# An entropy-based persistence barcode ☆

CrossMark

Harish Chintakunta [a], Thanos Gentimis [a], Rocio Gonzalez-Diaz [b,*],
Maria-Jose Jimenez [b], Hamid Krim [a]

[a] *ECE Dept., NCSU, Raleigh, NC USA*
[b] *Applied Math Dept., School of Computer Engineering, U. Seville, Spain*

## ARTICLE INFO

## ABSTRACT

In persistent homology, the persistence barcode encodes pairs of simplices meaning birth and death of homology classes. Persistence barcodes depend on the ordering of the simplices (called a filter) of the given simplicial complex. In this paper, we define the notion of "minimal" barcodes in terms of entropy. Starting from a given filtration of a simplicial complex $K$, an algorithm for computing a "proper" filter (a total ordering of the simplices preserving the partial ordering imposed by the filtration as well as achieving a persistence barcode with small entropy) is detailed, by way of computation, and subsequent modification, of maximum matchings on subgraphs of the Hasse diagram associated to $K$. Examples demonstrating the utility of computing such a proper ordering on the simplices are given.

## 1. Introduction

In recent years persistent homology has successfully been used to characterize topological properties of given sets and structures [10]. Persistence and ZigZag barcodes [3] have helped identify characteristic properties of the underlying space in data mining, network coverage and social networks. The theory of persistence has proven to be robust and invariant under small perturbations of the original state of the data. The algebraic structure of the set of persistence barcodes viewed as a module is well understood [27], and various stability theorems are proven in [4] when the space is equipped with the Wasserstein distance. In particular, it is proven that the persistence barcode is stable under "noise" that is, small changes in the function used to create a filtration to compute persistent homology imply only small changes in the persistence barcode.

It is obvious that a persistence barcode depends heavily on the filter considered for such computation. An "interval" in a persistence barcode representation is a horizontal line segment in a plane whose length is equal to the lifespan of the corresponding homology class. A typical persistence barcode contains short-lived intervals representing topological holes (homology classes) which may not represent real features of the space being analyzed. In such cases one has to set a "threshold" of significance on the length of the intervals of the persistence barcode, and this is normally carried out a posteriori (see, for example, [14] where authors consider a "simplified" barcode). On the other hand, if someone examines in depth typical cases where persistent homology is used, one will be faced with the inherent problem of "noise" in the persistence barcodes. A similar problem was posed in [6,12] where various algorithmic results were presented. Those problems take the form of non-significant topological holes in sensor networks and inefficiencies of the filtration coming from the construction of the Rips (or Čech) complex in data sets.

Motivated by practical applications of persistent homology computation, our starting point is a given simplicial complex and an initial filtration. Although the total number of intervals in a persistence barcode remains invariant (as we will see later), the lengths of its intervals depend on the selected filter. Since non-significant intervals (i.e., intervals with short length) may not imply relevant homological information, we are interested in looking for a filter which preserves the partial ordering imposed by the given filtration, in order to minimize the number of significant intervals and maximize their lengths.

From an information-theoretic viewpoint, and if we interpret the number of significant intervals as the coding length of a complex, our goal is to select the most "parsimonious" representation (also by Occam's razor principle). As is also well known, the coding length is intimately related to the notion of entropy (i.e., a

topological entropy of the barcodes in our case). While one would ideally want to balance the minimization with a penalty term of the number of non-significant intervals [18], further knowledge of statistical distribution of the long intervals is required and postponed to future work.

This paper completes the work proposed in [12] by providing additional insights, examples, results and proofs. A new definition of "minimal barcode" is given using the notion of entropy, and an algorithm for computing persistence barcodes with small entropy starting from a given filtration is provided.

The remainder of this paper is organized as follows. Section 2 covers the relevant background material. In Section 3, we give the definition of minimal barcodes based on the notion of entropy. In Section 4, an algorithm for computing a filter which preserves the partial ordering imposed by the filtration and whose associated persistence barcode has small entropy, is computed on the Hasse diagram of the poset of faces of a given complex. Section 5 is devoted to relations between minimal barcodes and discrete Morse theory. Conclusions and future work are presented in Section 6.

## 2. Preliminaries

Homology theory uses algebraic groups to encode the topological structure of a simplicial complex $K$. In general, we will always consider connected, finite simplicial complexes $K$ (i.e., with $m$ simplices, $m < \infty$) unless stated otherwise.

### 2.1. Homology

The set of $i$-simplices, $\sigma^i$ (superscript denotes dimension), will be denoted by $K^i$. The number of simplices in a set $S$ is denoted by $|S|$. The dimension of the simplicial complex is $n \geq 0$ if $K^n \neq \emptyset$ and $K^m = \emptyset, \forall m > n$. Finite formal sums of simplices of $K^i$ with coefficients in a field (called $i$-chains), define an additive Abelian group structure on $K^i$. In our case, the coefficients for those sums belong to $\mathbb{Z}_2$ thus the group of $i$-chains, $C_i(K; \mathbb{Z}_2)$ is a vector space over $\mathbb{Z}_2$ with basis elements the simplices of dimension $i$.

If a simplex $\sigma$ is a face of another simplex $\sigma'$, we write $\sigma \prec \sigma'$. We say that $\sigma'$ is a *coface* of $\sigma$. A *proper face* of $\sigma \in K^i$, is a face of $\sigma$ of dimension $i - 1$. The *boundary* of $\sigma$, denoted by $\partial(\sigma)$ is the formal sum (with coefficients in $\mathbb{Z}_2$) of the proper faces of $\sigma$. The boundary operator is extended to all chains of $K$ by linearity.

An $i$-chain $a$ is an $i$-cycle if $\partial_i(a) = 0$ i.e., $a \in \text{Ker } \partial_i$; it is an $i$-boundary if there is an $(i+1)$-chain $b$ such that $\partial_{i+1}(b) = a$ i.e., $a \in \text{Im}(\partial_{i+1})$. Two $i$-cycles $a$ and $a'$ are *homologous* if $a + a'$ is an $i$-boundary. Since $\partial_i \partial_{i+1} = 0$, $\text{Im}(\partial_{i+1}) \subseteq \text{Ker } \partial_i$. The quotient of $i$-cycles over $i$-boundaries is the $i$-th *homology group of $K$* i.e., $H_i(K) = \text{Ker}(\partial_i)/\text{Im}(\partial_{i+1})$. The elements of $H_i(K)$ are called *homology classes*.

Since the considered field of coefficients is $\mathbb{Z}_2$, the $i$-th *Betti number* (denoted by $\beta^i$) is the rank of the $i$-th homology group of $K$.

Then, the basic topological structure of $K$ is quantified by the number of independent classes in each homology group. See [25,15].

Given a simplicial complex $K$, a nested sequence of simplicial complexes

$$\emptyset = K_0 \subset K_1 \subset \cdots \subset K_{p-1} \subset K_p = K$$

is called a *filtration* of $K$. An ordering of the simplices of a simplicial complex $K = \{\sigma_1, \ldots, \sigma_m\}$ is called a *filter* if it satisfies the property that $s < t$ whenever $\sigma_s \prec \sigma_t$. Then we can create a filtration by setting:

$$K_t = \{\sigma_1, \ldots, \sigma_t\}, \quad \text{for } 1 \leq t \leq m.$$

### 2.2. Metric filtrations

Many applications of Computational Topology start with a cloud of points embedded in $\mathbb{R}^n$. Using a specific radius $r$ one can then define Alpha-complexes $A(r)$, Čech complexes $C(r)$, and Rips-complexes $R(r)$. Furthermore, one can obtain a filtration

$$\emptyset = K_0 \subset K_1 \subset \cdots \subset K_{p-1} \subset K_p = K$$

by gradually increasing $r$ where $K_i$ is $A(r_i)$, $C(r_i)$ or $A(r_i)$ depending on the complex $K$ we are creating (Alpha, Čech or Rips-complex) and $r_i < r_j$ if $i < j$ (see [7, page 70] and [10]). In particular, all vertices enter at $K_0$, and $K_t$ and $K_{t+1}$ differ by at least one simplex.

