

Persistent Homology: An Introduction and a New Text Representation for Natural Language Processing

Xiaojin Zhu

Department of Computer Sciences, University of Wisconsin-Madison
 Madison, Wisconsin, USA 53706
 jerryzhu@cs.wisc.edu

Abstract

Persistent homology is a mathematical tool from topological data analysis. It performs multi-scale analysis on a set of points and identifies clusters, holes, and voids therein. These latter topological structures complement standard feature representations, making persistent homology an attractive feature extractor for artificial intelligence. Research on persistent homology for AI is in its infancy, and is currently hindered by two issues: the lack of an accessible introduction to AI researchers, and the paucity of applications. In response, the first part of this paper presents a tutorial on persistent homology specifically aimed at a broader audience without sacrificing mathematical rigor. The second part contains one of the first applications of persistent homology to natural language processing. Specifically, our Similarity Filtration with Time Skeleton (SIFTS) algorithm identifies holes that can be interpreted as semantic “tie-backs” in a text document, providing a new document structure representation. We illustrate our algorithm on documents ranging from nursery rhymes to novels, and on a corpus with child and adolescent writings.

1 Introduction

Imagine dividing a document into smaller units such as paragraphs. A paragraph can be represented by a point in some space, for example, as the bag-of-words vector in \mathbb{R}^d where d is the vocabulary size. All paragraphs in the document form a point cloud in this space. Now let us “connect the dots” by linking the point for the first paragraph to the second, the second to the third, and so on. What does the curve look like? Certain structures of the curve capture information relevant to Natural Language Processing (NLP). For instance, a good essay may have a conclusion paragraph that “ties back” to the introduction paragraph. Thus the starting point and the ending point of the curve may be close in the space. If we further connect all points within some small ϵ diameter, the curve may become a loop with a hole in the middle. In contrast, an essay without any tying back may not contain holes, no matter how large ϵ is.

There has been geometric methods for visualizing documents and information flow, e.g. based on differential geometry [Lebanon *et al.*, 2007; Lebanon, 2006; Gous, 1999; Hall and Hofmann, 2000]. In contrast, we introduce an *algebraic* method based on persistent homology. As a branch of topological data analysis, persistent homology has the advantage of capturing novel invariant structural features of documents. Intuitively, persistent homology can identify clusters (0-th order holes), holes (1st order, as in our loopy curve), voids (2nd order holes, the inside of a balloon), and so on in a point cloud. Considering the importance of clustering today, the value of these higher order structures is tantalizing. Indeed, in the last few years persistent homology has found applications in data analysis, including neuroscience [Singh *et al.*, 2008], bioinformatics [Kasson *et al.*, 2007], sensor networks [de Silva and Ghrist, 2007a; de Silva and Ghrist, 2007b], medical imaging [Chung *et al.*, 2009], shape analysis [Gamble and Heo, 2010], and computer vision [Freedman and Chen, 2011].

Unfortunately, existing homology literature requires advanced mathematical background not easily accessible to a broader audience. Our first contribution is an accessible yet rigorous tutorial that contains many unpublished materials. Although a tutorial is unconventional in a technical paper, we feel that there is value to the AI community as it paves the way to further interdisciplinary research. Our second contribution is a novel text representation using persistent homology. It formalizes the curve-and-loop intuition based on Vietoris-Rips filtration over semantic similarity. We hope this paper inspires future innovations on topology and AI.

2 Persistent Homology

We aim for mathematical rigor and intuition, but have to sacrifice completeness. Readers can follow up with [Singh *et al.*, 2008; Giblin, 2010; Freedman and Chen, 2011; Zomorodian, 2001; Rote and Vegter, 2006; Edelsbrunner and Harer, 2010; Hatcher, 2001; Carlsson, 2009; Edelsbrunner and Harer, 2007; Balakrishnan *et al.*, 2012; 2013] for detailed treatment.

Persistent homology finds “holes” by identifying equivalent cycles: Consider the following space in yellow with a small white hole. Imagine the blue cycle as a rubber band. It can be stretched and bent within the space into the green cycle, but not the red one without tearing itself.



There are two equivalent classes of rubber bands: some surround the hole and others do not. Conversely, two equivalent classes indicate one hole. To formalize this idea, we need to introduce some algebraic concepts.

2.1 Group Theory

Definition 1. A group $\langle G, * \rangle$ is a set G with a binary operation $*$ such that (1. associative) $a * (b * c) = (a * b) * c$ for all $a, b, c \in G$. (2. identity) $\exists e \in G$ so that $e * a = a * e = a$ for all $a \in G$. (3. inverse) $\forall a \in G, \exists a' \in G$ where $a * a' = a' * a = e$.

