Topological Data Analysis, Basics and Computation

Jean-Daniel Boissonnat, *Siddharth Pritam*

DataShape INRIA, Sophia-Antipolis, France

February 10, 2020

FGV EMAp

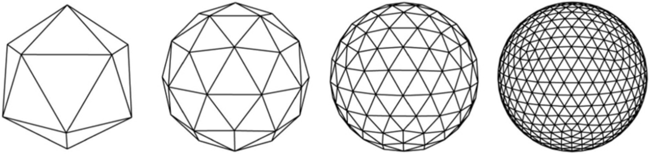
Table of Contents

What is Topology?

- Topology $\sim$ Qualitative Geometry.
- Geometry is about quantities like, Distances and Angles.

What is Topology?

- Topology $\sim$ Qualitative Geometry.
- Geometry is about quantities like, Distances and Angles.



- Different geometrically, same topologically.

Topology

- Topology is the study of continuous deformations (rubber sheet geometry).

Topology

- Topology is the study of continuous deformations (rubber sheet geometry).

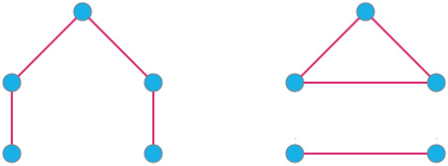
- Allowed: Stretching and Shrinking.
- Not Allowed : Cutting and Gluing.

Different topological spaces

- What are the things Topologists can differentiate?

Different topological spaces

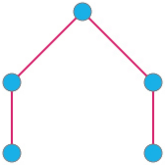
- What are the things Topologists can differentiate?

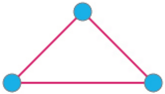


- Above two graphs are topologically different.
- Different number of components (therefore different connectivity).
- More formally: They have different zero-dimensional **Homology**, $H_0()$.

The idea of Homology

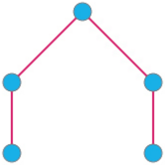
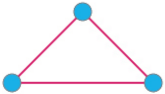
- Different number of cycles (1-dim-holes).





The idea of Homology

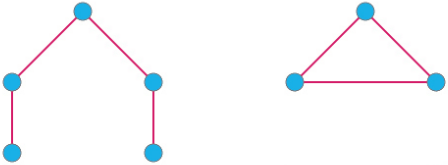
- Different number of cycles (1-dim-holes).

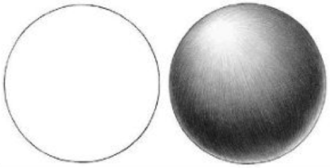
- Different one-dimensional Homology, $H_1()$

The idea of Homology

- Different number of cycles (1-dim-holes).

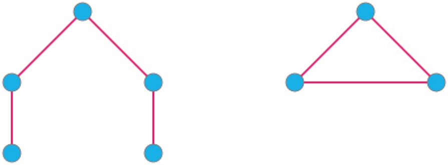


- Different one-dimensional Homology, $H_1()$
- 1-dim Cycle, 2-dim Hole.

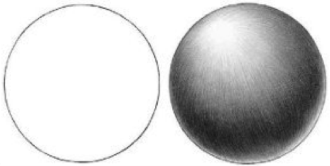


The idea of Homology

- Different number of cycles (1-dim-holes).



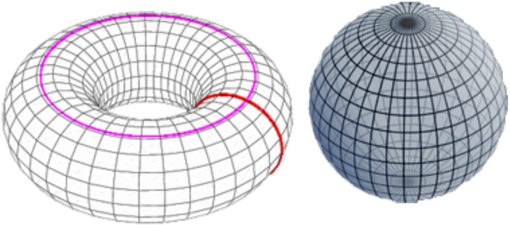
- Different one-dimensional Homology, $H_1()$
- 1-dim Cycle, 2-dim Hole.



