A fast and robust algorithm to count topologically persistent holes in noisy clouds

Vitaliy Kurlin
Durham University
Department of Mathematical Sciences, Durham, DH1 3LE, United Kingdom
vitaliy.kurlin@gmail.com, http://kurlin.org

Abstract

Preprocessing a 2D image often produces a noisy cloud of interest points. We study the problem of counting holes in noisy clouds in the plane. The holes in a given cloud are quantified by the topological persistence of their boundary contours when the cloud is analyzed at all possible scales.

We design the algorithm to count holes that are most persistent in the filtration of offsets (neighborhoods) around given points. The input is a cloud of n points in the plane without any user-defined parameters. The algorithm has $O(n \log n)$ time and $O(n)$ space. The output is the array (number of holes, relative persistence in the filtration).

We prove theoretical guarantees when the algorithm finds the correct number of holes (components in the complement) of an unknown shape approximated by a cloud.

1. Introduction: counting holes in noisy clouds

We apply methods from the new area of topological data analysis to counting persistent holes in a noisy cloud of points. Such a cloud can be obtained by selecting interest points in a gray scale or RGB image. Our region-based method uses global topological properties of contours.

By a *shape* we mean any subset $X \subset \mathbb{R}^2$ that can be split into finitely many (topological) triangles. Hence X is bounded, but may not be connected. Then a *hole* in a shape $X \subset \mathbb{R}^2$ is a bounded connected component of the complement $\mathbb{R}^2 - X$. Such a hole can be a disk, a ring or may have a more complicated topological form, see Fig. 1.

The α*-offset* X^α is the union $\cup_{p\in X} B(p;\alpha)$ of disks with the radius $\alpha \geq 0$ and centers at all $p \in X$. For instance, X^0 is the original shape $X \subset \mathbb{R}^2$. When α is increasing, the holes of $\mathbb{R}^2 - X^\alpha$ are shrinking, may split into smaller newborn holes and will eventually die, each at its own death time α, see Fig. 3. The *persistence* of a hole is its life span $\mathrm{death} - \mathrm{birth}$ in the filtration $\{X^\alpha\}$ of all α-offsets. So we quantify holes by their persistence at different scales α.

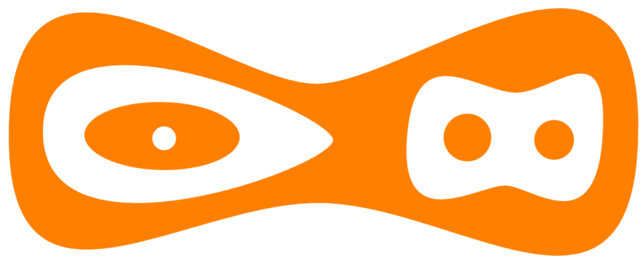

Figure 1. The orange shape $X \subset \mathbb{R}^2$ with 3 white holes of different forms: a small disk, a ring-like hole, a 'figure-eight' hole.
.

Hole counting problem. Let a shape X be represented by a finite sample C of points in $\mathbb{R}^2$. Find conditions on X and its sample when one can quickly count persistent holes.

We solve the problem by the algorithm HoCToP : Hole Counting based on Topological Persistence. The only input is a finite cloud C of n points approximating an unknown shape $X \subset \mathbb{R}^2$. The algorithm outputs the relative persistence of k holes in the filtration $\{C^\alpha\}$ for all $k \geq 0$. If the scale α is random and uniform, this output gives probabilities $P(k \text{ holes})$. The boundary edges of persistent holes can be quickly post-processed to extract all boundary contours.

Figure 2. Input: cloud C of 1251 points uniformly sampled from the shape with 3 holes in Fig. 1. Output probabilities of HoCToP : $P(3 \text{ holes}) \approx 24\%$, $P(2 \text{ holes}) \approx 13\%$, $P(8 \text{ holes}) \approx 11\%$.

Theorems 1, 4 say that the algorithm HoCToP quickly and correctly finds all persistent holes using only a good enough sample C of an unknown shape X, see section 2.

2. Main results: the algorithm and guarantees

We start from a high-level description of our algorithm.

The topological persistence of contours in the filtration $\{C^\alpha\}$ is computed by using a Delaunay triangulation $\mathrm{Del}(C)$ of a given cloud $C \subset \mathbb{R}^2$ of n points. By Nerve Lemma 8 the α-offsets C^α can be continuously deformed to the α-complexes $C(\alpha)$, which filter $\mathrm{Del}(C)$ as follows: $C = C(0) \subset \cdots \subset C(\alpha) \subset \cdots \subset C(+\infty) = \mathrm{Del}(C)$. Each $C(\alpha)$ has some edges and triangles from $\mathrm{Del}(C)$.