### 2.3. Lower-star filtrations

In lecture 11 of their course[1] in computational geometry and topology, Edelsbrunner and Kerber argue that if we choose a reasonable filtration, we can learn more about a complex than just analyzing its Betti numbers. They propose the creation of a filtration given some function on the vertices. Examples of these are the grayscale value of images, or height information in geographical data.

Let $K$ be a simplicial complex with distinct real values specified at their vertices $h : K^0 \to \mathbb{R}$. We can then order the vertices by an increasing function value as $h(v_1) < \cdots < h(v_{m^0})$ where $m^0 = |K^0|$. Each simplex $\sigma$ has a unique maximum vertex $v_{max}$, i.e.,

$$h(v_{max}) = \max\{h(v) : v \in K^0 \text{ and } v \prec \sigma\}$$

The *lower star filtration of $h$* [7, Section VI.3] is the nested sequence of complexes $\emptyset = K_0 \subset K_1 \subset \cdots \subset K_{m^0} = K$ such that:

$$K_t \backslash K_{t-1} = \{\sigma \in K : \text{maximum vertex of } \sigma \text{ is } v_t\}$$

In particular, $K_t$ and $K_{t+1}$ differ by at least one simplex since each simplex has a unique maximum vertex.

### 2.4. Persistent homology

*Persistent homology* [6,27] studies homology classes and their "lifetimes" (persistence) along a nested sequence of objects (simplicial complexes in our case).

Given a filter of $K$, the algorithm for computing persistence barcodes that appears in [6], marks an $i$-simplex $\sigma_t$ as positive (birth) if it belongs to an $i$-cycle in $K_t = \{\sigma_1, \ldots, \sigma_t\}$ (i.e., $\sigma_t$ creates a new homology class at time $t$) and negative (death) otherwise (i.e., $\sigma_t$ destroys a homology class created at some time $s$ for $0 \leq s < t$).

Given a filter $\{\sigma_1, \ldots, \sigma_m\}$, a *persistence barcode* [3] is a graphical representation of pairs of birth and death times, as a collection of horizontal line segments (*intervals*) in a plane. If a simplex $\sigma_s$ creates a homology class at time $s$ (the index in the filter) which is destroyed at time $t$, $0 \leq s < t \leq m$, then the interval $[s, t)$ is added to the corresponding persistence barcode. If a simplex $\sigma_s$, $0 \leq s \leq m$ creates a homology class at time $s$ which survives along the process, then the interval $[s, \infty)$ is added to the persistence barcode. For a fixed $i$, the *$i$-barcode* is the set of intervals of a given persistence barcode corresponding to the pairs of positive $i$-simplices and negative $(i+1)$-simplices of $K$. See [3]. The following lemma holds.

**Lemma 1.** *Independent of the selected filter of a simplicial complex $K$, the number of intervals in an $i$-barcode, $0 \leq i \leq n$, is*

$$\beta^i + \sum_{j=i+1}^{n} (-1)^{i+1-j}(|K^j| - \beta^j),$$

*where $|K^j|$ is the number of j-simplices in K and n is the dimension of K.*

**Proof.** Notice first that, the number of intervals of infinite length in the i-barcode is independent of the filter since it coincides with the i-th Betti number of K, $\beta^i$. Furthermore, recall that each i-simplex $\sigma_t^i$ in the given filter, $0 \leq t \leq m$, is marked as positive or negative. No simplex can remain unmarked after the whole process. Therefore, counting the number of finite intervals in an i-barcode, is equivalent to counting the number of negative $(i+1)-$ simplices. Let $b^i$ (resp. $d^i$) be the number of positive (resp. negative) i-simplices. We then have that $d^0 = d^{n+1} = 0$; and $|K^i| = b^i + d^i$ and $\beta^i = b^i - d^{i+1}$, for $0 \leq i \leq n$ (see [5]).

Hence $d^n = |K^n| - b^n = |K^n| - \beta^n$ and by induction, we have:

$$d^i = |K^i| - b^i = |K^i| - \beta^i - d^{i+1}$$
$$= |K^i| - \beta^i - \sum_{j=i+1}^{n} (-1)^{i+1-j}(|K^j| - \beta^j) = \sum_{j=i}^{n} (-1)^{i-j}(|K^j| - \beta^j). \quad \square$$

In addition, since $d_0 = 0$ we have that $\sum_{j=0}^{n} (-1)^j (|K^j| - \beta^j) = 0$, which yields the known expression for the Euler characteristic [5, page 777]:

$$\chi(K) = \sum_{j=0}^{n} (-1)^j |K^j| = \sum_{j=0}^{n} (-1)^j \beta^j.$$

## 3. Minimal barcodes

Our general aim is to find, under some constraints depending on the nature of the application, filters that minimize the number of long-life homological classes which are associated with significant intervals in the persistence barcodes. A "minimality" in the construction of these persistence barcodes is needed since the ultimate goal is to be able to use them to compare two different topological spaces, for example simplicial complexes X and Y coming from grayscale images [19]. If there is no "uniform" way of constructing a filter, we may end up having similar persistence barcodes describing completely different spaces.

In this paper our starting point is a given filtration of a simplicial complex K, and our aim is to find a "proper" filter preserving the partial ordering imposed by the filtration trying to answer the following question: If we have a shape in 3D, which filtration should we choose for persistent homology computation? The example shown in Fig. 1 of a conical shape with "empty bottom" demonstrates the importance of choosing the appropriate filtration. If we choose an ordering of the vertices using the function F1, and obtain the lower-star filtration, then no holes appear throughout the process. On the other hand, if we choose an ordering of the vertices using F2 and the lower-star filtration, then a hole survives until the last vertex on the top is added, and we thereby unveil much more information about the 3D shape.

We first rigorously formulate a notion of minimality. Consider that a simplicial complex K of dimension n with $m+1$ simplices, is given.

### 3.1. Entropy of a persistence barcode

A general idea in the study of topological persistence is that significant topological attributes must have long life-times, and topological features with short life-times are considered to be "noise".

Following this idea, a notion of k-significant intervals was introduced in [12]: Fix $k > 0$, an interval $[s, t)$ is *k-significant* if
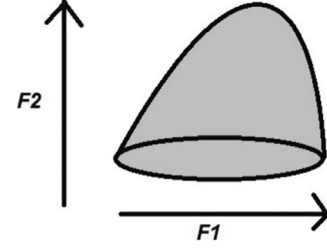


**Fig. 1.** Two orderings of the vertices for the same 3D shape using the functions F1 (increasing weight) and F2 (increasing height).

$k < t - s$. Persistence barcodes associated to different filters of K can be compared by the amount of k-significant intervals they contain.

Another form of comparison between persistence barcodes can be given in terms of entropy. To that end, we need a convention about intervals that persist to "infinity":

**Remark 2.** In persistent homology theory, intervals that extend all the way to the end of the filter are denoted by $[a, \infty)$. In our case we denote them by $[a, m+1)$. This way all intervals have finite length.

**Definition 3.** Given a filter F and the corresponding persistence barcode $B = \{[a_j; b_j] : j \in J\}$, we define the entropy (Shannon entropy) of F as

$$E(F) = - \sum_{j \in J} p_j \cdot \log(p_j) \qquad (1)$$

where $p_j = \ell_j / L$, $\ell_j = b_j - a_j$, and $L = \sum_{j \in J} \ell_j$.

Observe that Formula (1) can also be written as

$$E(F) = \log(L) - \frac{1}{L} \sum_{j \in J} \ell_j \log(\ell_j).$$

**Remark 4.** Since $p_j \leq 1$, then $\log(p_j) \leq 0$ and the entropy of a filter (or barcode) is always positive.

Intuitively, entropy measures how different bars of the barcodes are in length. A barcode with uniform lengths has small entropy. For example, consider the barcodes B1, B2 and B3 shown in Fig. 2 and the filters F1, F2 and F3 associated to them. We have that $E(F1) = 2.198816406$, $E(F2) = 1.474560793$ and $E(F3) = 1.45962243$.