For example, integer addition $\langle \mathbb{Z}, + \rangle$, real number addition $\langle \mathbb{R}, + \rangle$ are groups with identity 0 and a 's inverse $-a$. Positive real numbers and multiplication is a group $\langle \mathbb{R}_+, \times \rangle$ with identity 1 and a 's inverse $\frac{1}{a}$. However, $\langle \mathbb{R}, \times \rangle$ is not a group since $0 \in \mathbb{R}$ does not have an inverse under \times . Real numbers except 0 is again a group $\langle \mathbb{R} \setminus \{0\}, \times \rangle$. \mathbb{Z}_2 is the only group (up to element renaming) of size two:

$+_2$	0	1
0	0	1
1	1	0

We can think of $+_2$ as the XOR function or mod-2 addition.

For any set $A = \{a_1, \dots, a_n\}$, its power set forms a group $\langle 2^A, +_2 \rangle$ where $+_2$ is the symmetric difference: $B +_2 C = (B \cup C) \setminus (B \cap C)$. The identity is the empty set \emptyset , and the inverse of any $B \subseteq A$ is B itself.

Definition 2. A group G is abelian if the operation $*$ is commutative: $\forall a, b \in G, a * b = b * a$.

All groups in this paper are abelian. For an example of non-abelian groups, consider $n \times n$ invertible matrices under matrix multiplication.

Definition 3. A subset $H \subseteq G$ of a group $\langle G, * \rangle$ is a subgroup of G if $\langle H, * \rangle$ is itself a group.

$\{e\}$ is the trivial subgroup of any group G (we often omit the operation when it is clear). $\langle \mathbb{R}_+, \times \rangle$ is a subgroup of $\langle \mathbb{R} \setminus \{0\}, \times \rangle$ by restricting multiplication to positive numbers. Note however multiplication on negative numbers $\langle \mathbb{R}_-, \times \rangle$ is not a subgroup because the result is not in \mathbb{R}_- .

Definition 4. Given a subgroup H of an abelian group G , for any $a \in G$, the set $a * H = \{a * h \mid h \in H\}$ is the coset of H represented by a .

Consider $H = \mathbb{R}_+$ and $G = \mathbb{R} \setminus \{0\}$. Then $3.14 \times \mathbb{R}_+$ is a coset which is the same as \mathbb{R}_+ . In fact for any $a > 0$, $a \times \mathbb{R}_+ = \mathbb{R}_+$, i.e., many different a 's represent the same coset. On the other hand, $-1 \times \mathbb{R}_+ = \mathbb{R}_-$, so \mathbb{R}_- is a coset represented by -1 (or any negative number, for that matter). Since \mathbb{R}_- is not a group, we see the cosets do not have to be subgroups. Also note that the two cosets, \mathbb{R}_+ and \mathbb{R}_- , have equal size and partition G . This fact will be important for counting cycles for homology later.

We now consider mappings from one group $\langle G, * \rangle$ to another $\langle G', \star \rangle$.

Definition 5. A map $\phi : G \mapsto G'$ is a homomorphism if $\phi(a * b) = \phi(a) \star \phi(b)$ for $\forall a, b \in G$.

For example, the groups $\langle \mathbb{R}_+, \times \rangle$ and $\langle \mathbb{Z}_2, +_2 \rangle$ do not look similar at all. But there is a trivial homomorphism $\phi(a) = 0, \forall a \in \mathbb{R}_+$. Note the last 0 is in \mathbb{Z}_2 . This simply says that we map all positive real numbers to the "0" in mod-2 addition. Obviously $0 = \phi(a \times b) = \phi(a) +_2 \phi(b) = 0 +_2 0 = 0$ for $\forall a, b \in \mathbb{R}_+$.

As another example, consider the group of (somewhat artificial) negation in natural language: $G_N = \{\sqcup, \text{not}\}$ with the following operation, where \sqcup stands for whitespace:

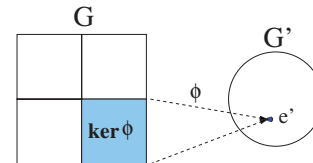
$*$	\sqcup	not
\sqcup	\sqcup	not
not	not	\sqcup

i.e., single negation stays while double negation cancels. There is a homomorphism between G_N and \mathbb{Z}_2 : $\phi(\sqcup) = 0, \phi(\text{not}) = 1$. In fact, G_N and \mathbb{Z}_2 are identical up to renaming. There is a name for such homomorphisms:

Definition 6. A homomorphism that is a one-to-one correspondence is called an isomorphism.

Definition 7. The kernel of a homomorphism $\phi : G \mapsto G'$ is $\ker \phi = \{a \in G \mid \phi(a) = e'\}$. In other words, the kernel is the elements that map to identity.

Theorem 1. For any homomorphism $\phi : G \mapsto G'$, $\ker \phi$ is a subgroup of G .



Because $\ker \phi$ is a subgroup (depicted as the blue square above), we can partition G into cosets of the form $a * \ker \phi$ for $a \in G$. These cosets are the white or blue squares. For example, $\phi : \langle \mathbb{R} \setminus \{0\}, \times \rangle \mapsto G_N$ with $\phi(a) = \sqcup$ if $a > 0$ and "not" if $a < 0$, then $\ker \phi = \mathbb{R}_+$ is one coset and \mathbb{R}_- is the only other coset.