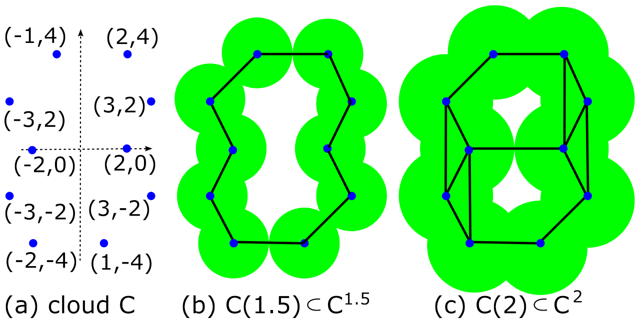

Figure 3. The big hole in the green offset C^α is born at $\alpha = 1.5$, splits into 2 smaller holes at $\alpha = 2$ and dies at $\alpha \approx 2.577$, so the topological persistence of this hole is $\mathrm{death} - \mathrm{birth} \approx 1.077$.

The graph dual to $\mathrm{Del}(C)$ is filtered by the subgraphs $C^*(\alpha)$ whose connected components correspond to holes in $C(\alpha)$. When α is decreasing, $C(\alpha)$ is shrinking, so its holes are growing and corresponding components of $C^*(\alpha)$ merge at critical values of α, see Fig. 6. The persistence of cycles in the filtration $\{C^\alpha\}$ corresponds to the persistence of components in $\{C^*(\alpha)\}$, see Duality Lemma 14.

The pairs $(\mathrm{birth}, \mathrm{death})$ of connected components in $\{C^*(\alpha)\}$ are found via a union-find structure by adding edges and merging components. So computing the 1-dimensional persistence of cycles in $\{C^\alpha\}$ reduces to the 0-dimensional persistence of components in $\{C^*(\alpha)\}$.

Starting from a given cloud $C \subset \mathbb{R}^2$ of n points with real coordinates (x_i, y_i), $i = 1, \ldots, n$, we find a Delaunay triangulation $\mathrm{Del}(C)$ in $O(n \log n)$ time with $O(n)$ space. Then we remove each edge of $\mathrm{Del}(C)$ one by one in the decreasing order of their length. Removing an edge may break a contour when adjacent regions in $C(\alpha)$ and the corresponding components of $C^*(\alpha)$ merge. In the case of a merger, a younger component of $C^*(\alpha)$ and the corresponding hole in $C(\alpha)$ die. We note the birth and death of each dead hole. We get the probability of k holes as the relative length of all intervals of the scale α when $C^\alpha \subset \mathbb{R}^2$ has k holes.

Theorem 1. *The algorithm* HOCTOP *counts all holes in a given cloud $C \subset \mathbb{R}^2$ of n points in $O(n \log n)$ time with $O(n)$ space. All holes are ordered by their topological persistence in the ascending filtration $\{C^\alpha\}$ of the α-offsets.*

Definition 2 (ε-sample)**.** *A cloud C is an* ε-sample *of a shape $X \subset \mathbb{R}^2$ if $X \subset C^\alpha$ and $C \subset X^\alpha$. So any point of C is within the distance ε from a point of X and any point of X is at most ε away from a point of C. Hence ε can be considered as the upper bound of some arbitrary noise.*

Definition 3 (min and max homological feature sizes)**.** *For any shape $X \subset \mathbb{R}^2$, let $\alpha = \mathrm{minhfs}(X)$ be the minimum* homological feature size *when a first hole is born or dies in X^α. Let $\alpha = \mathrm{maxhfs}(X)$ be the maximum* homological feature size *after which no holes are born or die in X^α.*

Theorem 4 gives sufficient (not necessary) conditions when the algorithm finds the correct number of holes in an unknown shape $X \subset \mathbb{R}^2$ that is represented by its finite sample C. We extend the Homology Inference Theorem [1] to the case when the upper bound ε of noise is unknown.

Theorem 4. *Let a cloud C be an ε-sample of a shape $X \subset \mathbb{R}^2$ with an unknown parameter ε such that $\mathrm{minhfs}(X) > \frac{1}{2}\mathrm{maxhfs}(X) + 2\varepsilon$. Then the algorithm* HOCTOP *finds the correct number of holes in X by using only the cloud C.*

The condition $\mathrm{minhfs}(X) > \frac{1}{2}\mathrm{maxhfs}(X) + 2\varepsilon$ means that all holes of X, which are bounded components of $\mathbb{R}^2 - X$, have comparable sizes (neither tiny nor huge).

Even if the conditions of Theorem 4 are not satisfied, we can always find the number k of holes with the highest probability. The algorithm HOCTOP can also accept a signal-to-noise ratio τ and output all holes whose persistence is larger than τ. Alternatively, the user may prefer to get most likely outputs ordered by the probability $P(k \text{ holes})$.

3. Previous work on computing persistence

The offsets C^α of a finite cloud C are usually studied through the Čech or Rips complexes, which may contain up to $O(n^k)$ simplices in all dimensions $k \leq n - 1$ even if $C \subset \mathbb{R}^2$. A Delaunay triangulation has the advantage of a smaller size up to $m = O(n^2)$ in dimensions $n = 2, 3, 4$.