The upper bound of the entropy of a probability mass function **p** is given by $\log |\mathbf{p}|$, where $|\mathbf{p}|$ is the cardinality of the support set of **p**. The upper bound is achieved for a uniform distribution. In the definition of entropy of a barcode, we use the normalized sequence of lengths of the bars in place of the probability mass function. In order to meaningfully compare different barcodes using the entropy function, we need the support sets of the corresponding probability mass functions to have the same cardinality (number of intervals in the barcode). Lemma 1 establishes this fact.

### 3.2. Minimal barcodes and filtrations

As stated in [7, page 149], the measurements of the scale or resolution of topological features using persistent homology make sense only if the function on the given simplicial complex does. This is the reason why we begin with a scenario in which a partial
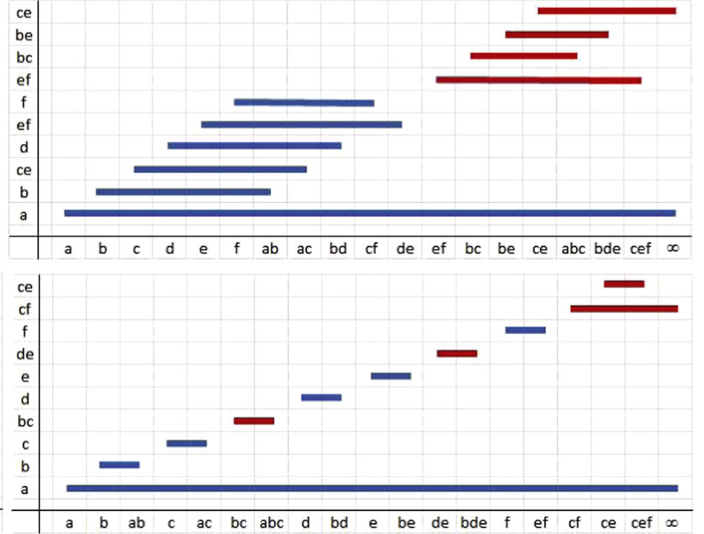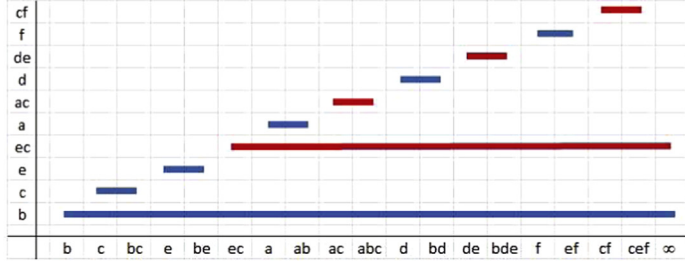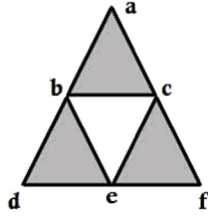
**Fig. 2.** An example of three different persistence barcodes obtained from the same simplicial complex. Intervals corresponding to 0-barcodes are in blue. Intervals corresponding to 1-barcodes are in red. Top-right: barcode *B1*; down-left: barcode *B2*; down-right: barcode *B3*. The filters can be read on the *x*-axis of each of the diagrams. For *B1*, it is $F1 = \{a, b, c, d, e, f, ab, ac, bd, cf, de, ef, bc, be, ce, abc, bde, cef\}$; For *B2*, $F2 = \{b, c, bc, e, be, ec, a, ab, ac, abc, d, bd, de, bde, f, ef, cf, cef\}$; For *B3*, $F3 = \{a, b, ab, c, ac, bc, abc, d, bd, e, bd, de, bde, f, ef, cf, ce, cef\}$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

ordering on the simplices of *K* is given (derived, for example, from a metric or a lower-star filtration). We then define

$$\mathcal{F}_K = \{\text{all filters on } K \text{ respecting the pre} - \text{imposed partial ordering}\}$$

Since in general there are many ways to construct a filter out of a collection of simplices satisfying the pre-imposed partial ordering, an important question arises as to which of these will result in a "minimal" barcode, and in what sense this is "minimal"?

A notion of minimality using *k*-significant intervals is the following:

**Definition 5** (*Rocio Gonzalez-Diaz et al. [12]*). A persistence barcode associated with a filter $F \in \mathcal{F}_K$ is *k*-minimal if the persistence barcode associated with any other filter in $\mathcal{F}_K$ contains greater or equal number of *k*-significant intervals.

Though not unique, such a filter always exists since there are finitely many filters with a fixed set of original conditions on a finite simplicial complex. See Fig. 2 for examples of 1-minimal barcodes. The drawback of this definition is that it depends on determining the right integer *k*. So, another question arises, which integer *k* should we choose? As before, it will depend on the nature of the application. However, this drawback may be entirely avoided by adopting an information theoretic definition as follows.

Define an equivalence relation on $\mathcal{F}_K$:

$$F_1 \sim F_2 \Leftrightarrow E(F_1) = E(F_2) \tag{2}$$

or in other words two filters with the same entropy will be considered equivalent. Consider the set of equivalence classes $\tilde{\mathcal{F}}_K$. The entropy function induces a partial ordering on this set as follows:

$$E(F) \le E(F') \Rightarrow F \le F'.$$

**Definition 6.** A filter $F_{min} \in \tilde{\mathcal{F}}_K$ is *E*-minimal if for any other filter $F \in \tilde{\mathcal{F}}_K$ we have $F_{min} \le F$ according to the entropy ordering.

We note that there is no direct relation between *k*-minimal and *E*-minimal barcodes since *k*-minimality minimizes the number of intervals with length greater than *k* while *E*-minimality looks for a barcode with short intervals being as short as possible and long intervals as long as possible.

## 4. Algorithm for computing persistence barcodes with small entropy

Given a simplicial complex *K*, and a filtration

$$\varnothing = K_0 \subset K_1 \subset \cdots \subset K_p = K,$$

our goal is to find a filter $\{\sigma_1, \ldots, \sigma_m\}$ of *K* with associated persistence barcode having small entropy and satisfying the pre-imposed partial ordering on the simplices of *K* given by the filtration.

In order for our algorithm to work, we need the following result. Suppose that the length of all the intervals in a filter $F(\ell)$ of *K* is fixed except for one interval $[s, t]$ with variable length $\ell$. This variable $\ell$ represents the different possibilities of inserting the simplices $\sigma_s$ and $\sigma_t$ in the filter. Therefore, $\ell \in [\min, \max]$ where $1 \le \min \le \ell \le \max \le m$. Observe that $\ell < m + 1$ since there always exists one interval with maximal length $m + 1$ (the interval representing the connected component of *K*).

**Lemma 7.** *The minimum value for $E(F(\ell))$ can be found in $\ell = \min$ or $\ell = \max$.*

**Proof.** Varying $\ell$ and fixing the rest of the parameters, the entropy $E(F(\ell))$ can be written as

$$E(F(\ell)) = \log(L + \ell) - \frac{c + \ell \log(\ell)}{L + \ell},$$

where *L* and *c* are positive constants. The solution for the equation $E'(F(\ell)) = 0$ is $\ell = e^{c/L}$. Furthermore, $E''(F(e^{c/L})) < 0$. Therefore $\ell = e^{c/L}$ is a maximum. Consequently, the minimum values for $E(F(\ell))$ are obtained in $\ell = \min$ or $\ell = \max$.  □

Since $\lim_{\ell \to +\infty} E(F(\ell)) = 0$, there always exists a positive value $x > e^{c/L}$ such that $E(F(x)) < E(F(\min))$. Observe that if $\max < x$, this value is not valid as a length of the interval.

### 4.1. Hasse diagrams

The *Hasse diagram H* of *K* is the graph whose node set is the set of simplices and whose arcs are the pairs $(\sigma, \sigma')$ if $\sigma$ is a proper face of $\sigma'$. We draw the Hasse diagram in the plane in such a way that,

if $\sigma \prec \sigma'$ then $\sigma$ is in a higher level than that of $\sigma'$. We thus get a multipartite graph where each level corresponds to the dimension of the simplices (see [9]).

