ , an isolated node  $v$  appears in  $\text{Forest}(\alpha)$ . Any non-acute triangle  $T$  is removed together with its longest edge  $e$  at the scale  $\alpha$  equal to the half-length of  $e$ . If  $e$  was shared by two triangles  $S, T$ , their corresponding nodes are linked in  $\text{Forest}(\alpha)$ , so  $T$  joins the region of  $\mathbb{R}^2 - C(\alpha)$  already containing  $S$ . Fig. 13 shows how the nodes  $v_1, v_2$  corresponding to right-angled triangles in the complex  $C(2)$  are joined in  $\text{Forest}(\sqrt{2})$ .

**Merging.** In the most interesting case we remove an edge  $e$  between two acute triangles  $S, T$  from different regions  $R, Q \subset \mathbb{R}^2 - C(\alpha)$ . Then  $R, Q$  merge into  $R \cup Q$  and the corresponding components of  $\text{Forest}(\alpha)$  are linked. The edge  $e$

between  $Q, R$  is critical by Definition 17. So we maintain a 1-1 correspondence between all connected regions of  $\mathbb{R}^2 - C(\alpha)$  and all trees of  $\text{Forest}(\alpha)$ . Fig. 13 shows red dashed arcs joining centers of adjacent triangles for simplicity. Real links in  $\text{Forest}(\alpha)$  are added to guarantee balanced trees for a faster search.

**Critical edges and MST.** In comparison with a standard algorithm for 0-dimensional persistence [3] we add to each pair (birth, death) the critical edge  $e$  whose removal from  $C(\alpha)$  caused the death of the younger region, say  $Q$ . The older region  $R$  survives and absorbs  $Q$  by the ‘elder’ rule of persistence. When only two remaining regions merge into one, the resulting complex  $C(\alpha)$  is connected if we keep all edges having the same region on both sides. So the final complex  $C(\alpha)$  contains  $\text{MST}(C)$  and no critical edges by construction.

**Output  $\text{HoPeS}'(C)$ .** We convert the array (birth, death) into the persistence diagram  $\text{PD}\{C^\alpha\}$ . Then we find all  $m$  points (birth, death) above the widest diagonal gap in  $\text{PD}\{C^\alpha\}$ . The corresponding  $m$  critical edges that were recorded together with the pairs (birth, death) and can now be added to  $\text{MST}(C)$  giving the derived graph  $\text{HoPeS}'(C)$  by Definition 6. If we add the critical edges corresponding to all points in  $\text{PD}\{C^\alpha\}$ , we get the full graph  $\text{HoPeS}(C)$ .

**Simplify  $\text{HoPeS}'(C)$ .** The critical scale  $\alpha(C)$  from Definition 6 is a lower estimate for the noise bound  $\varepsilon$  by Theorem 7. Then  $2\alpha(C)$  is a lower estimate for the distance between  $\varepsilon$ -perturbations of the same point. Hence we can simplify  $\text{HoPeS}'(C)$  using the critical scale  $\alpha(C)$  without any ad-hoc heuristics. We remove all paths to degree 1 vertices that have a length up to  $2\alpha(C)$  or whose endpoints are within  $2\alpha(C)$ -offset of non-splitting edges of  $\text{HoPeS}'(C)$ .

Then we collapse all short edges of length up to  $2\alpha(C)$  between vertices of  $\text{deg} \neq 2$ . For simplicity, all remaining paths between vertices of  $\text{deg} \neq 2$  are approximated by polygonal lines with edge-lengths of about  $2\alpha(C)$ . Better curve-fittings will certainly give smoother simplifications.

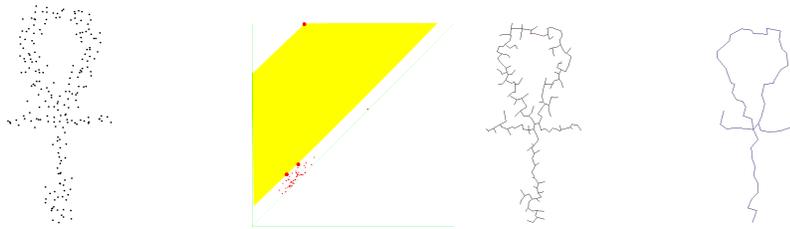
We shall upload the C++ code at <http://kurlin.org> in May 2015.