We need one more piece of definition. Let $\langle H, * \rangle$ be a subgroup of an abelian group $\langle G, * \rangle$. We can introduce a new binary operation \star not on the elements of G but on the cosets of H : $(a * H) \star (b * H) = (a * b) * H, \forall a, b \in G$. The operation \star is well-defined and does not depend on the particular choice of representer.

Definition 8. The cosets $\{a * H \mid a \in G\}$ under the operation \star form a group, called the quotient group G/H .

It is useful to think of quotient groups as "higher level" groups defined on the squares in the previous picture. $\ker \phi$ (the blue square) is a subgroup of G . The elements of the quotient group $G/\ker \phi$ are the cosets of $\ker \phi$, i.e. all the squares. In a previous example $G = \mathbb{R} \setminus \{0\}$ and $\ker \phi = \mathbb{R}_+$, and there were two cosets: \mathbb{R}_+ and \mathbb{R}_- . Thus the quotient group $(\mathbb{R} \setminus \{0\})/\mathbb{R}_+$ is a small group with those two cosets as elements. Furthermore, note $R_- \star R_- = (-1 \times R_+) \star (-1 \times R_+) = (-1 \times -1) \times R_+ = 1 \times R_+ = R_+$. Therefore, this quotient group $(\mathbb{R} \setminus \{0\})/\mathbb{R}_+$ is isomorphic to \mathbb{Z}_2 .

Definition 9. Let $S \subset G$. The subgroup generated by S , $\langle S \rangle$, is the subgroup of all elements of G that can be expressed as the finite operation of elements in S and their inverses.

For example, \mathbb{Z} is itself the subgroup generated by $\{1\}$, the group of even integers is the subgroup of \mathbb{Z} generated by $\{2\}$.

Definition 10. The rank of a group G is the size of the smallest subset that generates G .

For example, $\text{rank}(\mathbb{Z}) = 1$ since $\mathbb{Z} = \langle \{1\} \rangle$. $\text{rank}(\mathbb{Z} \times \mathbb{Z}) = 2$ since $\mathbb{Z} \times \mathbb{Z} = \langle \{(0, 1), (1, 0)\} \rangle$. Note there is no one-element basis for $\mathbb{Z} \times \mathbb{Z}$.

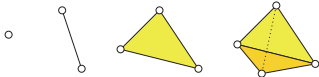
Group theory is important because when counting “holes” in homology, G will be the group of cycles (the rubber bands). The blue square will be the subgroup of “uninteresting rubber bands” that do not surround holes, similar to the earlier blue and green rubber bands. The quotient group “all rubber bands”/“uninteresting rubber bands” will identify holes. However, the rubber bands are continuous and difficult to compute. We first need to discretize the space into a simpler structure called simplicial complex.

2.2 Simplicial Homology

The building blocks of our discrete space are simplices.

Definition 11. A p -simplex σ is the convex hull of $p + 1$ affinely independent points $x_0, x_1, \dots, x_p \in \mathbb{R}^d$. We denote $\sigma = \text{conv}\{x_0, \dots, x_p\}$. The dimension of σ is p .

Affinely independent means the p vectors $x_i - x_0$ for $i = 1 \dots p$ are linearly independent, i.e., they are in general position. The convex hull is simply the solid polyhedron determined by the $p+1$ vertices. A 0-simplex is a vertex, 1-simplex an edge, 2-simplex a triangle, and 3-simplex a tetrahedron:



Definition 12. A face of σ is $\text{conv} S$ where $S \subset \{x_0, \dots, x_p\}$ is a subset of the $p + 1$ vertices.

For example, a tetrahedron has four triangle faces corresponding to the four subsets S obtained by removing one vertex at a time from σ . These four triangle faces are 2-simplices themselves. It also has six edge faces and four singleton vertex faces.

Our space of interest is properly arranged simplices:

Definition 13. A simplicial complex K is a finite collection of simplices such that $\sigma \in K$ and τ being a face of σ implies $\tau \in K$, and $\sigma, \sigma' \in K$ implies $\sigma \cap \sigma'$ is either empty or a face of both σ and σ' .

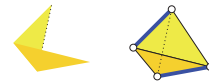
The intuition of simplicial complex is that if a simplex is in K , all its faces need to be in K , too. In addition, the simplices have to be glued together along whole faces or be separate. The figure on the left is a simplicial complex, while the one on the right is not:



Simplicial complex plays the role of the yellow space in the rubber band example. We next introduce the discrete version of the rubber bands.

Definition 14. A p -chain is a subset of p -simplices in a simplicial complex K .

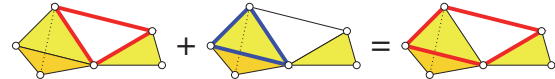
For example, let K be a tetrahedron. By definition the four triangle faces (i.e., 2-simplices) are in K , too. A 2-chain is a subset of these four triangles, e.g., all four triangles, the bottom triangle face only, or the empty set. There are 2^4 distinct 2-chains. Similarly, by definition all six edges of the tetrahedron are in K , too. Thus, there are 2^6 distinct 1-chains. Despite the name “chain,” a p -chain does not have to be connected. The figure below shows a 2-chain on the left and a 1-chain (the blue edges) on the right:



Recall for any set A , its power set forms a group $\langle 2^A, +_2 \rangle$.