- $H_k()$s are vector spaces, **Betti numbers** $\beta_k = Rank(H_k())$

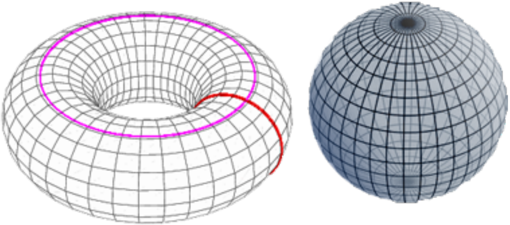
Homology of Sphere and Torus

- Different number of cycles (1-dim-holes).



Homology of Sphere and Torus

- Different number of cycles (1-dim-holes).

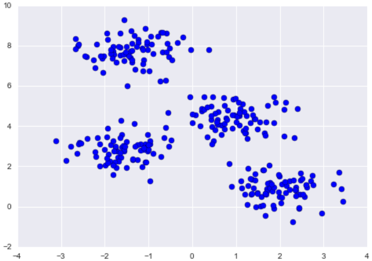


- Sphere : $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$, $\beta_k = 0$ for $k > 2$
- Torus : $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$, $\beta_k = 0$ for $k > 2$
- Betti numbers β_ks could be used to distinguish topological spaces.

Table of Contents

Data

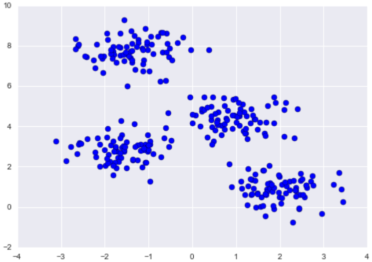
- Data could be financial, biological, material...
- Visualized as point cloud in $\mathbb{R}^d$



- Task: Analyze it topologically.

Data

- Data could be financial, biological, material...
- Visualized as point cloud in $\mathbb{R}^d$



- Task: Analyze it topologically.

It's not all new

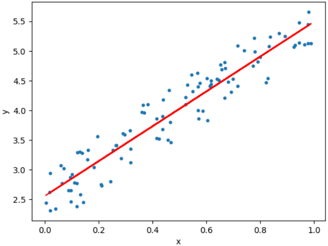
- Clustering.

- Topologically: Connected Component; Compute the 0-th homology group or β_0

It's not all new

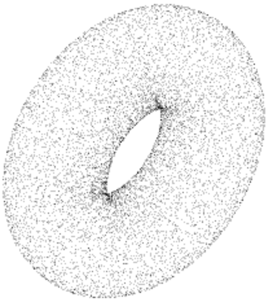
- Clustering.

- Topologically: Connected Component; Compute the 0-th homology group or β_0
- Regression. Hypothesis: Linear shape; Compute the line that fits the best.



The new approach

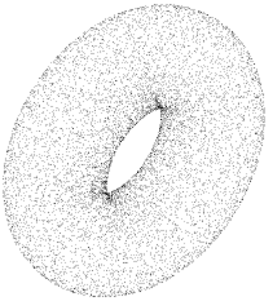
- Hypothesis: Data P is a finite sample of some underlying topological space X.



- Goal: To infer the homology groups or betti numbers of X.

The new approach

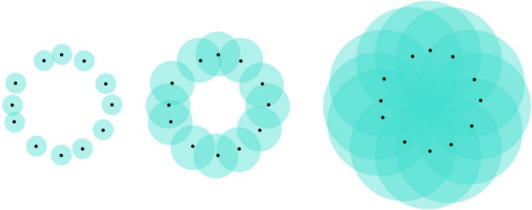
- Hypothesis: Data P is a finite sample of some underlying topological space X.



- Goal: To infer the homology groups or betti numbers of X.
- Problem $P \neq X$, P is discrete, finite, X could be continuous.

