Foundations of Geometric Methods in Data Science

Jean-Daniel Boissonnat
Mathieu Carrière
Frédéric Cazals
INRIA

Master Science des Données et Intelligence Artificielle
2023

Lecture 1 : Elements of Computational Geometry and Topology

Motivation for Geometric Data Analysis

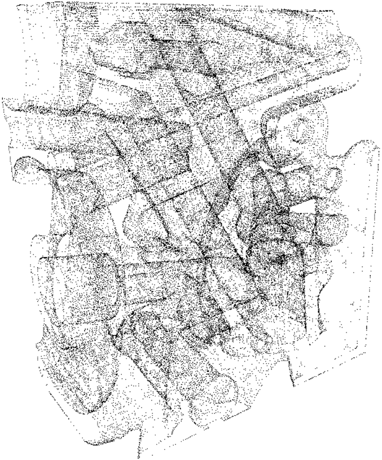
Combinatorial models

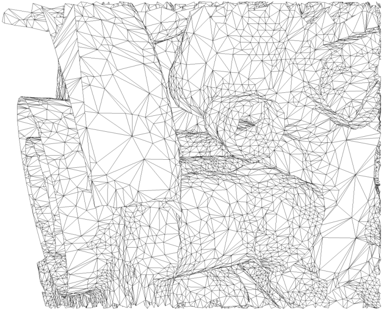
Delaunay complexes

Nets, sampling and clustering

Data structures

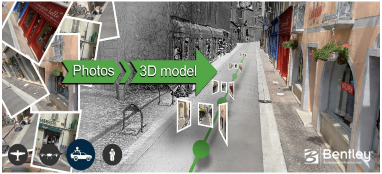
Reconstructing surfaces from point clouds

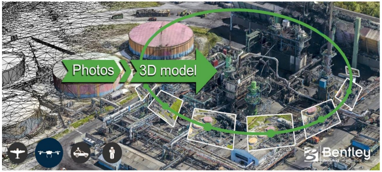




One can reconstruct a surface from 10^6 points within 1mn [CGAL]

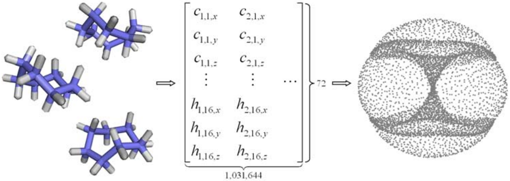
3D Reconstruction from images





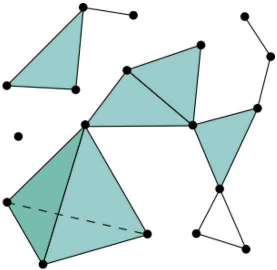
Acute3D, Bentley Systems

Geometric data analysis



Geometrisation :	Data = points + distances between points
Manifold Hypothesis :	Data lie close to a structure of “small” intrinsic dimension
Problem :	Infer the structure from the data

Simplicial complexes



H. Poincaré (1854-1912)

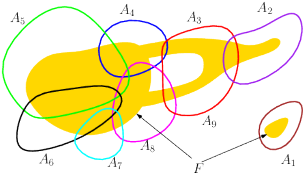
Let V be a finite set. A simplicial complex (abstract) on V is a finite set of subsets of V called the simplexes or faces of K that satisfy :

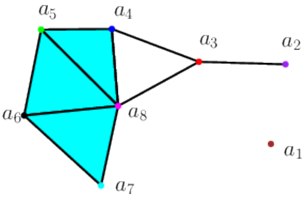
1. The elements of V belong to K (vertices)

2. If $\tau \in K$ and $\sigma \subseteq \tau$, then $\sigma \in K$ $(\dim(\sigma) \stackrel{\text{def}}{=} |\sigma| - 1)$

Nerve of a good cover

A simplicial complex to represent the topology of an object



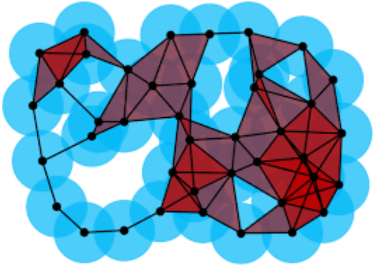


Nerve theorem (J. Leray, 1945)

If the intersection of any subset of elements in the cover is contractible, then the nerve and the union of the elements of the cover have the same homotopy type.