## Appendix D: more practical experiments with $\text{HoPeS}'(C)$

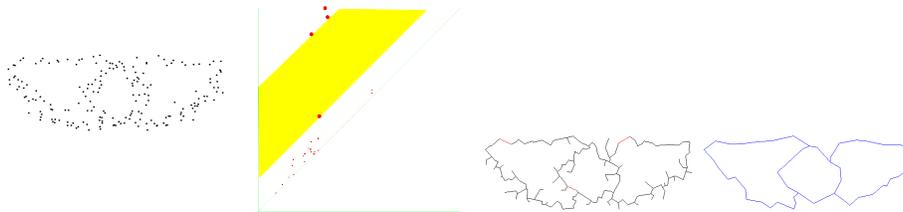
The outputs for hieroglyphs O45, D33 are in section 6. Fig. 7–18 show results for other Egyptian hieroglyphs in Fig. 14 from [http://en.wikipedia.org/wiki/List\\_of\\_Egyptian\\_hieroglyphs\\_by\\_alphabetization](http://en.wikipedia.org/wiki/List_of_Egyptian_hieroglyphs_by_alphabetization). A few red dots with highest persistence are fat in  $\text{PD}\{C(\alpha)\}$  and widest diagonal gaps are yellow.



Fig. 14. Images of hieroglyphs O45, D33, S34, W3, O42, N24.



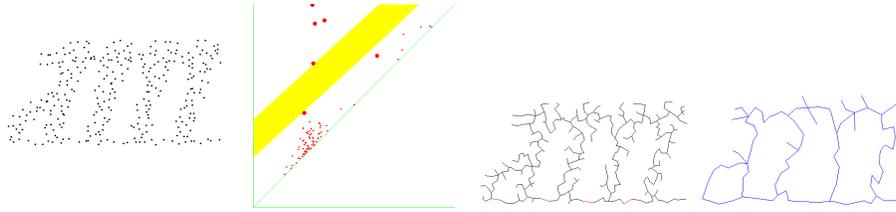
**Fig. 15.** A sample  $C$  of S34, diagram  $PD\{C(\alpha)\}$ ,  $HoPeS'(C)$  and its simplification.



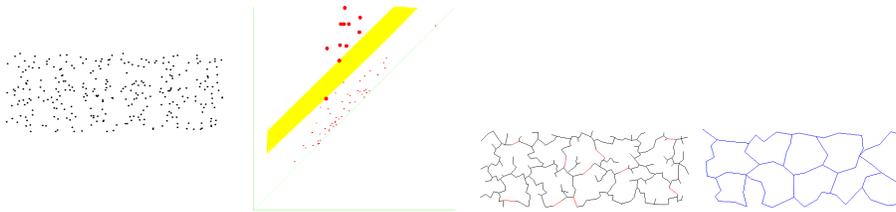
**Fig. 16.** A sample  $C$  of W3, diagram  $PD\{C(\alpha)\}$ ,  $HoPeS'(C)$  and its simplification.

In Fig. 19 and 20 we used lower thresholds for selecting feature points in the image BSD42049 from Fig. 9 in section 6. So  $C$  has more points than in Fig. 9, but the persistence diagrams  $PD\{C(\alpha)\}$  still have a widest gap separating 2 red points from the noise near the diagonal. The derived skeletons  $HoPeS'(G)$  and their simplifications have more branches presenting finer details of the input.

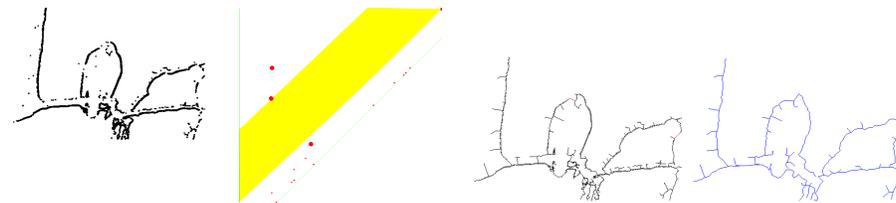
The derived skeleton  $HoPeS'(C)$  in Definition 6 is based on a widest gap in the diagram  $PD\{C^\alpha\}$ . The red points above this 1st widest gap correspond to critical red edges that we added to  $MST(C)$  to get  $HoPeS'(C)$ . Instead of the 1st widest gap, we may take the 2nd gap and so on to get more derived graphs  $HoPeS''(C)$  etc. Our algorithm produces a hierarchy of skeletons ordered by persistence of cycles hidden in  $C$ . All these skeletons are extracted from  $PD\{C^\alpha\}$  in the same time  $O(n \log n)$ . Indeed, all critical edges were found in our data structure  $Map(\alpha)$  when computing  $PD\{C^\alpha\}$ . Fig. 21 and 22 show the 1st and 2nd derived skeltons for more challenging images.