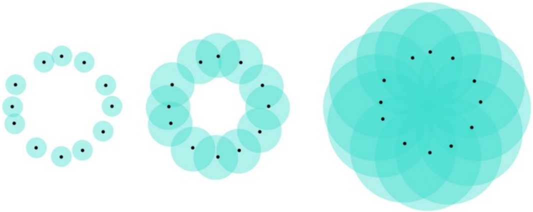
Union of Balls

- Probable Solution: Inflate the points.
- Imagine a ball $B(p, \epsilon)$ of some radius ϵ around all $p \in P$.
- Such that $P^{\epsilon} := \bigcup_{p \in P} (B(p, \epsilon)) \supset X$.



Union of Balls

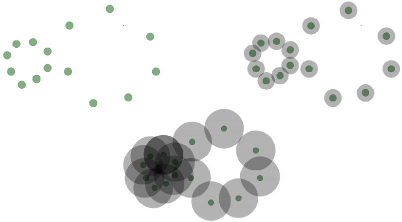
- Probable Solution: Inflate the points.
- Imagine a ball $B(p, \epsilon)$ of some radius ϵ around all $p \in P$.
- Such that $P^{\epsilon} := \bigcup_{p \in P} (B(p, \epsilon)) \supset X$.



- Problem, how to find the right ϵ!.

Union of Balls

- Bigger problem, There might not be one right ϵ.

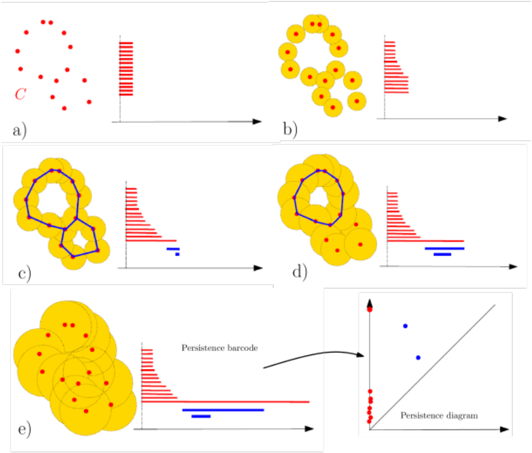


At no scale, can the union of balls determine the two loops simultaneously.

- Solution: Compute homology at all scales, i.e. **Persistent Homology**

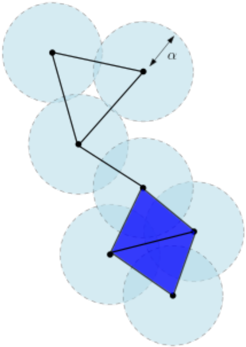
Persistent Homology

- We keep track of the birth and the death of cycles.



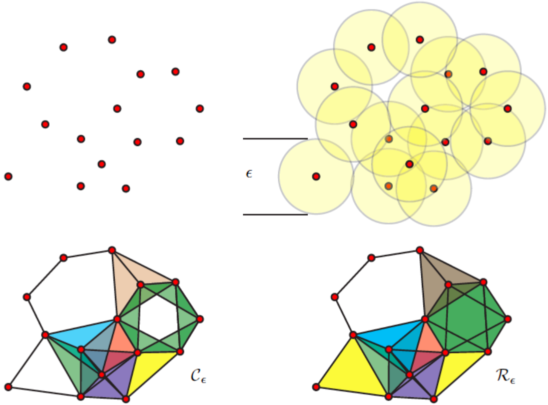
Simplicial Complex

- To perform computation, transform continuous to discrete space.



- Union of balls $\sim$ Čech complex.
- The discrete space is known as **Simplicial Complex**.
- Simplicial Complex : Nicely glued edges, triangle, tetrahedron...