Čech complex
Nerve of a set of balls

A finite set of points $\mathcal{P} \in \mathbb{R}^d$



J. Leray
(1906-1998)

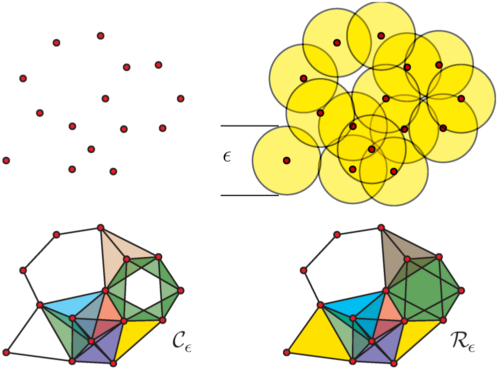
Corollary of the nerve theorem (J. Leray, 1945)

The Čech complex has the same homotopy type as the union of balls.

Čech and Rips-Vietoris complexes

$$\sigma \subseteq \mathcal{P} \in CR(\mathcal{P}, \alpha) \quad \Leftrightarrow \quad \bigcap_{p \in \sigma} B(p, \alpha) \neq \emptyset$$

$$\sigma \subseteq \mathcal{P} \in R(\mathcal{P}, \alpha) \quad \Leftrightarrow \quad \forall p, q \in \sigma \; \|p - q\| \leq 2\alpha \quad \Leftrightarrow \quad B(p, \alpha) \cap B(q, \alpha) \neq \emptyset$$



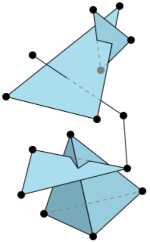
Interleaving : $R(\mathcal{P}, \frac{\alpha}{2}) \subseteq C(\mathcal{P}, \alpha) \subseteq R(\mathcal{P}, \alpha)$

Geometric simplicial complexes

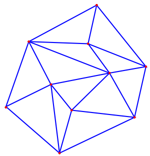
Geometric simplex of dimension k : the convex hull of $k+1$ (independent) points

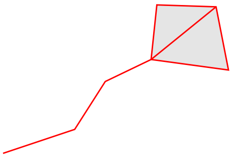
Geometric simplicial complex :

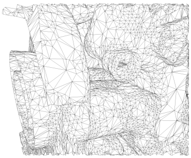
A finite collection of geometric simplices K called the faces of K such that



- $\forall \sigma \in K$, σ is a simplex
- $\sigma \in K, \tau \subset \sigma \Rightarrow \tau \in K$
- $\forall \sigma, \tau \in K$, either $\sigma \cap \tau = \emptyset$ or $\sigma \cap \tau$ is a common face of both







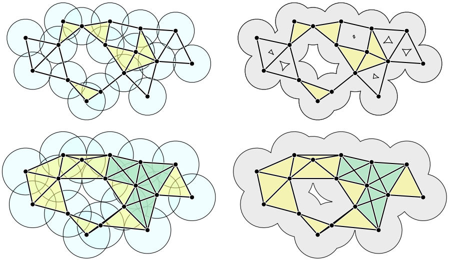
Filtration of a simplicial complex

A filtration of K is a sequence of nested subcomplexes of K

$$\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$$

such that: $K^{i+1} = K^i \cup \sigma^{i+1}$, where σ^{i+1} is a simplex of K

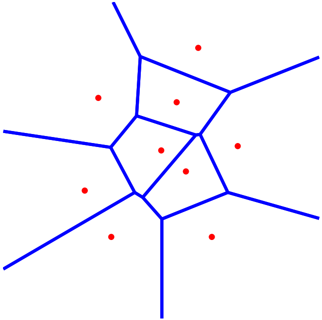
Example : Čech filtration



Filtrations play a central role in topological persistence

Voronoi diagrams

A set of points $\mathcal{P}$ in $(\mathbb{R}^d, \|.\|)$

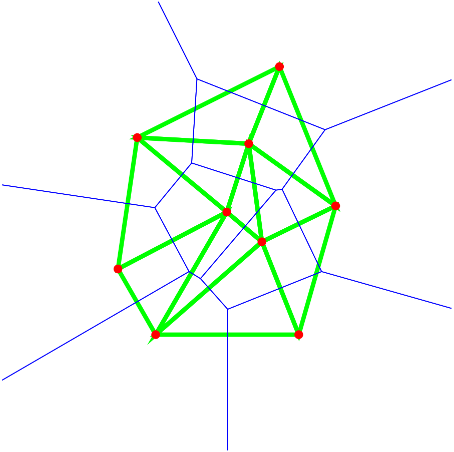


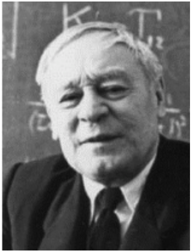
Voronoi cell $\quad V(p_i) = \{x : \|x - p_i\| \leq \|x - p_j\|, \ \forall j\}$