Definition 15. The set of p -chains of a simplicial complex K form a p -chain group C_p .

When adding two p -chains we get another p -chain with duplicate p -simplices cancel out. We have a separate chain group for each dimension p . Below is an example of 1-chain addition:

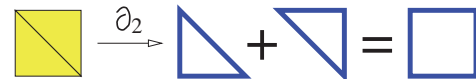


Definition 16. The boundary of a p -simplex is the set of $(p - 1)$ -simplices faces.

The boundary of a tetrahedron is the set of four triangles faces; the boundary of a triangle is its three edges; the boundary of an edge is its two vertices.

Definition 17. The boundary of a p -chain is the $+_2$ sum of the boundaries of its simplices. Taking the boundary is a group homomorphism ∂_p from C_p to C_{p-1} .

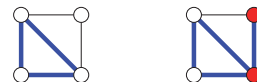
Note faces shared by an even number of p -simplices in the chain will cancel out:



We have finally reached our discrete p -dimensional rubber bands: the p -cycles.

Definition 18. A p -cycle c is a p -chain with empty boundary: $\partial_p c = 0$ (the identity in C_{p-1}).

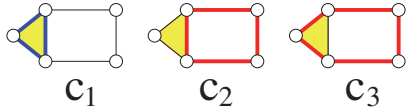
The figure below shows a 1-cycle in blue on the left, and a 1-chain on the right that is not a cycle because it has the red boundary vertices.



Let Z_p be all the p -cycles, i.e., all the “rubber bands.” Since $\partial_p Z_p = 0$, by definition 7 Z_p is the kernel $\ker \partial_p$, which is a subgroup of C_p .

We now identify the “uninteresting rubber bands.” It may not be obvious but the boundary of any higher order $(p + 1)$ -chain is always a p -cycle. For example, the left figure below shows a simplicial complex containing a $(p + 1) = 2$ chain (the yellow tri-

angle). Its boundary c_1 (blue) is indeed a 1-cycle.



Theorem 2. For every p and every $(p + 1)$ -chain c , $\partial_p(\partial_{p+1}c) = 0$.

Definition 19. A p -boundary-cycle is a p -cycle that is also the boundary of some $(p + 1)$ -chain.

Let $B_p = \partial_{p+1}C_{p+1}$, namely all the p -boundary-cycles. B_p are the uninteresting rubber bands. In the example above, $B_1 = \{0, c_1\}$, none surrounding any holes. It is easy to see that B_p is a group, therefore a subgroup of Z_p (all rubber bands).

Are there “interesting rubber bands”? In other words, do we have anything in Z_p besides B_p ? It depends on the structure of the simplicial complex. In the example above, the 1-cycles c_2 and c_3 (red) are not in B_1 since the rectangle does not contain any 2-simplices. These are interesting because they surround the hole in the rectangle. In fact, we can drag the rubber band c_2 over the yellow triangle and turn it into c_3 . Formally, we do this by $c_3 = c_2 + c_1$. Intuitively, c_2 and c_3 are equivalent in the hole they surround. More generally, such equivalence class is obtained by $c + B_p$: we are allowed to drag a p -cycle rubber band c over any $(p + 1)$ -simplices without changing the holes (or the lack thereof) it surrounds.

Returning to the example, we now see all the 1-cycles for this simplicial complex: $Z_1 = \{0, c_1, c_2, c_3\}$. The uninteresting ones are $B_1 = \{0, c_1\}$, a subgroup of Z_1 . The interesting ones are $c_2 + B_1 = c_3 + B_1 = \{c_2, c_3\}$: this should remind us of cosets and quotient group.

Definition 20. The p -th homology group is the quotient group $H_p = Z_p/B_p$. The p -th Betti number is its rank: $\beta_p = \text{rank}(H_p)$.

We have arrived at the core of homology. In our example, $H_1 = \{0, c_1, c_2, c_3\}/\{0, c_1\}$ which is isomorphic to \mathbb{Z}_2 . The first Betti number is $\beta_1 = \text{rank}(\mathbb{Z}_2) = 1$, indicating one independent 1st-order hole not filled in by triangles.

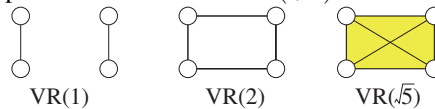
In general, β_p is the number of independent p -th holes. For example, a tetrahedron has $\beta_0 = 1$ since the shape is connected, $\beta_1 = \beta_2 = 0$ since there is no holes or voids. A hollow tetrahedron has $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$ because of the void. Further removing the four triangle faces but keeping the six edges, the skeleton has $\beta_0 = 1, \beta_1 = 3$ (there are 4 triangular holes but one is the sum of the other three), $\beta_2 = 0$ (no more void). Finally removing the edges but keeping the four vertices, $\beta_0 = 4$ (4 connected components each a single vertex) and $\beta_1 = \beta_2 = 0$.