The fastest algorithm [7] for computing persistence of a filtration in all dimensions has the same running time $O(m^{2.376})$ in the number m of simplices as the best known time for the multiplication of two $m \times m$ matrices.

The only faster (almost linear) algorithm was known in dimension 0 for any filtration [4, p. 6–8] and in dimension 1 for a filtration of sublevels of a function on a fixed triangulation of a closed surface [4, p. 159–160]. In the latter case functions are monotonic and Morse, so vertices join a filtration one by one, not all at once as in the α-offsets C^α.

There is an $O(n \log n)$ algorithm [5] reconstructing the Reeb graph (hence the 1-dimensional homology) from a k-neighbor complex that ε-approximates an unknown metric graph G. This algorithms needs k as an input parameter.

Two extra parameters were used in a Delaunay-based image segmentation [6]: α for the radius of disks centered at points of a cloud C and p for a desired level of persistence.

4. Delaunay triangulation and α-complexes

For simplicity, we introduce only 2-dimensional simplicial complexes that are relevant to clouds in the plane.

Definition 5 (simplicial complex)**.** *A simplicial 2-complex is a finite set of* simplices *(vertices, edges, triangles):*

- *the sides of any triangle are included in the complex;*
- *the endpoints of any edge are included in the complex;*
- *two triangles can intersect only along a common edge;*
- *edges can meet only at a common endpoint (a vertex);*
- *an edge can not pierce through the interior of a triangle.*

If a complex S is drawn in $\mathbb{R}^n$ without self-intersections, we may call this image $|S|$ a *geometric realization* of S. We have defined a shape $X \subset \mathbb{R}^2$ as a geometric realization of a 2-complex. For instance, a round disk whose boundary is split into 3 edges by 3 vertices is a topological triangle.

A *cycle* in a complex is a sequence of edges $e_1, \dots, e_m$ such that any consecutive edges e_i, e_{i+1} (in the cyclic order) have a common vertex. Any loop in a geometric realization $|S|$ continuously deforms to a cycle of edges in S.

Definition 6 (Delaunay triangulation Del)**.** *For a point p_i in a cloud $C = \{p_1, \dots, p_n\} \subset \mathbb{R}^2$, the* Voronoi cell $V(p_i) = \{q \in \mathbb{R}^2 : d(p_i, q) \leq d(p_j, q)\ \forall j \neq i\}$ *is the set of all points q that are (non-strictly) closer to p_i than to other points of C. The* Delaunay triangulation $\mathrm{Del}(C)$ *is the nerve of the Voronoi diagram $\cup_{p \in C} V(p)$. Namely, $p, q, r \in C$ span a triangle if and only if $V(p) \cap V(q) \cap V(r) \neq \emptyset$.*

By another definition [2, section 9.1] the circumcircle of any *Delaunay triangle* in $\mathrm{Del}(C)$ encloses no points of C.

For a cloud $C \subset \mathbb{R}^2$ of n points, let $\mathrm{Del}(C)$ have k triangles and b boundary edges in the external region. Counting all E edges over triangles, we get $3k + b = 2E$. Euler's formula $n - E + (k + 1) = 2$ implies that $k = 2n - b - 2$, $E = 3n - b - 3$. So $\mathrm{Del}(C)$ has $O(n)$ edges and triangles.

Definition 7 (α-complex $C(\alpha)$)**.** *For a scale parameter $\alpha > 0$, the* α-complex $C(\alpha)$ *is the nerve of $\cup_{p \in C}(V(p) \cap B(p; \alpha))$, see [4, section III.4]. Points $p, q \in C$ are connected by an edge if $V(p) \cap B(p; \alpha)$ meets $V(q) \cap B(q; \alpha)$. Three points $p, q, r \in C$ span a triangle if the intersection $V(p) \cap B(p; \alpha) \cap V(q) \cap B(q; \alpha) \cap V(r) \cap B(r; \alpha) \neq \emptyset$.*

If $\alpha > 0$ is very small, all points of C are disjoint in $C(\alpha)$, while $C(\alpha) = \mathrm{Del}(C)$ for any large enough α, see examples in Fig. 3. So all α-complexes form the filtration $C = C(0) \subset \dots \subset C(\alpha) \subset \dots \subset C(+\infty) = \mathrm{Del}(C)$. Edges or triangles are added only at *critical values* of α.

Lemma 8 (Nerve of a ball covering [3])**.** *The union of balls $C^\alpha = \cup_{p \in C} B(p; \alpha)$ continuously deforms to (has the homotopy type of) a geometric realization of $C(\alpha)$.*

5. Persistent homology: definitions, examples

Definition 9 (1-dimensional homology H_1)**.** *We consider the 1-dimensional homology group $H_1(S)$ only with coefficients in $\mathbb{Z}/2\mathbb{Z} = \{0, 1\}$. Cycles of a 2-dimensional complex S can be algebraically written as linear combinations of edges (with coefficients 0 or 1) and generate the vector space C_1 of cycles. The boundaries of all triangles in S (as cycles of 3 edges) generate the subspace $B_1 \subset C_1$. The quotient group C_1/B_1 is the* homology *group* $H_1(S)$.