Voronoi diagram $\quad \mathrm{Vor}(\mathcal{P}) = \{\text{set of cells } V(p_i), p_i \in \mathcal{P}\}$

Delaunay Triangulations

Sur la sphère vide (On the empty sphere), Boris Delaunay (1934)

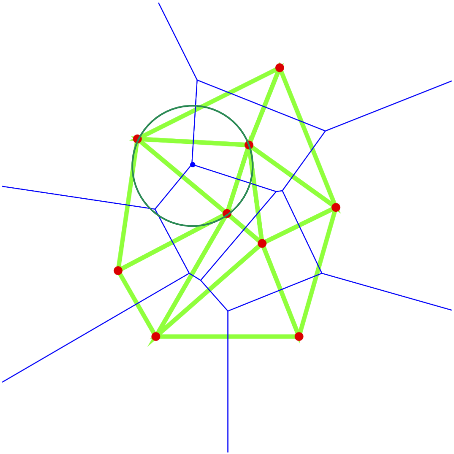


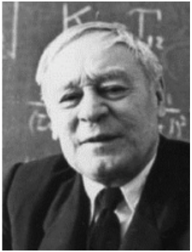


The Delaunay complex $\mathrm{Del}(\mathcal{P})$ is the nerve of $\mathrm{Vor}(\mathcal{P})$

Delaunay Triangulations

Sur la sphère vide (On the empty sphere), Boris Delaunay (1934)

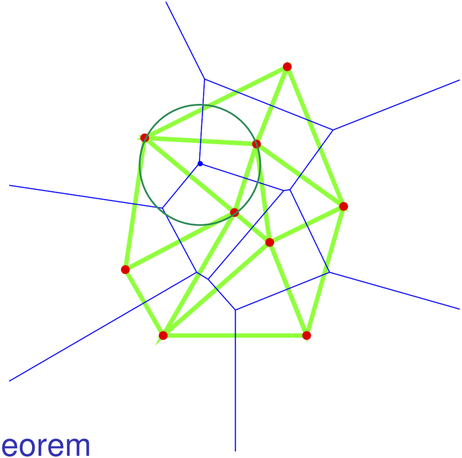


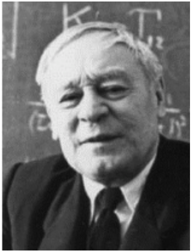


The Delaunay complex $\mathrm{Del}(\mathcal{P})$ is the nerve of $\mathrm{Vor}(\mathcal{P})$

Delaunay Triangulations

Sur la sphère vide (On the empty sphere), Boris Delaunay (1934)





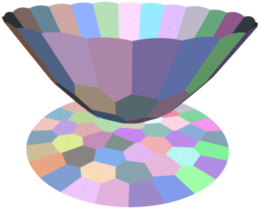
The Delaunay complex $\mathrm{Del}(\mathcal{P})$ is the nerve of $\mathrm{Vor}(\mathcal{P})$

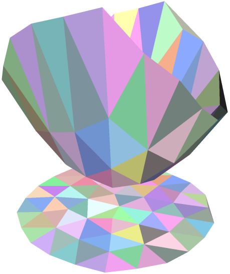
Theorem

If $\mathcal{P}$ contains no subset of $d+2$ points on a same hypersphere, then $\mathrm{Del}(\mathcal{P})$ is a triangulation of $\mathcal{P}$

Correspondence between structures

$h_{p_i} : x_{d+1} = 2p_i \cdot x - p_i^2$ $\qquad$ $\hat{p}_i = (p_i, p_i^2) = h^*_{p_i}$





$$\begin{array}{ccc} \mathcal{V}(\mathcal{P}) = h^+_{p_1} \cap \ldots \cap h^+_{p_n} & \xrightarrow{\text{duality}} & \mathcal{D}(\mathcal{P}) = \text{conv}^-(\{\hat{p}_1, \ldots, \hat{p}_n\}) \\ \uparrow & & \downarrow \\ \text{Voronoi diagram of } \mathcal{P} & \xrightarrow{\text{nerve}} & \text{Delaunay triang. of } \mathcal{P} \end{array}$$