2.3 Persistent Homology

Usually we are given data as a point cloud $x_1, \dots, x_n \in \mathbb{R}^d$. Where does the simplicial complex come from in the first place? One way to create it is to examine all subsets of points. If any subset of $p + 1$ points are “close enough,” we add a p -simplex σ with those points as vertices to the complex:

Definition 21. A Vietoris-Rips complex of diameter ϵ is the simplicial complex $VR(\epsilon) = \{\sigma \mid \text{diam}(\sigma) \leq \epsilon\}$.

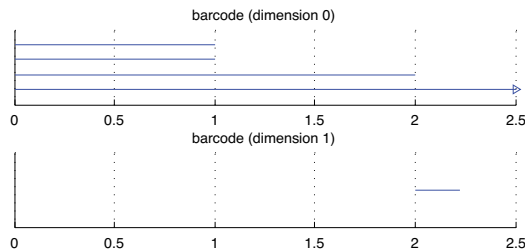
Here $\text{diam}(\sigma)$ is the largest distance between two points in σ . Note if $\sigma \in VR(\epsilon)$, all its faces are, too. The following figure shows four points $(0,0), (0,1), (2,1), (2,0)$ and the Vietoris-Rips complex with different ϵ . $VR(\sqrt{5})$ is a flat tetrahedron.



A natural question is what best ϵ to use for any data set. Persistent homology examines all ϵ 's to see how the system of holes change.

Definition 22. An increasing sequence of ϵ produces a filtration, i.e., a sequence of increasing simplicial complexes $VR(\epsilon_1) \subseteq VR(\epsilon_2) \subseteq \dots$, with the property that a simplex enters the sequence no earlier than all its faces.

Persistent homology tracks homology classes along the filtration: at what value of ϵ does a hole appear, and how long does it persist till it is filled in? A convenient way to visualize persistent homology is the barcode plot shown below. The x -axis is ϵ . Each horizontal bar represents the birth-death of a separate homology class. Longer bars correspond to more robust topological structure in the data.



The top panel shows H_0 (0-th order holes or clusters). At $\epsilon = 0$ there are four bars for the four disconnected vertices in $VR(0)$. The Betti number at any given ϵ is the number of bars above it, in this case $\beta_0 = 4$. At $\epsilon = 1$ two edges appear in $VR(1)$, reducing the number of connected components to two. This is why the top two bars die and β_0 reduces to 2. At $\epsilon = 2$, $VR(2)$ forms a rectangle and becomes fully connected, so one more bar dies and $\beta_0 = 1$ thereafter. The remaining bar represents the one vertex that grabs everything to eventually become the fully connected component. It never dies (represented by the arrow at the end of the bar). We note that the clusters are precisely those obtained from hierarchical clustering with single-linkage.

The bottom panel shows H_1 (1st order holes). In the example above, a homology class corresponding to the hole is born at $\epsilon = 2$ when the rectangle becomes connected. It persists until $\epsilon = \sqrt{5}$ and dies because the Vietoris-Rips complex becomes the solid tetrahedron. This is represented by the single short bar. The Betti number is $\beta_1 = 1$ in the interval $[2, \sqrt{5})$ and 0 otherwise.

3 A Natural Language Processing Application

We all have the intuition that some documents tell a straight story while others twist and turn. We hope persistent homology captures such structures. We assume that a document has been divided into small units x_1, \dots, x_n . We are given a distance function $D(x_i, x_j) \geq 0$ so that similar units have small

distance. We will focus on the 0-th (clusters) and 1st (holes) order homology classes. We introduce two algorithms: SIF and SIFTS.

Similarity Filtration (SIF). SIF is a simple method to compute persistent homology by creating a Vietoris-Rips complex over x_1, \dots, x_n , where the diameter measures the similarity between text units:

1. $D_{max} = \max D(x_i, x_j), \forall i, j = 1 \dots n$
2. **FOR** $m = 0, 1, \dots, M$
3. Add $VR\left(\frac{m}{M} D_{max}\right)$ to the filtration
4. **END**
5. Compute persistent homology on the filtration

The growing diameter corresponds to allowing looser tie-backs: more dissimilar text units are linked together to form simplices in the Vietoris-Rips complex. Note the order of $x_1 \dots x_n$ is ignored.