Let $\{S(\alpha)\}$ be an ascending filtration of complexes, where edges or triangles are added at finitely many critical values $\alpha_1, \dots, \alpha_m$. The inclusions $S(\alpha_1) \subset \dots \subset S(\alpha_m)$ induce the linear maps $H_1(S(\alpha_1)) \to \dots \to H_1(S(\alpha_m))$.

Definition 10 (persistence diagram PD)**.** *In a filtration $\{S(\alpha)\}$ a cycle γ is* born *at* birth time α_i *and dies at α_j if the homology class of γ belongs to the homology $H_1(S(\alpha))$ only for $\alpha_i \leq \alpha < \alpha_j$. So γ has* $(\mathrm{birth}, \mathrm{death}) = (\alpha_i, \alpha_j)$ *and the persistence* $\mathrm{pers}(\gamma) = \mathrm{death} - \mathrm{birth}$. *The* persistence diagram $\mathrm{PD}\{S(\alpha)\}$ *consists of the diagonal $\{x = y\}$ and the points* $(\mathrm{birth}(\gamma), \mathrm{death}(\gamma))$ *over all cycles γ.*

Pairs with a low persistence $\mathrm{death} - \mathrm{birth}$ (close to the diagonal $\{x = y\}$ in PD) are treated as noise. Pairs with a high persistence represent persistent cycles in $\{S(\alpha)\}$.

We shall consider the filtrations of α-offsets $\{X^\alpha\}$ and $\{C^\alpha\}$ for a shape $X \subset \mathbb{R}^2$ and a finite cloud $C \subset \mathbb{R}^2$. Figures 4 and 5 show the persistence diagram PD for the filtration of the α-offsets C^α equivalent to $C(\alpha)$ by Lemma 8.

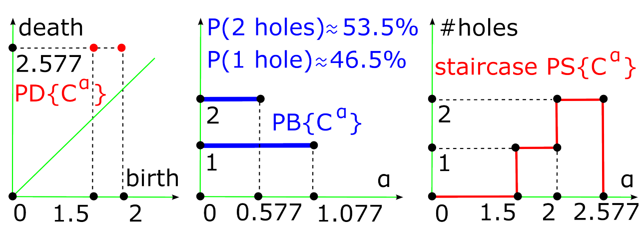

Figure 4. Extra outputs for the cloud C of 10 points in Fig. 3. Left: persistence diagram, middle: barcode, right: persistence staircase.

We can convert the persistence diagram into the persistence barcode $\mathrm{PB}\{C^\alpha\}$. All pairs $(\mathrm{birth}, \mathrm{death})$ give horizontal bars ordered by their length $\mathrm{death} - \mathrm{birth}$. Usually the bars are drawn from the left endpoint 0 to the right endpoint $\mathrm{death} - \mathrm{birth}$, see the middle picture in Fig. 4.

We suggest one more way to visualize persistence. Each pair $(\mathrm{birth}, \mathrm{death})$ defines the function $f(\alpha) = 1$ for $\mathrm{birth} \leq \alpha < \mathrm{death}$ and $f(\alpha) = 0$ otherwise. The sum of

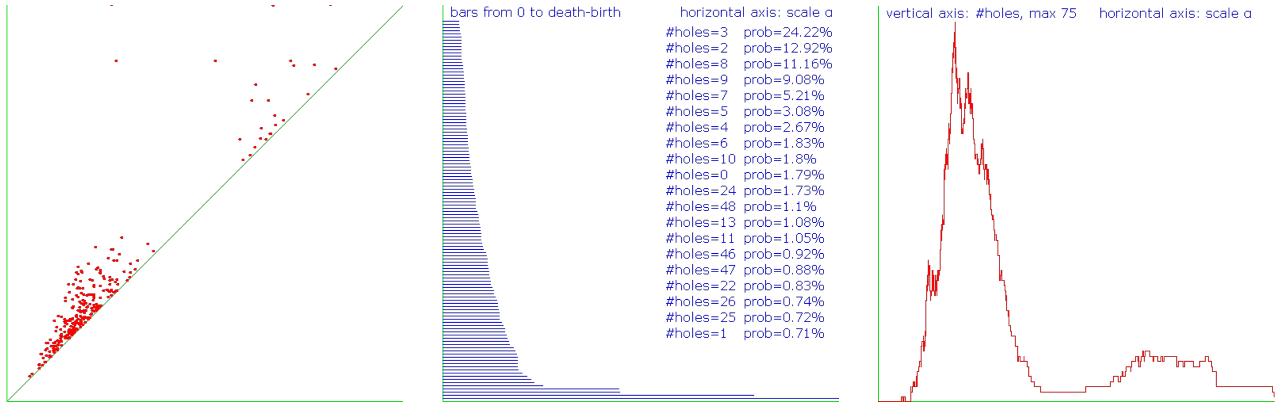

Figure 5. Extra outputs for the cloud C of 1251 points in Fig. 2. Left: persistence diagram PD, middle: barcode PB, right: staircase PS.