The diagram commutes if $\mathcal{P}$ is in general position wrt spheres

Corollaries

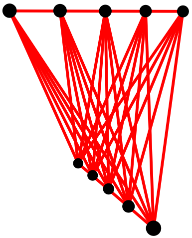
Combinatorial complexity

The Voronoi diagram of n points of $\mathbb{R}^d$ has the same combinatorial complexity as the intersection of n half-spaces of $\mathbb{R}^{d+1}$

The Delaunay triangulation of n points of $\mathbb{R}^d$ has the same combinatorial complexity as the convex hull of n points of $\mathbb{R}^{d+1}$

The two complexities are the same (duality): $\Theta(n^{\lceil \frac{d}{2} \rceil})$ [Mc Mullen 1970]

Worst-case: points on the moment curve $\Gamma(t) = \{t, t^2, ..., t^d\} \subset \mathbb{R}^d$



Quadratic in $\mathbb{R}^3$

Corollaries

Algorithmic complexity

Construction of $\mathrm{Del}(\mathcal{P})$, $\mathcal{P} = \{p_1, ..., p_n\} \subset \mathbb{R}^d$

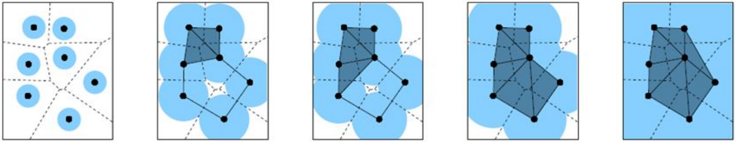
1. Lift the points of $\mathcal{P}$ onto the paraboloid $x_{d+1} = x^2$ of $\mathbb{R}^{d+1}$: $p_i \to \hat{p}_i = (p_i, p_i^2)$
2. Compute $\mathrm{conv}(\{\hat{p}_i\})$
3. Project the lower hull $\mathrm{conv}^-(\{\hat{p}_i\})$ onto $\mathbb{R}^d$

Complexity : $\Theta(n \log n + n^{\lceil \frac{d}{2} \rceil})$ [Clarkson & Shor 1989] [Chazelle 1993]

Alpha-shapes and the Delaunay filtration

Let $U(\alpha)$ be the union of the balls $B(p, \alpha)$, $p \in P$.

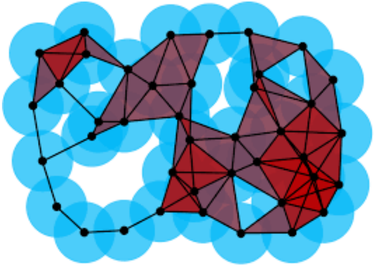
The alpha-shape of P, noted alpha(P), is the nerve of the restriction of $\mathrm{Del}(P)$ to $U(\alpha)$.



The alpha-shape is a deformation retract of the union of balls

The Delaunay filtration is the nested sequence of alpha(P) for $\alpha \in [0, \infty]$

Čech complex versus alpha-shape



- Both complexes are homotopy equivalent to $U(\alpha)$
- The size of $\check{\mathrm{C}}\mathrm{ech}(P, \alpha)$ is $\Theta(n^d)$
- The size of the alpha-shape(P) is $\Theta(n^{\lceil \frac{d}{2} \rceil})$
- the alpha-shape naturally embeds in $\mathbb{R}^d$ but not $\check{\mathrm{C}}\mathrm{ech}(B)$ (general position)

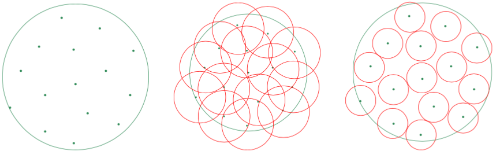
Definition and existence of nets

Definition

Let Ω be a bounded subset of $\mathbb{R}^d$. A finite set of points P is called an $(\varepsilon, \bar{\eta})$-net of Ω iff

Covering/density : $\forall x \in \Omega, \exists p \in P : \|x - p\| \leq \varepsilon$

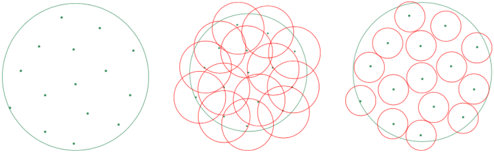
Packing/separation : $\forall p, q \in P : \|p - q\| \geq \bar{\eta}\,\varepsilon \stackrel{\text{def}}{=} \eta$



Lemma Ω admits an $(\varepsilon, 1)$-net.