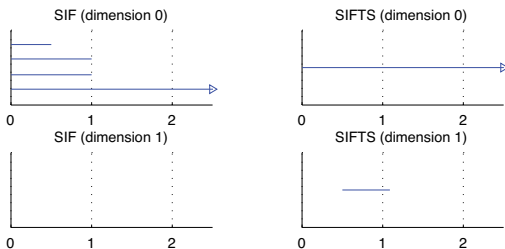
Similarity Filtration with Time Skeleton (SIFTS). We may be more interested in the *flow* of the document. Recall we “connect the dots” in the introduction. This prompts us to add “time edges” $(x_i, x_{i+1}), i = 1 \dots n - 1$ to the simplicial complex before any similarity filtration. These edges form a “time skeleton” by connecting units in document order. The SIFTS algorithm implements time skeleton by adding the following preprocessing step before the SIF algorithm in section 3:

$$0. D(x_i, x_{i+1}) = 0 \text{ for } i = 1, \dots, n - 1$$

The key difference between SIF and SIFTS is that a time-skeleton edge can be arbitrarily long as measured by $D()$. By adding the time skeleton upfront, we enable “tie-back” holes in SIFTS. This is illustrated by the toy document $(0, 0), (1, 0), (2, 0), (-\frac{1}{2}, 0)$ below, with the Vietoris-Rips complex $VR(0.5)$:



SIF sees the Vietoris-Rips complex on the left as four vertices and an edge between $(0, 0), (-\frac{1}{2}, 0)$. Even though the edge represents a tie-back between the first and last units, no hole has formed. In contrast, SIFTS sees the combined complex on the right with time skeleton in red. The similarity and time edges together form a hole (i.e., $\beta_1 = 1$). The complete barcodes for SIF and SIFTS are presented below. SIF detects no hole at all ($\beta_1 = 0$ always): as ϵ increase the filtration fills the complex with solid triangles, preventing holes. The hole detected by SIFTS persists until ϵ is large enough to cover $(1, 0)$ and $(-\frac{1}{2}, 0)$. Also note SIFTS complex is trivially connected by the time skeleton, hence $\beta_0 = 1$ always.



3.1 On Nursery Rhymes and Other Stories

We now illustrate persistent homology as computed by SIF and SIFTS on a few nursery rhymes. Nursery rhymes are repetitive and familiar, ideal for homology examples. Each unit is a sentence. We perform minimum tokenization by case-folding and punctuation removal only. The distance $D()$ is the Euclidean distance between sentence-level bag-of-words count vectors. All filtrations has $M = 100$ steps.

Figure 1(a) shows *Itsy Bitsy Spider*. Its homology is strikingly similar to the previous toy document, as the spider climbed up the water spout in both the 1st and the 4th sentences. This hole is detected by SIFTS but not SIF.

Figure 1(b) shows *Row Row Row Your Boat*. Its four sentences are distinct from each other, forming a “linear progression.” Both SIF and SIFTS give $\beta_1 = 0$: there is no hole.

Figure 1(c) shows *London Bridge is Falling Down*. The lyric has $n = 48$ sentences; The sentence “My fair Lady” repeats 12 times. With the time skeleton, SIFTS therefore detects 11 independent holes ($\beta_1 = 11$) right away in $VR(0)$. These holes are not detected by SIF. Both SIF and SIFTS detect more holes later, some are caused by the near-repetition “Build it up with X and Y ”, where X, Y vary from wood and clay to silver and gold.

We now move on to longer documents. Here and in next section, the text units are natural paragraphs (or chapters for *Alice*). We perform Penn Treebank tokenization, case-folding, punctuation removal, and SMART stopword removal [Salton, 1971]. Each text unit is converted to a tf.idf vector, where idf is computed within the document. We compute the cosine similarity then take the angular distance:

$$D(x_i, x_j) = \cos^{-1} \left(\frac{x_i^\top x_j}{\|x_i\| \|x_j\|} \right).$$

Figure 1(d,e,f) show the barcodes on three stories. In general, SIFTS detects more holes and detects them earlier than SIF. The homology classes that persist the longest tend to be reappearance of salient words. For example, in *Red-Cap* the first SIFTS hole is between the sentences “The better to see you with, my dear” and “The better to eat you with!”

3.2 On Child and Adolescent Writing

As a real world example, we quantitatively study whether children’s writing become structurally richer as they grow up. Specifically, our hypothesis is that older writers have more 1-homology groups than younger writers.

We use the LUCY corpus which contains roughly matched child and adolescent writing [Sampson, 2003]. We merge the F,H,K,M groups (ages 9–12, 150 essays) to form a child-writing set. We use the E group (undergraduates, 48 essays) as the adolescent-writing set. The main differences between the two sets are age and average article length (child=11.6 sentences, adolescent=25.8 sentences), see LUCY documentation for other minor differences.

We compute each essay’s SIFTS barcode. To facilitate comparison, we extract two summary statistics. The first is $|H_1|$, the total number of 1st-order persistent homology classes (holes) over the whole ϵ range. This is obtained by counting the number of bars. Note $|H_1| \geq \beta_1$ since the Betti number is for a specific ϵ . The second is ϵ^* , the smallest ϵ