these functions over all pairs gives the *persistence staircase* $\mathrm{PS}\{C^\alpha\}$. The value of this piecewise constant function of α is the number of holes in the offset C^α. We have connected consecutive horizontal segments of $\mathrm{PS}\{C^\alpha\}$ to get a 'continuous' staircase as in the right picture of Fig. 4.

For the cloud C of 10 points in Fig. 3, the full range of the scale α is from the smallest critical value $\alpha = 1.5$ (when a first hole is born) to the largest critical value $\alpha = \frac{5}{8}\sqrt{17} \approx 2.577$ (when both final holes die). The output probability $P(1 \text{ hole}) \approx 46.5\%$ is the contribution of the interval $(1.5, 2)$ to the full range $1.5 \le \alpha \le \frac{5}{8}\sqrt{17}$. The largest probability $P(2 \text{ holes}) \approx 53.5\%$ is the contribution of the interval $(2, \frac{5}{8}\sqrt{17})$ when C^α has exactly 2 holes.

For the cloud C of 1251 points in Fig. 2, we scaled $\mathrm{PB}\{C^\alpha\}$ and $\mathrm{PS}\{C^\alpha\}$ along the horizontal α-axis and kept only the longest bars in the barcode $\mathrm{PB}\{C^\alpha\}$ in Fig. 5.

6. Persistent homology: stability and duality

Definition 11 (bottleneck distance d_B)**.** *Let the distance between* $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$ *be* $\max\{|x_1 - x_2|, |y_1 - y_2|\}$. *The* bottleneck distance $d_B(\mathrm{PD}_1, \mathrm{PD}_2)$ *between persistence diagrams* $\mathrm{PD}_1, \mathrm{PD}_2$ *is the smallest* ε *such that* PD_1 *is in the* ε*-offset of* PD_2 *and* PD_2 *is in the* ε*-offset of* PD_1.

We quote the simple version of the Persistence Stability Theorem [1] only for persistence diagrams of α-offsets.

Theorem 12. *[1] If a finite cloud C of points is an ε-sample of a shape $X \subset \mathbb{R}^2$, then $d_B(\mathrm{PD}\{X^\alpha\}, \mathrm{PD}\{C^\alpha\}) \le \varepsilon$.*

Stability Theorem 12 implies for barcodes PB that the endpoints of all bars are perturbed by at most ε. So a long bar can become only a bit shorter after adding noise.

To every triangle in the Delaunay triangulation $\mathrm{Del}(C)$, let us associate a single abstract vertex v_i, $i = 1, \ldots, k$. It will be convenient to call the external region of $\mathrm{Del}(C)$ also a 'triangle' and represent it by an extra vertex v_0.

Definition 13 (graphs $C^*(\alpha)$)**.** *For any vertices v_i, v_j representing adjacent triangles in* $\mathrm{Del}(C)$*, let d_{ij} be the length of the (longest) common edge of the triangles. The metric graph C^* dual to* $\mathrm{Del}(C)$ *has the vertices $v_0, v_1, \ldots, v_k$ and edges of the length d_{ij} connecting vertices v_i, v_j that represent adjacent triangles, see Fig. 6. The graph C^* is filtered by the subgraphs $C^*(\alpha)$ that have only the edges of a length $d_{ij} > 2\alpha$. We remove any isolated node v (except v_0) from $C^*(\alpha)$ if the corresponding triangle T_v is not acute or has a small* circumradius $\mathrm{rad}(v) \le \alpha$. *We get the filtration* $C^* = C^*(0) \supset \cdots \supset C^*(\alpha) \supset \cdots \supset C^*(+\infty) = \{v_0\}$.

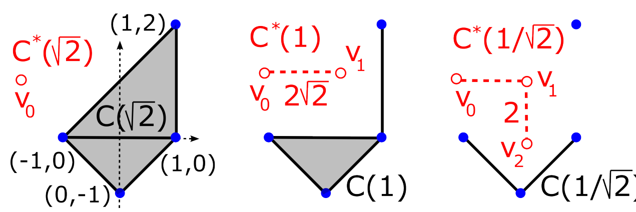

Figure 6. The complexes $C(\alpha)$ have solid edges and gray triangles. The graphs $C^*(\alpha)$ have circled vertices and red dashed edges.