Čech and Rips Complex

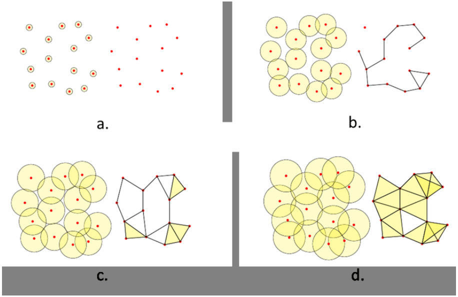


- Rips Complex: An example of flag complex.
- Čech and Rips complex are interleaved.

$$R(P,\epsilon) \subseteq C(P,\sqrt{2}\epsilon) \subseteq R(P,\sqrt{2}\epsilon)$$

Filtration

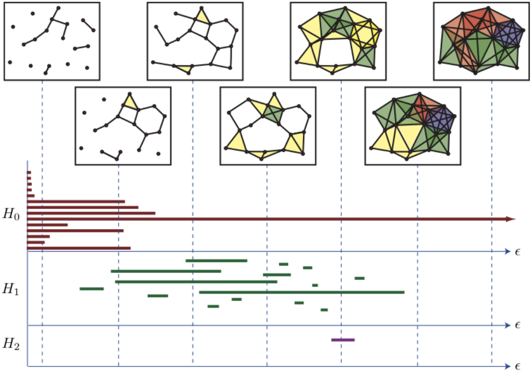
- As radius increases, the čech complex grows with inclusion of new simplices.



- Filtration : A sequence of nested simplicial complexes.
- The **filtration value** of a simplex is the radius at which appears first.

Persistent Homology of a filtration

- Barcode of a filtration.

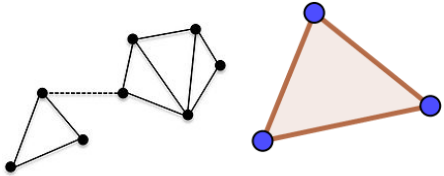


- Small bars → Noise, Big bars → Real features.

Table of Contents

Negative Simplex

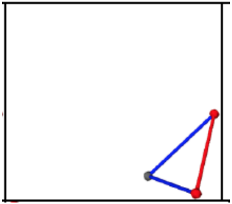
- Negative Simplex: A simplex which destroys a homology group.



- A negative k-simplex destroys a $(k-1)$-cycle.

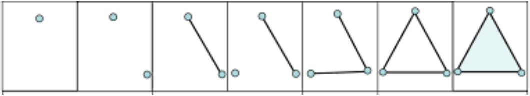
Positive Simplex

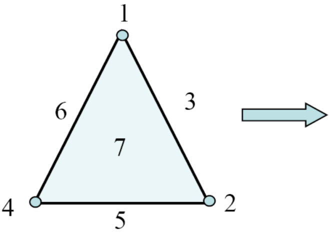
- Positive Simplex: A simplex which creates a homology group.



- A positive k-simplex creates a k-cycle.
- Persistence diagram: Pairings of positive and negative simplices.

Boundary Matrix





$$\begin{array}{cc}
\begin{matrix} & & & & j & & \end{matrix} & \\
\begin{pmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & \boxed{1} & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} & l(j)
\end{array}$$

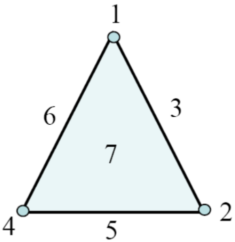
Persistence Algorithm

Algorithm 1 Reduction Algorithm

procedure REDUCE(∂) ▷ ∂ is the boundary matrix.
    $R = \partial$;
    for j = 1 to m **do**
        while there exists $j' < j$ with $low(j') = low(j)$ **do**
            add column j' to column j ▷ mod 2 operation.
        end while
    end for
end procedure ▷ Return R

- Run time complexity: $\mathcal{O}(n^3)$; $n =$ filtration size.

Simple Example



$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & \boxed{1} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \boxed{1} & \boxed{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

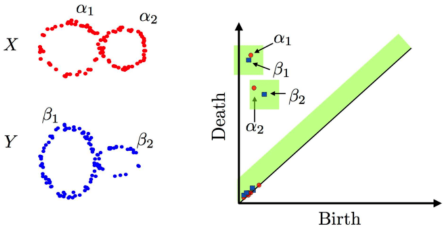
$$\begin{array}{c} \begin{array}{ccccccc} & & & & & C_5 + C_6 & \\ & & & & & \downarrow & \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & \boxed{1} & 0 & 1 & \boxed{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$\begin{array}{c} \begin{array}{ccccccc} & & & & & C_3 + C_6 & \\ & & & & & \downarrow & \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \boxed{1} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Pairs: $(2,3)$ $(4,5)$ $(6,7)$

Stability

- A hallmark result in persistence theory.
- Stablity theorem: Two close(similiar) point sets will have close(similar) barcodes/persistence diagrams.