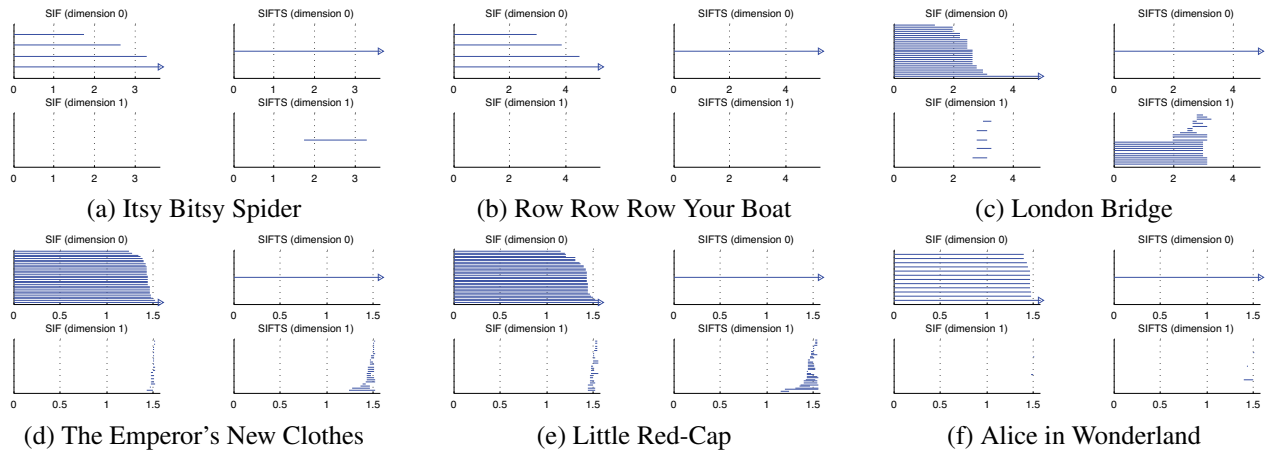


Figure 1: Persistent homology on nursery rhymes and other stories

	child	adolescent	adol. trunc.
holes?	87%	100%*	98%*
$ H_1 $	3.0 (± 0.2)	17.6 (± 0.9)*	3.9 (± 0.2)*
ϵ^*	1.35 (± 0.02)	1.27 (± 0.02)*	1.38 (± 0.01)

Table 1: Statistics on child vs. adolescent writing. Entries significantly different from *child* are marked by *

when the first hole in H_1 forms. If there is no hole we set $\epsilon^* = \pi/2$, the largest angular distance possible.

The first two columns in Table 1 show a marked difference between child vs. adolescent writing. Only 87% of child essays have holes while all adolescent essays do ($p = 0.01$, Fisher’s test). The average child essay has 3 holes while adolescent has 17.6 ($p = 10^{-55}$, t -test). First hole appears earlier in adolescent ($p = 0.01$, t -test).

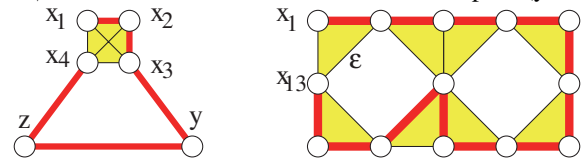
One has reason to suspect that the homology differs solely because adolescent essays are about twice as long. We thus create a third “adolescent truncated” data set, where we keep the first 11 sentences in each adolescent essay to match child writing. This perhaps removed many later tie-backs in the essays. The third column in Table 1, however, still shows some differences compared to child writing: more truncated adolescent essays contain holes ($p = 0.03$, Fisher’s test). On average a truncated essay has one more hole ($p = 0.03$, t -test). But the first-birth ϵ^* is no longer significantly different ($p = 0.2$, t -test).

We conclude that persistent homology detects significant differences between child and adolescent writing using only structural features. The point is not that classifying the two classes requires such sophisticated machinery – simpler features such as word usage probably suffice. Rather, our experiment shows that there is useful information in homology. Incorporating such information into existing text representation for NLP tasks such as discourse structure modeling or parsing can potentially enhance these tasks. This remains future work.

4 Discussion: Merely Counting Repeats?

Our nursery rhyme examples may give the impression that persistent homology computed by SIFTS is simply finding repeated (ϵ -close) text units. After all, in a document $x_1 \rightsquigarrow x_2 \rightsquigarrow x_3$ where x_1, x_2, x_3 are within ϵ of each other and \rightsquigarrow represents long sequence of mutually dissimilar units, SIFTS will identify exactly two independent holes: $x_1 \rightsquigarrow x_2$ where x_2 ties back to x_1 , and similarly $x_2 \rightsquigarrow x_3$. k such repeats of x will generate $k - 1$ holes. It seems one can just count k the number of repeats to get the Betti number $\beta_1 = k - 1$.

This impression is incomplete. Consider the document $x_1 x_2 x_3 y z x_4$ depicted on left, where y and z are distant. The SIFTS time skeleton is in red. There are $k = 4$ repeats of x but $\beta_1 = 1$ not 3, since the x ’s form a 3-simplex (yellow).



Perhaps such problem can be dealt with by preprocessing, where one merges *contiguous* units within ϵ ? Surely with $x_1 x_2 x_3$ merged into a super unit x' , we can use counting again to detect two repeats x', x_4 and correctly infer one hole. However, consider another document $x_1 x_2 \dots x_{13}$ on the right, where all contiguous unit pairs are within ϵ (the short diagonal length). The preprocessing will merge all units into a single super unit, thus incorrectly predicting 0 holes. In contrast, SIFTS can correctly identify the two holes. Homology is not just counting repeated text units.