Given a filtration of a simplicial complex $K$, consider the subgraphs $\{H_i\}_{1 \le i \le p}$ of the *Hasse diagram* $H$ generated by the simplices $\sigma \in K_i \backslash K_{i-1}$.

We thus could construct a function $T : K \rightarrow [0, m]$ satisfying the following properties:

(1) For $\sigma, \sigma'$ in $K$ such that $\sigma \prec \sigma'$, we have $T(\sigma) < T(\sigma')$.
(2) For any $i$, the entry times of simplices in $H_i$ will surjectively lie in a single closed interval in $\mathbb{N}$.
(3) Given two simplices $\sigma, \sigma' \in K$ such that $\sigma \in K_j$ and $\sigma' \in K_i \backslash K_j$ for some $j < i$, we have $T(\sigma) < T(\sigma')$.

For each $\sigma \in K$, $T(\sigma)$ denotes the *entry time* of $\sigma$.

An order on the simplices of $K$, $\{\sigma_1, \ldots, \sigma_m\}$ can be given by $T(\sigma_t) = t$. The first property on $T$ ensures that $\{\sigma_1, \ldots, \sigma_m\}$ is a valid filter of $K$, and the second property resembles that of consecutive labeling. The third property states that the order prescribed by the given filtration is preserved in the filter.

### 4.2. Algorithm

Now, starting from a given filtration of a simplicial complex $K$, an algorithm for computing a "proper" filter (preserving the partial ordering imposed by the filtration as well as achieving a persistence barcode with small entropy) is detailed, using an adaptation of a maximum matching on the subgraphs $H_i$ associated to $K$.

Fix $i$, let $e_i$ be the maximum level of the subgraph $H_i$ and let $h_i$ be the number of nodes of $H_i$. Given a matching in $H_i$ between levels $j$ and $j+1$, let the *completion number* of a matched $(j+1)$-simplex $\sigma$, denoted by $c(\sigma)$, be the number of its proper faces each of which is matched to a $(j+1)$-simplex. If $\sigma$ is matched to a $j$-simplex, then $c(\sigma)$ is at least 1.

**Algorithm 8.** Assigning a time value to the simplices on each subgraph $H_i$.

```
1:   for each subgraph H_i, 1 ≤ i ≤ p do
2:       set j = 0, t = 1
3:       while j < e_i do
4:           if there are no j-simplices in H_i, then
5:               increment j by one
6:           else
7:               compute a maximum matching between levels j and
                 j+1 in H_i.
8:               for each unmatched j-simplex σ ∈ H_i do
9:                   remove σ from the subgraph H_i
10:                  assign t(σ) = t
11:                  increment t by one.
12:              end for
13:
14:              while there are matched j-simplices in H_i do
15:                  for each matched (j+1)-simplices σ in H_i do
16:                      compute c(σ)
17:                  end for
18:                  if there is no matched (j+1)-simplex σ with
                     c(σ) = 1 then
19:                      find a (j+1)-simplex μ with c(μ) maximum
20:                      for the j-simplex σ matched with μ, assign
                         t(σ) = t
21:                      unmatch μ
22:                      increment t by one
```
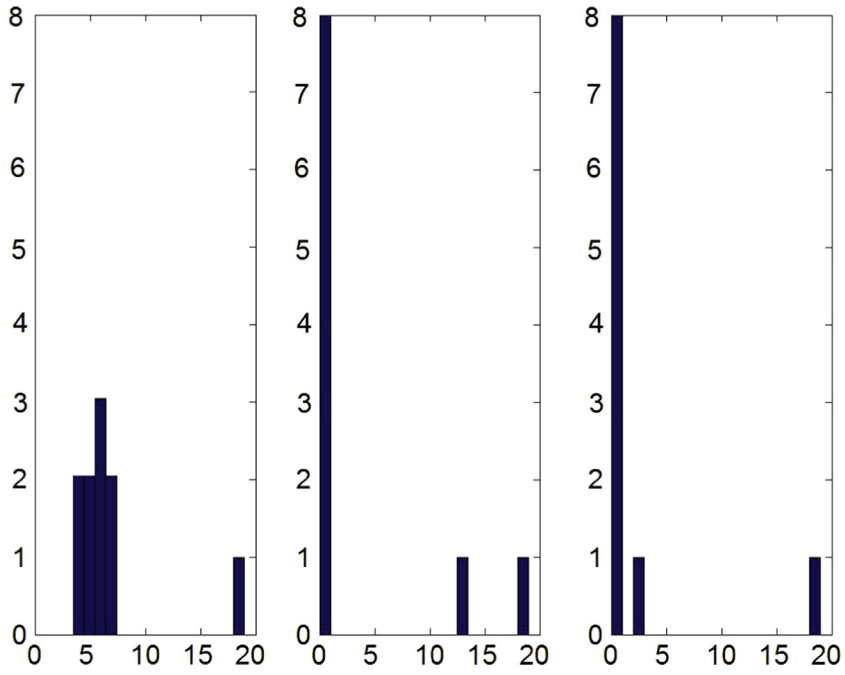
**Fig. 3.** (a) (resp. (b), (c)) The histogram of the lengths of the barcode *B1* (resp. *B2*, *B3*) shown in Fig. 2.
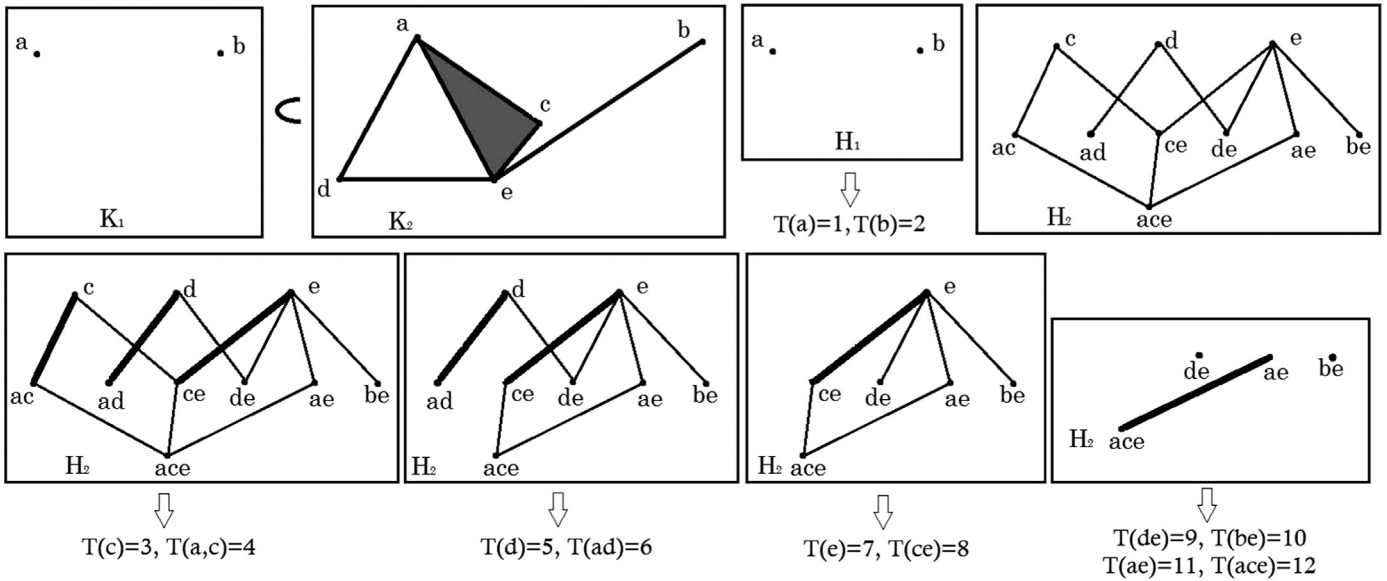


**Fig. 4.** Example of execution of the algorithm to construct a filter, starting from a filtration $\varnothing = K_0 \subset K_1 \subset K_2$.
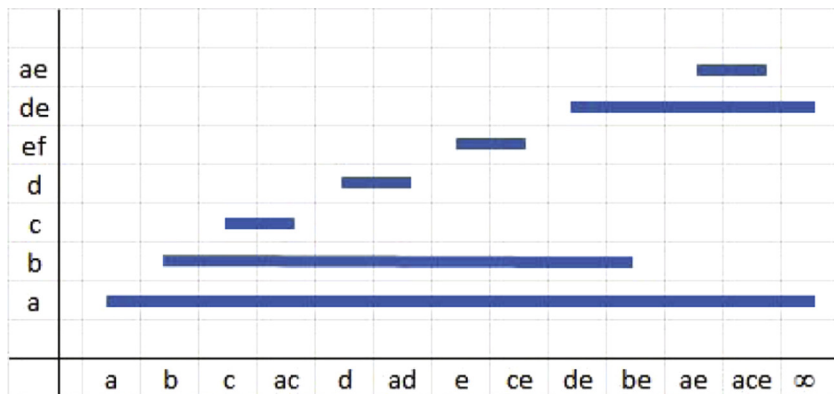


**Fig. 5.** Persistence barcode associated to the filter provided in Fig. 4.

correspond to a matching between two simplices of $K_i \backslash K_{i-1}$, for some $i$ and every interval of length greater than 1 will be associated to non-matched simplices.