- Proven by David Cohen-Steiner et. al.

Table of Contents

Motivation

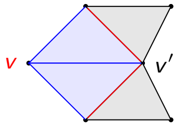
- Computing persistent homology (PH) has $\mathcal{O}(n^{\omega})$ time complexity, n is the filtration size, $\omega \leq 2.4$.
- For massive and high-dimensional datasets, n may be very large (of order of billions).
- Rips complex : Widely used, Easy to compute, however n grows exponentially with dimension.
- Our work reduces the size the filtration by order of magnitude 3-4 using Strong Collapses and Edge Collapse.

Motivation

- Computing persistent homology (PH) has $\mathcal{O}(n^{\omega})$ time complexity, n is the filtration size, $\omega \leq 2.4$.
- For massive and high-dimensional datasets, n may be very large (of order of billions).
- Rips complex : Widely used, Easy to compute, however n grows exponentially with dimension.
- Our work reduces the size the filtration by order of magnitude 3-4 using Strong Collapses and Edge Collapse.
- Major Advantages:
  - Reduction is done on the 1-skeleton $\implies$ Extremely Fast and Memory Efficient.
  - We can compute the exact PH, and a substantially faster approximate PH at a very minimal cost.

Table of Contents

Strong Collapse

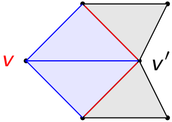


Definition

Dominated vertex: If the link $lk_K(v)$ is a simplicial cone. i.e $lk_K(v) = v'L$.

- Vertex v is said to be dominated by v'.

Strong Collapse



Definition

Dominated vertex: If the link $lk_K(v)$ is a simplicial cone. i.e $lk_K(v) = v'L$.

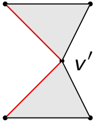
- Vertex v is said to be dominated by v'.

Definition

An elementary strong collapse consists of removal of a *dominated vertex* v from K.

$$K \searrow\searrow^{e} \{K \setminus v\}$$

.



- A series of elementary strong collapses from K to L (subcomplex) is called a strong collapse.

$$K \searrow\searrow L$$

.

- K and L are said to have the same *strong homotopy type*.

Theorem

Strong homotopy type $\implies$ *Simple homotopy type* $\implies$ *Homotopy type.*

- A series of elementary strong collapses from K to L (subcomplex) is called a strong collapse.

$$K \searrow\searrow L$$

.

- K and L are said to have the same *strong homotopy type*.

Theorem

Strong homotopy type $\implies$ *Simple homotopy type* $\implies$ *Homotopy type.*

Lemma

v is dominated by v' iff all the maximal simplices of K that contain v also contain v'.

- A series of elementary strong collapses from K to L (subcomplex) is called a strong collapse.

$$K \searrow\searrow L$$

.

- K and L are said to have the same *strong homotopy type*.

Theorem

Strong homotopy type $\implies$ *Simple homotopy type* $\implies$ *Homotopy type.*

Lemma

v is dominated by v' iff all the maximal simplices of K that contain v also contain v'.

- Retraction map: The vertex map $\mathbf{r} : K \to K \setminus v$ defined as: $r(w) = w$ if $w \neq v$ and $r(v) = v'$.
- Minimal complex : A complex without any dominated vertex.
- Core: K_0 is a core of K, if $K \searrow\searrow K_0$ and K_0 is a minimal complex.
- Every simplicial complex has a unique core upto isomorphism.