Components of $C^*(\alpha)$ are called *white*, because they represent regions in $\mathbb{R}^2 - C(\alpha)$ (or holes in $\mathbb{R}^2 - C^\alpha$). A cycle $\gamma \subset C(\alpha)$ is called a *contour* if γ bounds a region in $\mathbb{R}^2 - C(\alpha)$, so γ 'encloses' the corresponding white component of $C^*(\alpha)$. Lemma 14 is an analogue of the Symmetry Theorem [4, p. 164] for a function on a closed manifold.

Lemma 14 (Duality)**.** *All contours of the complex $C(\alpha)$ are in a 1-1 correspondence with all connected components of the graph $C^*(\alpha)$ not containing the vertex v_0. When α is decreasing, the contours of $C(\alpha)$ and the white components of $C^*(\alpha)$ have the corresponding critical moments:*

- *a birth of a contour $\leftrightarrow$ a birth of a white component,*
- *a death of a contour $\leftrightarrow$ a death of a white component.* □

7. The algorithm HoCToP for counting holes

We build the union-find structure $\mathrm{Forest}(\alpha)$ on the vertices of the graph $C^*(\alpha)$. All nodes and trees of $\mathrm{Forest}(\alpha)$ will be in a 1-1 correspondence with all vertices and white components of $C^*(\alpha)$. Every node v in $\mathrm{Forest}(\alpha)$ has

- a pointer to a unique parent of the node v in $\mathrm{Forest}(\alpha)$;
- a pointer to the Delaunay triangle dual to the node v;
- the weight (the number of nodes below v in its tree);
- the critical value (birth) $\alpha_v = \sup\{\alpha : v \in C^*(\alpha)\}$.

If a node v is a self-parent, we call v a *root*. We can find $\mathrm{root}(v)$ of any node v by going up along parent links. If α is decreasing, α_v can be considered as the birth time when the vertex v joins $C^*(\alpha)$. The algorithm initializes $\mathrm{Forest}(\alpha)$ as the set of isolated nodes $v_0, \dots, v_k$. If the triangle corresponding to v_k is acute, the birth time of v_k is the circumradius of the triangle, otherwise 0. We will go through all edges of $\mathrm{Del}(C)$ in the decreasing order of their length and will update α_v when v enters the ascending filtration $\{v_0\} = C^*(+\infty) \subset \dots \subset C^*(\alpha) \subset \dots \subset C^*(0) = C^*$.

All triangles of $C(\alpha)$ and the corresponding nodes of $\mathrm{Forest}(\alpha)$ are called *gray*. The remaining triangles and the external region of $\mathrm{Del}(C)$ are called *white*. The external region has birth time $+\infty$ and is called a 'triangle' for simplicity. Initially all triangles with birth time 0 are gray.

The *while loop*. For each edge $e \subset \mathrm{Del}(C)$ arriving in the decreasing order of length, we find two triangles T_u, T_v attached to e and check if they are gray or white. To determine if a triangle T_v represented by a node v is gray, we go up along parent links from v to $\mathrm{root}(v)$. If the birth time of $\mathrm{root}(v)$ is 0, the triangle T_v is still gray, otherwise white.

To distinguish Cases 1 and 4 below, we also check if the triangles T_u, T_v attached to the current edge e are in the same region of $\mathbb{R}^2 - C(\alpha)$. Case 1 means that the nodes $u, v \in \mathrm{Forest}(\alpha)$ belong to the same tree, so $\mathrm{root}(u) = \mathrm{root}(v)$. In all 4 cases the scale α goes down through the half-length $\frac{1}{2}\mathrm{length}(e)$ of the current edge e from $\mathrm{Del}(C)$.

Case 1: e has the same white region on both sides of e. $C(\alpha)$ loses only the open edge e. The white components of $C^*(\alpha)$ are unchanged. Fig. 6 illustrates Case 1 for $\alpha = 1$ when $C(\alpha)$ loses the edge connecting $(1, 0)$ to $(1, 2)$.

Case 2: the edge e is in 1 gray triangle and 1 white triangle. Let $u, v \in C^*(\alpha)$ be the vertices dual to the gray triangle T_u and the white triangle T_v attached to the current edge e in $\mathrm{Del}(C)$. Then the birth times are $\alpha_u = 0$, $\alpha_{\mathrm{root}(v)} > 0$.

Since α is decreasing, the descending filtration $C(\alpha)$ loses the (open) edge e and the gray (open) triangle T_u. So the vertex u becomes connected by an edge with v and joins the white component of $C^*(\alpha)$ containing v. Then we link the isolated node u to the tree containing the older node v in $\mathrm{Forest}(\alpha)$. So $\mathrm{root}(v)$ becomes the parent of u and the weight of $\mathrm{root}(v)$ jumps by 1. Fig. 7 illustrates Case 2 for $\alpha = \frac{\sqrt{17}}{2}$ when $C(\alpha)$ loses the 2 edges of length $\sqrt{17}$.