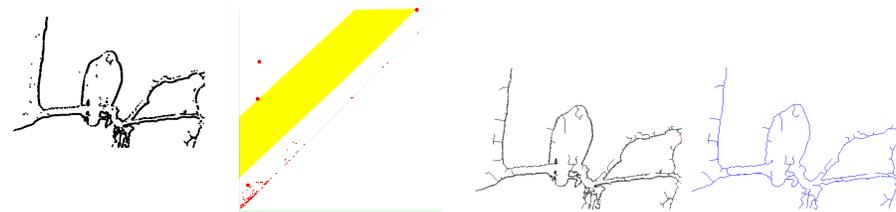
**Fig. 17.** A sample  $C$  of O42, diagram  $\text{PD}\{C(\alpha)\}$ ,  $\text{HoPeS}'(C)$  and its simplification.



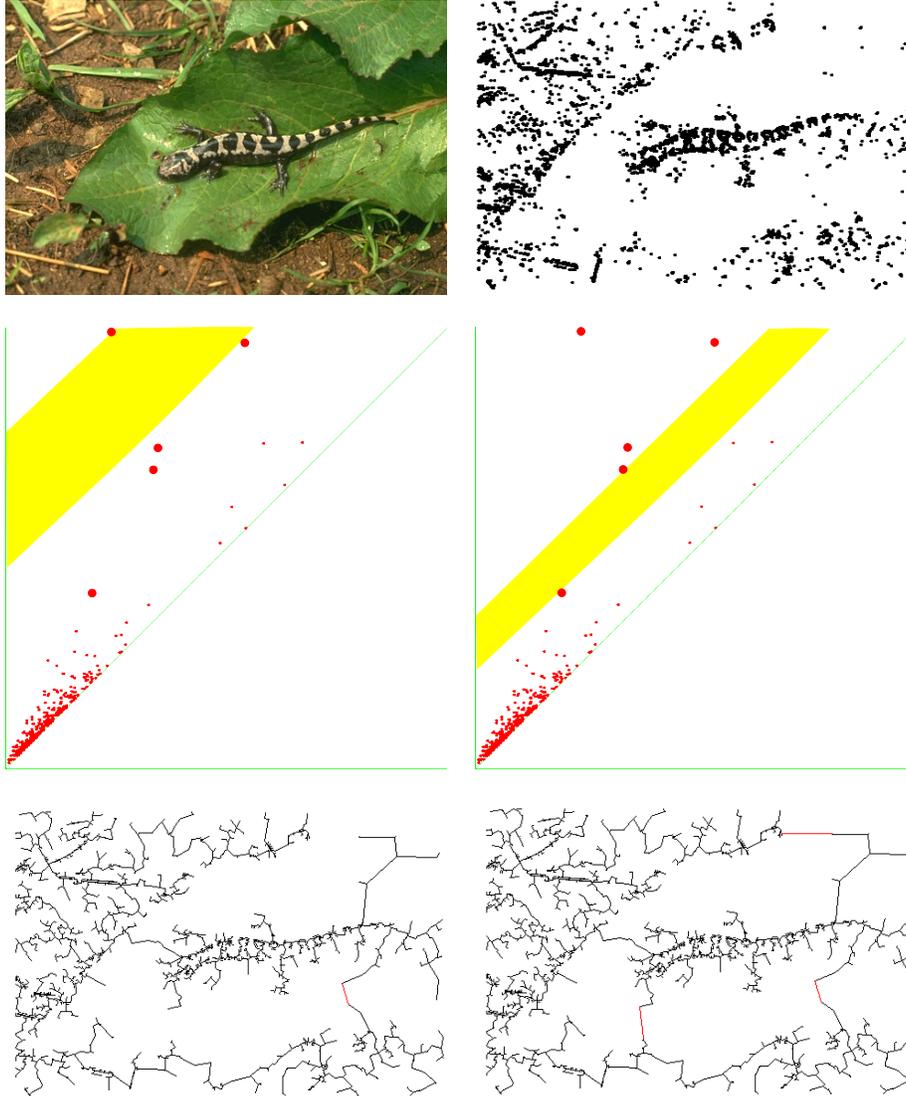
**Fig. 18.** A sample  $C$  of N24, diagram  $\text{PD}\{C(\alpha)\}$ ,  $\text{HoPeS}'(C)$  and its simplification.