Table of Contents

Strong collapse of a Flag Complex

- **Open neighborhood** $N_G(v)$ of v in G is defined as $N_G(v) := \{u \in G \mid [uv] \in E\}$.
- The **closed neighborhood** $N_G[v] := N_G(v) \cup \{v\}$.

Strong collapse of a Flag Complex

- **Open neighborhood** $N_G(v)$ of v in G is defined as $N_G(v) := \{u \in G \mid [uv] \in E\}$.
- The **closed neighborhood** $N_G[v] := N_G(v) \cup \{v\}$.

Lemma

Let K be a flag complex. A vertex $v \in K$ is dominated by v' if and only if $N_G[v] \subseteq N_G[v']$.

Strong collapse of a Flag Complex

- **Open neighborhood** $N_G(v)$ of v in G is defined as $N_G(v) := \{u \in G \mid [uv] \in E\}$.
- The **closed neighborhood** $N_G[v] := N_G(v) \cup \{v\}$.

Lemma

Let K be a flag complex. A vertex $v \in K$ is dominated by v' if and only if $N_G[v] \subseteq N_G[v']$.

Lemma

Core of a flag complex is a flag complex.

Strong collapse of a Flag Complex

- **Open neighborhood** $N_G(v)$ of v in G is defined as $N_G(v) := \{u \in G \mid [uv] \in E\}$.
- The **closed neighborhood** $N_G[v] := N_G(v) \cup \{v\}$.

Lemma

Let K be a flag complex. A vertex $v \in K$ is dominated by v' if and only if $N_G[v] \subseteq N_G[v']$.

Lemma

Core of a flag complex is a flag complex.

- $\implies$ The skeleton of the core can computed using only the graph G of K .

Table of Contents

Dominated Edge

Definition

Dominated edge: If the link $lk_K(e)$ is a simplicial cone. i.e $lk_K(e) = v'L$.

- Edge e is said to be dominated by v'.

Dominated Edge

Definition

Dominated edge: If the link $lk_K(e)$ is a simplicial cone. i.e $lk_K(e) = v'L$.

- Edge e is said to be dominated by v'.
- An elementary edge-collapse consists of removal of a *dominated edge* e from K.

$$K \searrow\searrow^{e} \{K \setminus e\}$$

.

Lemma

Let K be a flag complex. A vertex $e \in K$ is dominated by v' if and only if $N_G[e] \subseteq N_G[v']$.

Dominated Edge

Definition

Dominated edge: If the link $lk_K(e)$ is a simplicial cone. i.e $lk_K(e) = v'L$.

- Edge e is said to be dominated by v'.
- An elementary edge-collapse consists of removal of a *dominated edge* e from K.

$$K \searrow\searrow^{e} \{K \setminus e\}$$

.

Lemma

Let K be a flag complex. A vertex $e \in K$ is dominated by v' if and only if $N_G[e] \subseteq N_G[v']$.

Lemma

1-*core of a flag complex is a flag complex.*

Dominated Edge

Definition

Dominated edge: If the link $lk_K(e)$ is a simplicial cone. i.e $lk_K(e) = v'L$.

- Edge e is said to be dominated by v'.
- An elementary edge-collapse consists of removal of a *dominated edge* e from K.

$$K \searrow\searrow^{e} \{K \setminus e\}$$

.

Lemma

Let K be a flag complex. A vertex $e \in K$ is dominated by v' if and only if $N_G[e] \subseteq N_G[v']$.

Lemma

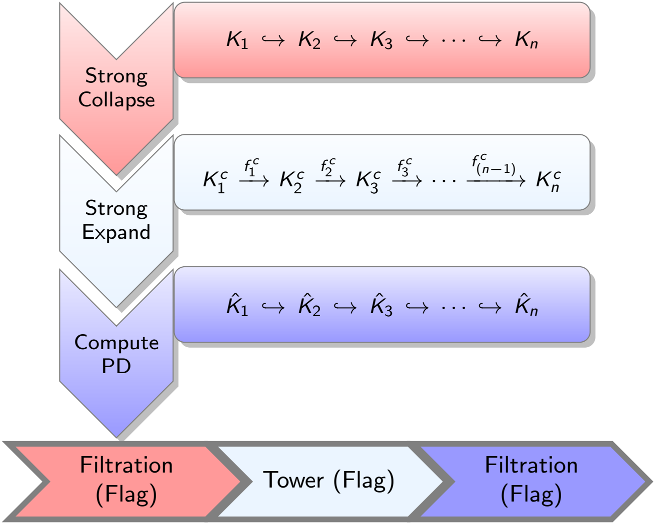
1-*core of a flag complex is a flag complex.*