By construction, the output of the algorithm is an ordering on the simplices that is in fact a filter of $K$ since properties (1), (2) and (3) enumerated in the introduction of this section are satisfied.

By convention, we have the vertices of $K$ in the top row of the Hasse diagram $H$. Fix $i$, let $H_i'$ be the directed graph obtained from $H_i$ by pointing the matched edges down, and all other edges up. Update $H_i'$ at the same time as $H'$ when running the algorithm. For a fixed $j$ in the algorithm, let $(\sigma, \sigma')$ be a matched pair of a $j$-simplex and a $(j+1)$-simplex. The pair $(\sigma, \sigma')$ can be "safely" added to the filter (preserving the simplicial structure) only when the completion number of $\sigma'$ is equal to 1 due to the following results:

**Lemma 9.** *If there exists a directed cycle in $H_i'$, then the ordering induced on the simplices by this matching is not a valid filter.*

**Proof.** Suppose there exists a directed cycle between levels $j$ and $j+1$. Let $\sigma_1, \sigma_2, ..., \sigma_n$ be the nodes of the directed cycle. Suppose that $\sigma_1$ is a $j$-simplex and $\sigma_s$ is matched to $\sigma_{s+1}$ for $s$ being odd. Observe that $n$ must be even, $\sigma_{n-1}$ matched to $\sigma_n$ and $\sigma_1 \prec \sigma_n$. Then, $\sigma_{n-1}$ must precede $\sigma_n$ in the filter. Since $\sigma_1$ is matched to $\sigma_2$, then $\sigma_1$ must precede $\sigma_2$ in the filter. If we iterate this procedure, we obtain that $\sigma_n$ must precede $\sigma_1$ and $\sigma_2$ but $\sigma_1 \prec \sigma_n$, so $\sigma_n$ cannot precede to $\sigma_1$.  □

The following theorem derives the condition (used in the algorithm) to identify the existence of such a directed cycle.

**Theorem 10.** *For fixed $i$ and $j$, all the matched $(j+1)$-simplices have completion number $>1$ in $H_i$ if and only if there exists a directed cycle between levels $j$ and $j+1$ in $H_i'$.*

**Proof.** First, take any $(j+1)$-simplex $\sigma_2$ matched to a $j$-simplex $\sigma_1$ such that $c(\sigma_2) > 1$. Suppose that $\sigma_2$ is matched with a $j$-simplex $\sigma_1$. Since $c(\sigma_2) > 1$, there exists a $j$-simplex $\sigma_3$ connected to $\sigma_2$ and matched to a $(j+1)$-simplex $\sigma_4$. If we iterate this procedure, for some odd number $s$, a $j$-simplex $\sigma_s$ is matched to a $(j+1)$-simplex $\sigma_{s+1}$ which is connected to the $j$-simplex $\sigma_1$, creating the directed cycle. Finally, if there exists a directed cycle between levels $j$ and $j+1$ in $H_i'$ then the completion number of the $(j+1)$-simplices of the cycle is at least 2.  □

**Theorem 11.** *Given a filtration of a simplicial complex $K$, the entry time function $T$ provides a filter of $K$ respecting the partial ordering on the simplices of $K$ given by the filtration.*

**Proof.** First, let $\sigma$ be a $j$-simplex and $\sigma < \sigma'$ then $T(\sigma) < T(\sigma')$ (i.e., the ordering of the simplices of $K$ is a filter). The following cases can occur:

1. If $\sigma$ is matched to $\sigma'$, then $T(\sigma') = T(\sigma) + 1$ by construction.
2. If $\sigma$ is not matched to $\sigma'$ and $\sigma, \sigma' \in K_i \backslash K_{i-1}$, then $\sigma, \sigma'$ are nodes of $H_i$, so:
   2.1. If $c(\sigma') = 0$, then $t(\mu) < t(\sigma')$ for any $j$-simplex $\mu$ of $H_i$, so in particular $t(\sigma) < t(\sigma')$ and therefore $T(\sigma) < T(\sigma')$.
   2.2. If $c(\sigma') = 1$, then $\sigma'$ is not matched to any other $(j+1)$-simplex, so $t(\sigma) < t(\mu)$ for any $(j+1)$-simplex $\mu \in H_i$. In particular, $t(\sigma) < t(\sigma')$ and therefore $T(\sigma) < T(\sigma')$.
3. If $\sigma \in K_i$ and $\sigma' \in K_j \backslash K_i$, for some $i < j$, then
$$T(\sigma') = t(\sigma') + \sum_{s=1}^{j-1} h_s \geq \sum_{s=1}^{i} h_s \geq T(\sigma).$$

Second, the ordering of the simplices in the filter satisfies the partial ordering imposed by the given filtration by the same arguments as in point 3 above.  □

Complexity of Algorithm 8 depends on the complexity of the algorithm selected to compute a maximum matching between consecutive levels in each Hasse diagram (see line 7 of Algorithm 8). The Hasse diagram is a bipartite graph, and in this case, the worst-case complexity of the classical algorithm for finding a maximum matching (Hopcroft–Karp algorithm) is $O(\sqrt{|V|}|E|)$, where $|V|$ is the number of vertices between two consecutive levels, and $E$ is the number of edges. Apart from this, the rest of the algorithm runs in $O(|V|^2)$ due to the two *for* instructions inside the *while* loop.

### 4.4. Persistence barcodes with small entropy from lower-star filtrations

In this subsection we will show using examples that given a lower-star filtration, the persistence barcode obtained using our algorithm has small entropy, and improves the barcode in the sense of "making short intervals very short and long intervals long".

Given a simplicial complex $K$, start with distinct real values specified at the vertices $h : K^0 \to \mathbb{R}$, which can be thought of as a set of "heights". Construct the lower-star filtration, $\varnothing = K_0 \subset K_1 \subset \cdots \subset K_{m^0}$, of $K$ (where $m^0 = |K^0|$) as explained in Section 2.3. Order the vertices by increasing function value as $h(v_1) < \cdots < h(v_{m^0})$ where $m^0 = |K^0|$. Construct the lower-star filtration,

$$\varnothing = K_0 \subset K_1 \subset \cdots \subset K_{m^0},$$

of $K$ (where $m^0 = |K^0|$) as explained in Section 2.3.

For each simplex $\sigma \in K$, assign a height to $\sigma$ which is equal to the maximum height on its faces. This way, a set of subgraphs $\{H_i\}_{1 \leq i \leq m^0}$ of $H$ associated with the lower-star filtration of $K$, can be constructed, where the nodes of $H_i$ are the simplices of $K$ with height equal to $h(v_i)$ for each $i$. After applying the procedure explained in Section 4.2, an entry time function $T : K \to \mathbb{N}$ is obtained satisfying the pre-imposed partial ordering on the simplices of $K$ given by the lower-star filtration.