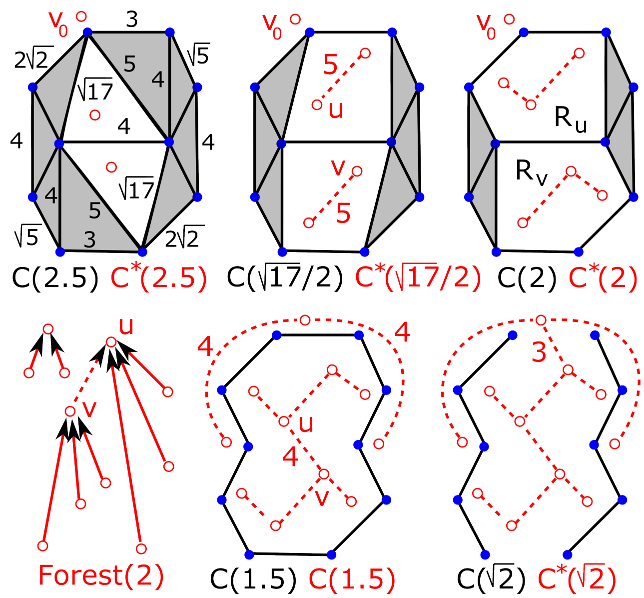

Figure 7. Complexes $C(\alpha)$ and graphs $C^*(\alpha)$ are shown for the cloud C from Fig. 3. Two trees in $\mathrm{Forest}(\alpha)$ merge at $\alpha = 2$.

Case 3: the edge e is in the boundary of 2 gray triangles. Let $u, v \in C^*(\alpha)$ be the vertices dual to the gray triangles T_u, T_v attached to the current edge $e \subset \mathrm{Del}(C)$. Then T_u, T_v are right-angled triangles with the common hypotenuse e. The birth time of both u, v is the half-length of e. Since α is decreasing, $C(\alpha)$ loses the (open) edge e and both (open) triangles T_u, T_v. The contour $\partial(T_u \cup T_v)$ appears in $C(\alpha)$. So we link the nodes u, v in $\mathrm{Forest}(\alpha)$.

Case 4: e has 2 different white regions on both sides. Let $u, v \in C^*(\alpha)$ be the vertices dual to the white triangles T_u, T_v attached to the current edge e in $\mathrm{Del}(C)$. The descending filtration $\{C(\alpha)\}$ loses the (open) edge e. The vertices u, v become connected by an edge, so their white components in $C^*(\alpha)$ merge into a new big component. By Duality Lemma 14, two contours enclosing regions R_u and R_v lose their common edge e and we get one larger contour $\partial(R_u \cup R_v)$ enclosing both regions. Fig. 7 illustrates Case 4 for $\alpha = 2$ when $C(\alpha)$ loses the middle edge of length 4. Then 2 white components (containing 4 vertices each) merge in the graph $C^*(\alpha)$ shown after merger at $\alpha = 1.5$.

To decide which white component dies, we find the roots $\mathrm{root}(u), \mathrm{root}(v) \in \mathrm{Forest}(\alpha)$ of the trees representing R_u, R_v and compare the birth times $\alpha_{\mathrm{root}(u)}, \alpha_{\mathrm{root}(v)}$ when a first node of each tree was born. By the elder rule [4, p. 150], the older white component (say, with u) survives and keeps its larger birth time $\alpha_{\mathrm{root}(u)}$. The younger white component R_v dies and we get $(\text{birth}, \text{death}) = (\frac{1}{2}\text{length}(e), \alpha_{\text{root}(v)})$ for the life of the white component in the ascending filtration $\{C^*(\alpha)\}$ and of the corresponding contour in the descending filtration $\{C(\alpha)\}$.

We swapped the birth and death times, because the persistence is usually defined when the scale α is increasing. However, we need the ascending filtration $\{C^*(\alpha)\}$ to use a union-find structure, so α is decreasing in the algorithm.

Finally, to merge the trees with $\text{root}(u), \text{root}(v)$ in $\text{Forest}(\alpha)$, we compare the weights of the roots and set the root of the (non-strictly) larger tree as the parent for the root of another tree. So the size of any subtree grows by a factor of at least 2 each time when we pass to the parent. We get

Lemma 15. *By the above construction the longest path in any tree of size k from* $\text{Forest}(\alpha)$ *has length* $O(\log k)$. □

8. Proofs of main results and our conclusion

Proof of Theorem 1. Constructing the Delaunay triangulation $\text{Del}(C)$ on a cloud of n points requires $O(n \log n)$ time [2, Chapter 9]. Sorting $O(n)$ edges of $\text{Del}(C)$ needs $O(n \log n)$ time. Then we go through the *while loop* analyzing each of the $O(n)$ edges of $\text{Del}(C)$. For the nodes $u, v \in \text{Forest}(\alpha)$ of triangles attached to each edge e, we find the roots of u, v by going up along $O(\log n)$ parent links by Lemma 15. All other steps in the *while loop* require only $O(1)$ time. Hence the total time is $O(n \log n)$. The sizes of all data structures are proportional to the numbers of edges or triangles in $\text{Del}(C)$, so we use $O(n)$ space. □

The careful analysis of a union-find structure [8] says that $\text{Forest}(\alpha)$ can be built in time $O(nA^{-1}(n,n))$ time, where $A^{-1}(n,n)$ is the extremely slowly growing inverse Ackermann function. Our time $O(n \log n)$ is dominated by the construction of $\text{Del}(C)$ and sorting all $O(n)$ edges.