- $\implies$ The skeleton of the core can computed using only the graph G of K .

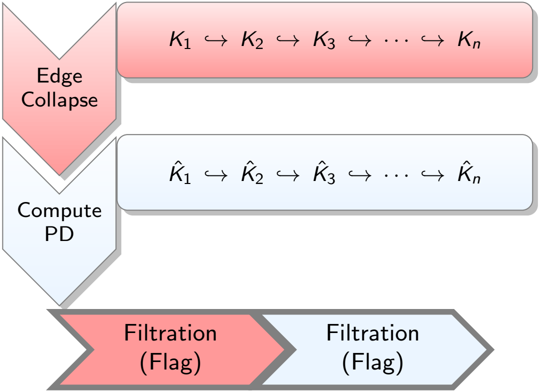
Table of Contents

Preprocessing Flow

Objective : To compute the PD of a filtration of a flag complex (flag filtration).



Edge collapse flow



Experiments

- VertexCollapser +PD(Gudhi)

Data	Pnt	VertexCollapser +PD(Gudhi)				
		dim	Pre-Time	Tot-Time	*Step(btl-dist)*	Snaps
netw-sc	379	∞	7.28	7.38	0.02	263
"	"	∞	13.93	14.03	0.01	531
"	"	∞	366.46	366.56	0	8420
senate	103	∞	2.53	2.54	0.001	403
"	"	∞	15.96	15.98	0	2728

- Ripser.

Data	Pnt	Threshold	Val		Val		Val	
			dim	Time	*dim*	Time	*dim*	Time
netw-sc	379	5.5	4	25.3	5	231.2	6	∞
senate	103	0.415	3	0.52	4	5.9	5	52.3
"	"	"	6	406.8	7	∞		
eleg	297	0.3	3	8.9	4	217	5	∞
HIV	1088	1050	2	31.35	3	∞		
torus	2000	1.5	2	193	3	∞		

Table: Time is the total time (in seconds) taken by Ripser. ∞ means that the experiment ran longer than 12 hours or crashed due to memory overload.

Experiments

- EdgeCollapser +PD(Gudhi)

Data	Pnt	Thrsld	EdgeCollapser +PD				
			Edge(I)/Edge(C)	Size/Dim	*dim*	Pre-Time	Tot-Time
netw-sc	379	5.5	8.4K/417	1K/6	∞	0.62	0.73
senate	103	0.415	2.7K/234	663/4	∞	0.21	0.24
eleg	297	0.3	9.8K/562	1.8K/6	∞	1.6	1.7
HIV	1088	1050	182K/6.9K	86.9M/?	6	491	2789
torus	2000	1.5	428K/14K	44K/3	∞	288	289

Table: Time (in seconds) taken by Edge-Collapser and total time (in seconds) including PD computation (Tot-Time).

- VertexCollapser +PD(Gudhi)

Data	Pnt	Thrsld	VertexCollapser +PD					
			Size/Dim	*dim*	Pre-Time	Tot-Time	*Step*	Snaps
netw-sc	379	5.5	175/3	∞	366.46	366.56	0	8420
senate	103	0.415	417/4	∞	15.96	15.98	0	2728
eleg	297	0.3	835K/16	∞	518.36	540.40	0	9850
HIV	1088	1050	127.3M/?	4	660	3,955	4	184
torus	2000	1.5		4	∞*	∞	0	428K

Thank You!