Consider the example shown in Fig. 6. The barcodes for filters obtained using random ordering on the simplices in a lower-star filtration and using the algorithm presented here are shown in Figs. 7 and 8 respectively. The intervals corresponding to the 0-homology are shown in blue and the intervals corresponding to 1-homology are shown in red. Even though the homology of the complex is trivial, the ordering on the vertices causes at least one long interval for the first homology as seen in both cases. However, if the ordering on the other simplices is not chosen properly, the resulting barcode may have several other long bars as seen in Fig. 7.

As another example, consider the simplicial complex shown in Fig. 9. Using the embedding in the plane, we consider the $y$-coordinate as the function on the vertices. Again, preserving the ordering on the vertices induced by this function, we obtain two filters, (1) using the algorithm presented here, and (2) using random ordering on the simplices. The resulting barcodes for the
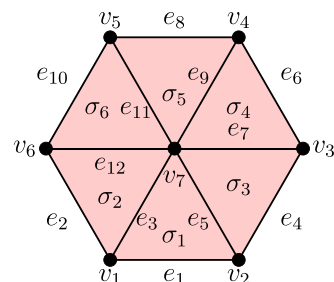


**Fig. 6.** The function value on a vertex $v_i$ is equal to $i$. The indices of other simplices are in the lexicographic order.
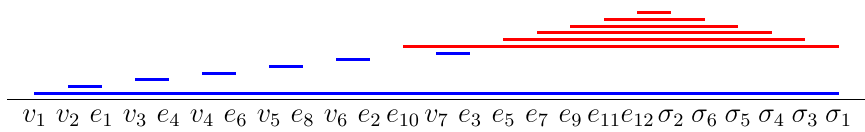
$v_1 \ v_2 \ e_1 \ v_3 \ e_4 \ v_4 \ e_6 \ v_5 \ e_8 \ v_6 \ e_2 \ e_{10} \ v_7 \ e_3 \ e_5 \ e_7 \ e_9 \ e_{11} e_{12} \ \sigma_2 \ \sigma_6 \ \sigma_5 \ \sigma_4 \ \sigma_3 \ \sigma_1$

**Fig. 7.** Barcode obtained using lower-star filtration with arbitrary ordering on the simplices at any stage.



$v_1 \ v_2 \ e_1 \ v_3 \ e_4 \ v_4 \ e_6 \ v_5 \ e_8 \ v_6 \ e_2 \ e_{10} \ v_7 \ e_3 \ e_5 \ \sigma_1 \ e_7 \ \sigma_3 \ e_9 \ \sigma_4 e_{11} \ \sigma_5 e_{12} \ \sigma_6 \ \sigma_2$
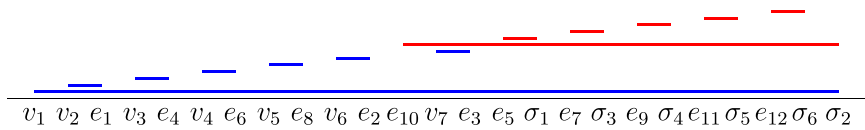
**Fig. 8.** Barcode obtained using the algorithm presented in this paper.
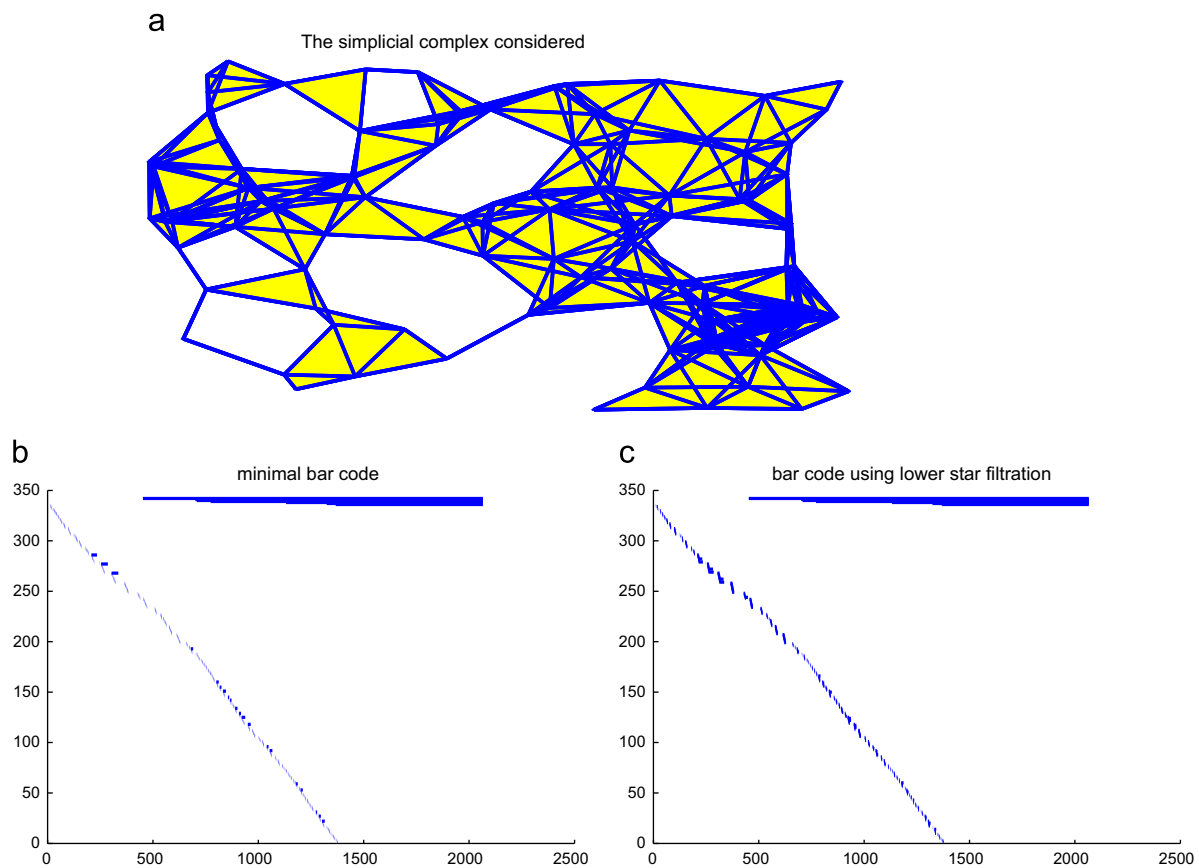


**Fig. 9.** (a) The simplicial complex being considered, (b) barcode obtained from our algorithm, and (c) barcode obtained using random ordering. The *y*-axis represents the index of the simplices in the given filter.

first homology are shown in Fig. 9. The following observations are in order.

(1) Both barcodes look similar, since we use same function on the vertices.
(2) The lengths of the intervals in the barcode obtained using our algorithm are smaller than that obtained using random ordering on simplices with equal height. This is because we are optimizing the ordering.

For this example, the entropy for the barcode obtained using our algorithm is 3.42 and that obtained using random ordering on simplices with equal height is 4.47. The two barcodes in the figure look very similar to each other due to scale. The difference is much clearer in Fig. 10(a), which compares the histograms of the lengths

of the intervals for the two barcodes. We can make the following observations from Fig. 10(a):

(1) For small values of *k* and $k > 1$, the number of *k*-significant intervals is larger for the barcode obtained using random ordering on simplices with equal height.
(2) For large values of *k*, the number *k*-significant intervals is larger for the barcode obtained using our algorithm.
(3) The above two points provide evidence for the fact that the barcode obtained using our algorithm minimizes the length of small intervals and maximizes the length of large intervals.