The barcodes in this paper were computed with the javaPlex software [Tausz *et al.*, 2011]. Our data and SIF, SIFTS code is online at <http://pages.cs.wisc.edu/~jerryzhu/publications.html>.

Acknowledgments: I thank Kevyn Collins-Thompson for discussions on corpora, the anonymous reviewers for helpful comments, and the support of NSF IIS-0953219, IIS-1216758, IIS-1148012, IIS-0916038.

References

- [Balakrishnan *et al.*, 2012] Sivaraman Balakrishnan, Alessandro Rinaldo, Don Sheehy, Aarti Singh, and Larry A. Wasserman. Minimax rates for homology inference. In *The fifteenth international conference on Artificial Intelligence and Statistics (AISTATS)*, pages 64–72, 2012.
- [Balakrishnan *et al.*, 2013] Sivaraman Balakrishnan, Britany Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Statistical inference for persistent homology. In *arXiv:1303.7117*, 2013.
- [Carlsson, 2009] Gunnar Carlsson. Topology and data. *Bulletin (New Series) of the American Mathematical Society*, 46(2):255–308, 2009.
- [Chung *et al.*, 2009] Moo K. Chung, Peter Bubenik, Peter T. Kim, Kim M. Dalton, and Richard J. Davidson. Persistence diagrams of cortical surface data. In *Information Processing in Medical Imaging*, pages 386–397, 2009.
- [de Silva and Ghrist, 2007a] Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7:339–358, 2007.
- [de Silva and Ghrist, 2007b] Vin de Silva and Robert Ghrist. Homological sensor networks. *Notices of the American Mathematical Society*, 54, 2007.
- [Edelsbrunner and Harer, 2007] H. Edelsbrunner and J. Harer. Persistent homology — a survey. In *Twenty Years After*, eds. J. E. Goodman, J. Pach and R. Pollack, AMS., 2007.
- [Edelsbrunner and Harer, 2010] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Applied mathematics. Amer Mathematical Society, 2010.
- [Freedman and Chen, 2011] Daniel Freedman and Chao Chen. Algebraic topology for computer vision. In Sota R. Yoshida, editor, *Computer Vision*, chapter 5, pages 239–268. Nova Science Pub. Inc., 2011.
- [Gamble and Heo, 2010] Jennifer Gamble and Giseon Heo. Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *J. Multivariate Analysis*, 101(9):2184–2199, 2010.
- [Giblin, 2010] P. Giblin. *Graphs, Surfaces and Homology*. Cambridge University Press, 2010.
- [Gous, 1999] Alan Gous. Spherical subfamily models. Technical report, 1999.
- [Hall and Hofmann, 2000] Keith Hall and Thomas Hofmann. Learning curved multinomial subfamilies for natural language processing and information retrieval. In *ICML*, pages 351–358, 2000.
- [Hatcher, 2001] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, first edition, December 2001.
- [Kasson *et al.*, 2007] Peter M. Kasson, Afra Zomorodian, Sanghyun Park, Nina Singhal, Leonidas J. Guibas, and Vijay S. Pande. Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23(14):1753–1759, 2007.
- [Lebanon *et al.*, 2007] Guy Lebanon, Yi Mao, and Joshua V. Dillon. The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research*, 8:2405–2441, 2007.
- [Lebanon, 2006] Guy Lebanon. Sequential document representations and simplicial curves. In *UAI*. AUAI Press, 2006.
- [Rote and Vegter, 2006] Günter Rote and Gert Vegter. Computational topology: an introduction. In Jean-Daniel Boissonnat and Monique Teillaud, editors, *Effective Computational Geometry for Curves and Surfaces*, Mathematics and Visualization, chapter 7, pages 277–312. Springer-Verlag, 2006.
- [Salton, 1971] G. Salton, editor. *The SMART Retrieval System Experiments in Automatic Document Processing*. Englewood Cliffs: Prentice-Hall, 1971.
- [Sampson, 2003] Geoffrey R. Sampson. The structure of children’s writing: moving from spoken to adult written norms. In S. Granger and S. Petch-Tyson, editors, *Extending the Scope of Corpus-Based Research*, pages 177–93. Rodopi, 2003. <http://www.grsampson.net/RLucy.html>.
- [Singh *et al.*, 2008] Gurjeet Singh, Facundo Memoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L. Ringach. Topological analysis of population activity in visual cortex. *J. Vis.*, 8(8):1–18, 6 2008.
- [Tausz *et al.*, 2011] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. Javaplex: A research software package for persistent (co)homology. Software available at <http://code.google.com/javaplex>, 2011.
- [Zomorodian, 2001] Afra Joze Zomorodian. *Computing and comprehending topology: persistence and hierarchical Morse complexes*. PhD thesis, University of Illinois at Urbana-Champaign, 2001.