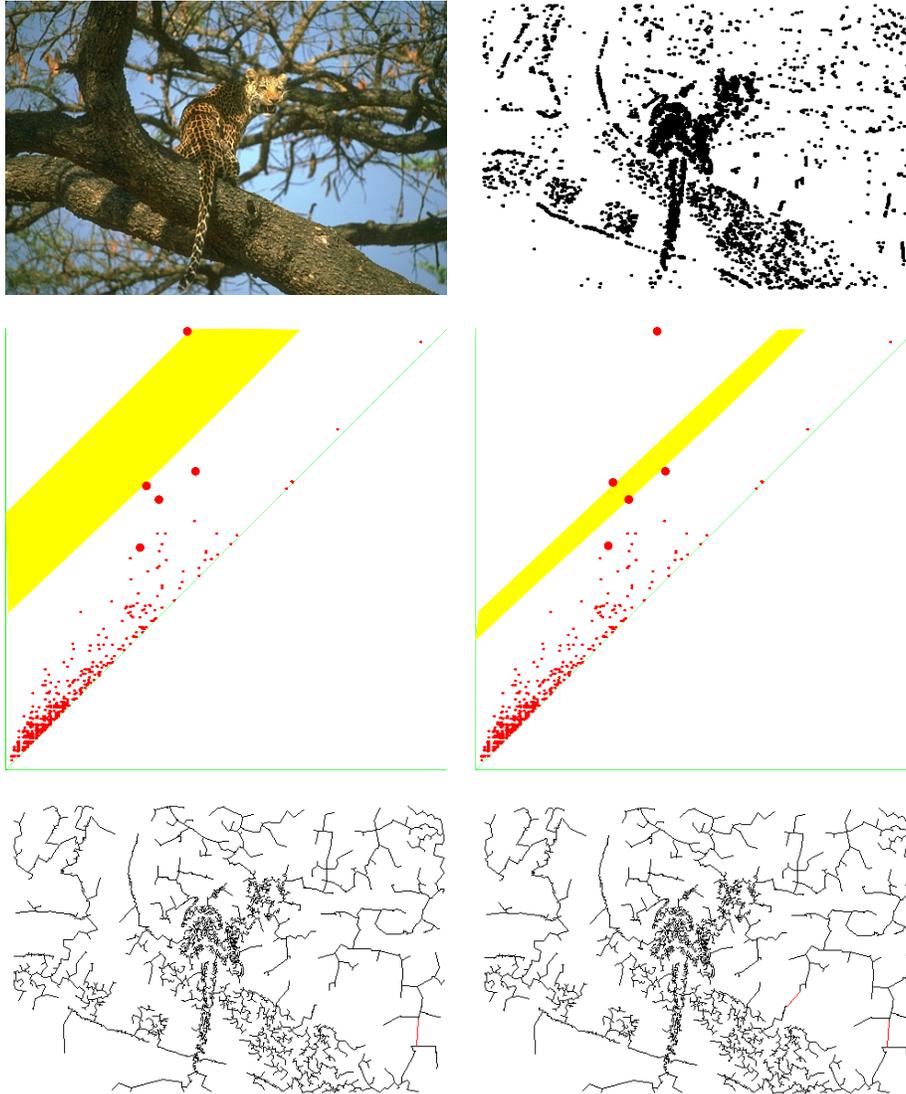
**Fig. 19.**  $C = \{2664 \text{ points in BSD42049}\}$ ,  $\text{PD}\{C(\alpha)\}$ ,  $\text{HoPeS}'(C)$  and simplification.



**Fig. 20.**  $C = \{3604 \text{ points in BSD42049}\}$ ,  $\text{PD}\{C(\alpha)\}$ ,  $\text{HoPeS}'(C)$  and simplification.



**Fig. 21.** First row: BSD175083 and cloud  $C$  of 3928 points. Second row:  $\text{PD}\{C(\alpha)\}$  with 1st and 2nd widest yellow gaps. Third row: derived graphs with 1 cycle (1 point above the 1st gap in  $\text{PD}\{C(\alpha)\}$ ) and with 4 cycles (4 points above the 2nd gap).



**Fig. 22.** First row: BSD134049 and cloud  $C$  of 5419 points. Second row:  $\text{PD}\{C(\alpha)\}$  with 1st and 2nd widest yellow gaps. Third row: the derived graphs with 1 cycle (1 point above the 1st gap in  $\text{PD}\{C(\alpha)\}$ ) and with 2 cycles (2 points above the 2nd gap).