Fig. 10(b) shows simulation results which further corroborates the effectiveness of the algorithm to produce barcodes with small entropies. The simplicial complexes considered in the experiments
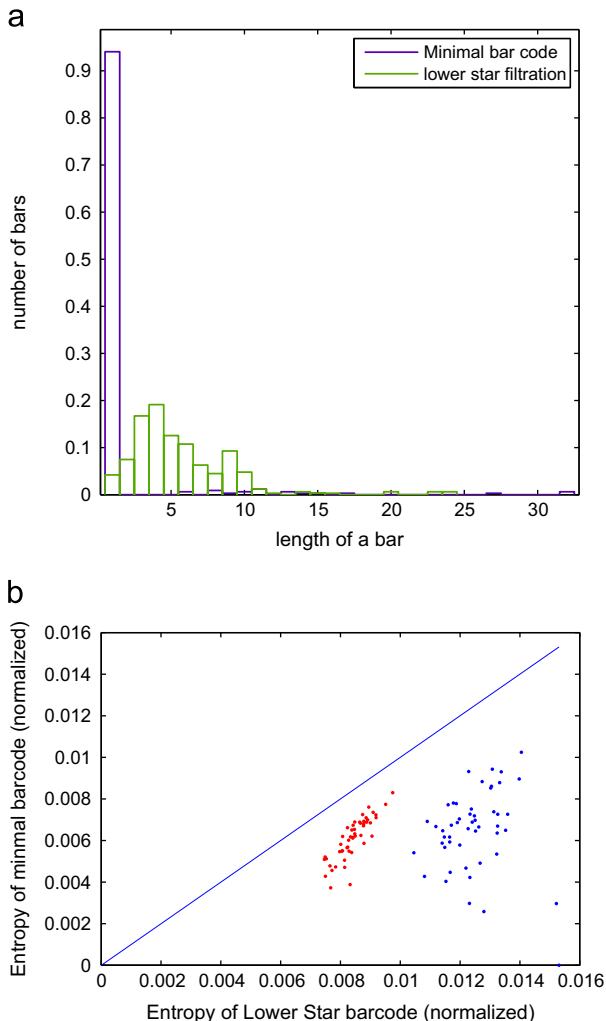
a



b



**Fig. 10.** (a) Comparison of the histograms of the lengths of the intervals in barcodes obtained from our algorithm to that obtained using random ordering on simplices with equal height. The $y$-axis ("number of bars") represents the probability of the length of the bars in the corresponding barcode and (b) figures show a comparison of entropies of the barcodes obtained from the lower star filtration ($E_{ls}$) with that obtained using the algorithm presented here ($E_m$). All the points lie below the $E_m = E_{ls}$ line demonstrating the algorithm produces barcodes with small entropies.

are random geometric complexes obtained from points on a plain. We generated 50 samples for each case of 100 points and 150 points, and created Rips complexes so that the underlying 1-skeleton has an average degree of 10 (these choices are arbitrary). The entropy for each sample complex was then computed, using the height function as a pre-ordering on the vertices, for the lower-star filtration and the filter obtained using the algorithm presented here. Since the upper bound for the entropy varies with the number of intervals $M$ as $\log(M)$, we normalize the entropies by $\log(M)$. In Fig. 10(b), the points in black are the entropies obtained for 100 points and those in red are entropies obtained for 150 points. The $x$-axis corresponds to entropy of lower star filtration ($E_{ls}$) and the $y$-axis to the entropy of the barcodes using our algorithm ($E_m$), which as seen, is lower.

## 5. In relation to discrete morse theory

It is clear that our work in this paper resembles that of the computation of optimal discrete Morse functions. We clarify here the differences between both problems.

Discrete Morse theory was developed by Forman in [8,9] as a combinatorial approach, on cell complexes, to Morse theory. Once a Morse function has been defined on a cell complex, then information about its topology can be deduced from its critical elements.

A function, $f : K \to \mathbb{R}$ is a *discrete Morse function* if for every $\sigma \in K$, $f$ takes a value less than or equal to $f(\sigma)$ on at most one coface of $\sigma$, or takes a value greater than or equal to $f(\sigma)$ on at most one face of $\sigma$. A cell $\sigma$ is *critical* if all of its cofaces take strictly greater values and all its faces are strictly lower. A cell which is not critical is called *regular*. A discrete vector field $V$ is a collection of pairs $(\sigma^i \prec \sigma^{i+1})$ of cells in $K$ such that each cell is in at most one pair of $V$. A discrete Morse function defines a discrete vector field by pairing $\sigma^i \prec \sigma^{i+1}$ whenever $f(\sigma^i) \geq f(\sigma^{i+1})$. The critical cells are precisely those which do not appear in any pair. Discrete vector fields which arise from Morse functions are called *gradient vector fields*. A *V-path* is a sequence of cells (of dimensions $j$ and $j+1$):

$$\sigma_0^j, \sigma_0^{j+1}, \sigma_1^j, \sigma_1^{j+1}, \ldots \sigma_{r-1}^j, \sigma_{r-1}^{j+1}, \sigma_r^j$$

where $(\sigma_t^j, \sigma_t^{j+1}) \in V$, $\sigma_{t+1}^j \prec \sigma_t^{j+1}$ and $\sigma_t^j \neq \sigma_{t+1}^j$. A *V*-path is a nontrivial closed *V*-path if $\sigma_r^j = \sigma_0^j$ for $r \geq 1$.

Forman showed that a discrete vector field is the gradient vector field of a discrete Morse function if and only if there are no nontrivial closed *V*-paths, so a discrete Morse function can be computed from a partial matching in the Hasse diagram with no directed cycles.

From this point of view, consider the final set of matchings computed in all the subgraphs $H_i$ through Algorithm 8 (removing the corresponding matching, as it is stated in the algorithm, when there is no matched $(j+1)$-simplex $\sigma$ with $c(\sigma) = 1$). Then this matching provides a way to construct a discrete Morse function for a cell complex of arbitrary finite dimension. The nonexistence of directed cycles in the subgraphs $H_i'$ implies that there cannot be any directed cycle in the Hasse diagram of the whole complex. This is due to the fact that all the faces of a given cell $\sigma \in H_i$ for some $i$ belong to $\bigcup_{\ell=1\ldots i} H_\ell$ (because every $K_i$ in the filtration is a cell complex itself). So a cycle in $H$ cannot be decomposed in paths in different subgraphs $H_i'$.

There have been several works in the literature addressing the problem of obtaining optimal discrete Morse functions, with the minimum possible number of critical cells in each dimension (see for example [20,22,16,21]) and perfect Morse functions, for which the number of critical $i$-cells coincides with the $i$-th Betti number of the complex [1].

An optimal Morse function can be interpreted in terms of a maximum matching in the corresponding Hasse diagram with no directed cycles. We cannot guarantee optimality in the global Morse matching provided by our algorithm since (1) the partial order imposed by the filtration may constrain the possible matchings, forcing some simplices to be left unmatched (that is the case of simplices $e$ and $be$ in example of Fig. 4); (2) when a directed cycle is detected in our algorithm, we eliminate a matching, but perhaps an alternative maximum matching with no cycles could be found. Since maximal matching at each $H_i$ cannot generally guarantee optimality in entropy, this heuristic is acceptable. This is particularly reasonable, in light of the fact that determining an optimal Morse function is computationally expensive [20,16,2].

In general, optimal discrete Morse functions cannot yield minimal barcodes (in fact not even valid filters) as resulting matchings between different subgraphs $H_i$, could violate the pre-imposed ordering of the vertices. Conversely, a set of matchings provided by a minimal barcode respecting a given filtration may not lead to an optimal discrete Morse function, since the matching may not even be maximal.

Numerous papers addressed the computation of discrete Morse functions in different contexts, also trying to respect some filtration on the set of vertices. In [17], an algorithm is proposed to

compute an optimal discrete Morse function on a simplicial complex $K$ to mirror the large-scale behavior of an injective map on the vertices of $K$, $h : K^0 \to \mathbb{R}$. The output of this algorithm also yields a valid filter of $K$ which preserves the partial ordering given by the lower-star filtration. This algorithm is valid for simplicial complexes, and may likely be extended to cubical complexes. Our algorithm computes filters which preserve the partial ordering given by an arbitrary filtration (not only lower-star filtration), and may be applied to any finite cell complex (since the computation is carried out on a Hasse diagram representation of the complex). In [26,13], the authors present algorithms related to three dimensional grayscale digital images utilizing the machinery of discrete Morse theory. The filtration is, then, given by the gray levels of the image. It is also worth to mention paper [24], which addresses the problem of reducing the subcomplexes in a given filtration, extending combinatorial Morse theory from complexes to filtrations. This way, an alternate filtration with smaller complexes is constructed from the original with the same persistent homology. Therefore, no filter of $K$ is provided. Finally, Lewiner studies the relation between Forman's and piecewise linear approaches to discrete Morse theory in [23], constructing a discrete vector field from a scalar function sampled on the vertices of a triangulated surface. The goal is to relate the critical sets of both approaches.