Proof of Theorem 4. The important critical values of α for the 1-dimensional homology of the filtration $\{X^\alpha\}$ are

- $\alpha = 0$ when the homology $H_1(X^0) = H_1(X)$ is correct;
- $\alpha = \text{minhfs}(X)$ is the 1st value when $H_1(X^\alpha)$ changes;
- $\alpha = \text{maxhfs}(X)$ is the last value when $H_1(X^\alpha)$ changes.

Then $H_1(X^\alpha) \cong H_1(X)$ for $0 \le \alpha < \text{minhfs}(X)$. If a cloud C is an ε-sample of a shape $X \subset \mathbb{R}^2$, then due to Stability Theorem 12 [1] we have $H_1(C^\alpha) \cong H_1(X)$ over (possibly) shorter interval $\varepsilon \le \alpha \le \text{minhfs}(X) - \varepsilon$. Then all changes of $H_1(C^\alpha)$ will stop at $\alpha = \text{maxhfs}(X) + \varepsilon$.

The condition $\text{minhfs}(X) > \frac{1}{2}\text{maxhfs}(X) + 2\varepsilon$ guarantees that the persistence interval $[\varepsilon, \text{minhfs}(X) - \varepsilon]$ is the longest over the full range $[0, \text{maxhfs}(X) + \varepsilon]$, because $\varepsilon < \text{minhfs}(X) - 2\varepsilon > \text{maxhfs}(X) + \varepsilon - (\text{minhfs}(X) - \varepsilon)$. Without using ε, we can find this interval of α, where $C(\alpha)$ has the same number of holes as the unknown shape X. □

Conclusion. Here are the key advantages of our approach:

- a cloud $C \subset \mathbb{R}^2$ of n points is simultaneously analyzed at all scales α without any extra user-defined parameters;
- the algorithm HoCToP counts persistent holes of any topological form in $O(n \log n)$ time, see Theorem 1;
- theoretical guarantees for a correct number of holes are proved for ε-samples of unknown shapes, see Theorem 4;
- the output is stable under perturbations of a cloud C and the only parameter of noise is an unknown upper bound ε.

Fig. 8 shows extracted contours (with our uniform noise) of images at http://www.lems.brown.edu/~dmc.

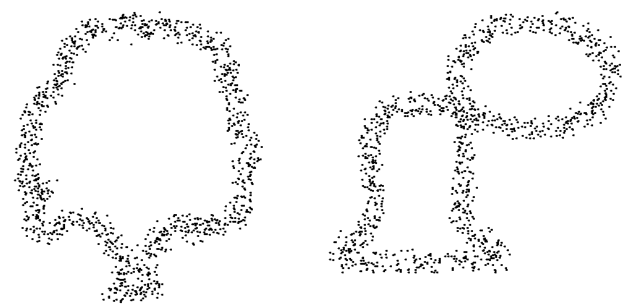

Figure 8. Output of HoCToP for real noisy contours. Left: $P(1 \text{ hole}) \approx 90.5\%$, $P(2 \text{ holes}) \approx 3\%$, $P(4 \text{ holes}) \approx 0.6\%$. Right: $P(2 \text{ holes}) \approx 74.2\%$, $P(1 \text{ hole}) \approx 13\%$, $P(3) \approx 1.3\%$.

More details and experimental results are in the extended version on author's website http://kurlin.org. We thank the anonymous reviewers for theirs helpful suggestions. We are open to collaboration on any related projects.

References

[1] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37:103–130, 2007. 2, 4, 6

[2] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications*. Springer, 2008. 3, 6

[3] H. Edelsbrunner. The union of balls and its dual shape. *Discrete Computational Geometry*, 13:415–440, 1995. 3

[4] H. Edelsbrunner and J. Harer. *Computational topology. An introduction*. AMS, Providence, 2010. 2, 3, 4, 6

[5] X. Ge, I. Safa, M. Belkin, and Y. Wang. Data skeletonization via reeb graphs. Proc. NIPS 2011, pages 837–845. 3

[6] Letscher, D., Fritts, J. Image segmentation using topological persistence. Proceedings of CAIP 2007: Computer Analysis of Images and Patterns, pages 587–595. 3

[7] N. Milosavljevic, D. Morozov, and P. Skraba. Zigzag persistent homology in matrix multiplication time. Proceedings of SoCG 2011, pages 216–225, ACM. 2

[8] R. Tarjan. *Data strustures and network algorithms*. SIAM, Philadelphia, 1983. 6