## 6. Conclusions and comments

We proposed a new filtering algorithm on a simplicial complex by optimizing the partial ordering preservation and achieving a desirable information measure defined on the barcode. Exploiting the Hasse Diagram of the poset of faces of a complex, such a filter indeed achieves a small entropy measure of the resulting barcode. Future work includes the minimization of such a measure which is to be balanced by the error incurred relative to the ideal representation as would be given by an oracle.

## Conflict of Interest

None declared.

## Acknowledgments

## References

[1] Rafael Ayala, Desamparados Fernández-Ternero, A. José, Vilches: perfect discrete Morse functions on 2-complexes, Pattern Recognit. Lett. 33 (2012) 1495–1500.

[2] Benjamin A. Burton, Thomas Lewiner, João Paixão, Jonathan Spreer, Parameterized complexity of discrete morse theory, in: Proceedings of the 29th Annual Symposium on Computational Geometry (SoCG '13). ACM, New York, NY, USA, 2013, pp. 127–136.

[3] Gunnar E. Carlsson, Vin de Silva, Zigzag persistence, Found. Comput. Math. 10 (4) (2010) 367–405.

[4] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, Stability of persistence diagrams, Discret. Comput. Geom. 37 (1) (2007) 103–120.

[5] Cecil J.A. Delfinado, Herbert Edelsbrunner, An incremental algorithm for Betti numbers of simplicial complexes on the 3-spheres, Comput. Aided Geom. Des. 12 (7) (1995) 771–784.

[6] Herbert Edelsbrunner, David Letscher, Afra Zomorodian, Topological persistence and simplification, in: Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS'00), IEEE Computer Society, 2000, pp. 454–463.

[7] Herbert Edelsbrunner, John Harer, Computational topology—an introduction, Am. Math. Soc. 2010, pp. I–XII, 1–241.

[8] Robin Forman, Morse theory for cell complexes, Adv. Math. 134 (1998) 90–145.

[9] Robin Forman: a user's guide to discrete Morse theory, in: Proceedings of the 2001 International Conference on Formal Power Series and Algebraic Combinatorics, Advances in Applied Mathematics, 2001, pp. 1–48.

[10] Robert Ghrist, Barcodes: the persistent topology of data, Bull. Am. Math. Soc. 45 (2008) 61–75.

[12] Rocio Gonzalez-Diaz, Maria-Jose Jimenez, Hamid Krim, Towards minimal barcodes, in: Proceedings of the 9th Workshop on Graph-based Representation in Pattern Recognition (GbRPR2013), Lecture Notes in Computer Science, vol. 7877, 2013, pp. 184–193.

[13] David Günther, Jan Reininghaus, Hubert Wagner, Ingrid Hotz, Efficient computation of 3D Morse-Smale complexes and persistent homology using discrete Morse theory, Vis. Comput. 28 (10) (2012) 959–969.

[14] Antonio Gutierrez, Maria-Jose Jimenez, David Monaghan, Noel E. O'Connor, Topological evaluation of volume reconstructions by voxel carving, Comput. Vis. Image Underst. 121 (2014) 27–35.

[15] Allen Hatcher, Algebraic Topology, Cambridge University Press, Cambridge, 2002, pp. I-XII, 1-544.

[16] Michael Joswig, Marc E. Pfetsch, Computing optimal morse Matchings, SIAM J. Discret. Math. 20 (1) (2006) 11–25.

[17] Henry King, Kevin P. Knudson, Neza Mramor, Generating discrete Morse functions from point data, Exp. Math. 14 (4) (2005) 435–444.

[18] Hamid Krim, John H. Cozzens, A data-based enumeration technique for fully correlated signals, IEEE Trans. Signal Process. 42 (7) (1994) 1662–1668.

[19] Javier Lamar-Leon, Edel B. Garcia-Reyes, Rocio Gonzalez-Diaz, Human gait identification using persistent homology, in: Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP2012), Lecture Notes in Computer Science, vol. 7441, 2012, pp. 244–251.

[20] Thomas Lewiner, Helio Lopes, Geovan Tavares, Optimal discrete Morse functions for 2-manifolds, Comput. Geom. 26 (3) (2003) 221–233.

[21] Thomas Lewiner, Helio Lopes, Geovan Tavares, Toward optimality in discrete Morse theory, Exp. Math. 12 (3) (2003) 271–285.

[22] Thomas Lewiner, Helio Lopes, Geovan Tavares, Applications of Forman's discrete Morse theory to topology visualization and mesh compression, IEEE Trans. Vis. Comput. Graph. 10 (5) (2004) 499–508.

[23] Thomas Lewiner, Critical sets in discrete Morse theories: relating Forman and piecewise-linear approaches, Comput. Aided Geom. Des. 30 (6) (2013) 609–621.

[24] Konstantin Mischaikow, Vidit Nanda, Morse theory for filtrations and efficient computation of persistent homology, Discret. Comput. Geom. 50 (2) (2013) 330–353.

[25] James R. Munkres, Elements of Algebraic Topology, Addison-Wesley, California, 1984, pp. I-IX, 1-454.

[26] Vanessa Robins, Peter J. Wood, Andrian P. Sheppard, Theory and algorithms for constructing discrete Morse complexes from grayscale digital images, IEEE TPAMI 33 (8) (2011) 1646–1658.

[27] Afra Zomorodian, Gunnar E. Carlsson, Computing persistent homology, Discret. Comput. Geom. 33 (2) (2005) 249–274.

**Harish Chintakunta** is a postdoctoral researcher in ECE department at North Carolina State University. His research focused mostly on fidelity of deployment and operation in sensor networks. The motivation for the research stems from the observation that problems in this field, such as tracking coverage and correlated failures, can be posed purely in topological terms. Sensor networks also pose unique technical challenges due to limited resources available, and due to the need for distributed algorithms. His work was to discover/develop appropriate mathematical and computational tools, to perform the required tasks efficiently, distributively, and with minimal information.

**Thanos Gentimis** is a postdoctoral researcher in ECE department at North Carolina State University and member of the VISSTA group. His field of expertise is Algebraic Topology and specifically Geometric Group Theory where he has most of his publications. Recently his interests shifted towards Applied Topology and he is currently working on problems like sensor network coverage, data mining, detection of respiratory abnormalities and social networks using algebraic topology tools (Persistent Homology etc).

**Rocio Gonalez-Diaz** is an Associate Professor in the Department of Applied Math (I) at the University of Seville. She is the head of the Computational Image Analysis group since 2010. Principal Researcher of the Spanish project Computational Topology for Combinatorial Image Analysis MTM2012-32706 and team member of the European

Science Foundation Program: Applied and Computational Topology. Her main interest is to investigate tools (methods) for combinatorial image analysis that build on the theory of (persistent-) homology and cohomology.

**Maria-Jose Jimenez** is an Associate Professor in the Department of Applied Math (I) at the University of Seville. She is a team member of the Computational Image Analysis group, the Spanish project Computational Topology for Combinatorial Image Analysis MTM2012-32706 and the European Science Foundation Program: Applied and Computational Topology. Her main interest is the application of discrete and combinatorial topology to provide topological analysis of 3D objects.

**Hamid Krim**. As a Member of Technical Staff at AT&T Bell Labs, he has worked in the areas of telephony and digital communication systems/subsystems. Following an NSF postdoctoral fellowship at Foreign Centers of Excellence, LSS/University of Orsay, Paris, France, he became a Research Scientist at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, performing and supervising research. He is presently on the faculty in the ECE Department, North Carolina State University, Raleigh, leading the Vision, Information and Statistical Signal Theories and Applications group, whose research interests are in statistical signal and image analysis and mathematical modeling with a keen emphasis on applied problems. Dr. Krim is also an Associate Editor of IEEE Transactions on Signal Processing.