Proof. While there exists a point $p \in \Omega$, $d(p, P) \geq \varepsilon$, insert p in P

Size of a net inside a unit ball Ω of $\mathbb{R}^d$



Lemma The number of points of an $(\varepsilon, \bar{\eta})$-net of Ω is at most

$$n(\varepsilon, \bar{\eta}) \leq \frac{\mathrm{vol}_d(B(1 + \frac{\eta}{2}))}{\mathrm{vol}_d(B(\frac{\eta}{2}))} = O\left(\frac{1}{\varepsilon^d}\right)$$

where the constant in the O depends on $\bar{\eta}^d$.

Size of the Delaunay complex of a net

Lemma Let Ω be a unit ball of $\mathbb{T}^d$, P an $(\varepsilon, \bar{\eta})$-net of Ω (of size $n = |P| = O(\frac{1}{\varepsilon^d})$) and assume that d and $\bar{\eta}$ are positive constants.

The Delaunay triangulation of P to Ω has linear size $O(n)$.

Proof.

1. The Delaunay neighbours of a point p are at distance $\leq 2\varepsilon$ from p
2. There number is $n_p = O(1)$ using a volume argument
3. The number of simplices incident on p is at most

$$\sum_{i=1}^{d+1} \binom{n_p}{i} \leq \sum_{i=0}^{n_p} \binom{n_p}{i} = 2^{n_p}.$$

Two problems about nets in discrete metric spaces

Input: a finite point set W. We know the distances between any 2 points but not the locations of the points.

Problems: Can we extract from W a subsample L such that

Subsampling : L is a $(\lambda, 1)$-net of W (assuming W is λ dense)

Clustering : $|L| = k$ and $\max_{w \in W} d(w, L)$ is minimized.

Farthest point insertion

Input: the distance matrix of a finite point set W and either a positive constant λ (Case 1) or an integer k (Case 2)

1. $L := \{w_1\}$ — initialize L sample with any point of W
2. $L(w) := w_1$ for all $w \in W$ — $L(w)$ stores the element of L closest to w
3. $\lambda^* := \max_{w \in W} \|w - L(w)\|$
4. $w^* :=$ a point $p \in W$ such that $\|p - L(p)\| = \lambda^*$ — point of W most distant from L
5. **while** either $\lambda^* > \lambda$ (Case 1) or $|L| < k$ (Case 2)

    5.1 add w^* to L
    5.2 **for** each point w of W such that $\|w - w^*\| < \|w - L(w)\|$ **do**
        5.2.1 $L(w) := w^*$
        5.2.2 update λ^* and w^*

6. **return :** $L \subseteq W$

Property : Case 1 : L is $(\lambda, 1)$-net of W
Case 2 : L is an approximate solution to the k-centers problem

Time complexity : $O(kn)$

Constructing nets by subsampling

Lemma 1 Let W be a finite set of points such that the distance of any point $q \in W$ to $W \setminus \{q\}$ is at most ε and let $\lambda \geq \varepsilon$. One can extract from W a subsample L that is a $(\lambda, 1)$-net of W.

Proof

For any $i > 0$, $L_i = \{l_1, ..., l_i\}$ and $\lambda_i = d(l_i, L_{i-1})$ (l_i indexed by insertion order)

Since L_i grows with i : $j \geq i \;\Rightarrow\; \lambda_j \leq \lambda_i$ (*)

We claim that at each iteration $i > 0$, L_i is a $(\lambda_i, 1)$-net of W.

1. L_i is λ_i-dense in W by (*)
2. L_i is λ_i-separated: $l_a l_b$ closest par in L_i, l_b (inserted after l_a)

$$\Rightarrow \quad \|l_a - l_b\| = \lambda_b \geq \lambda_i \qquad \text{by (*)}$$

k-centers clustering

Problem : Select from W a subset L of k points so as to maximize the minimum pairwise distance between the points of L.

Lemma 2 The farthest insertion algorithm (Case 2) provides a 2-approximation to the k-centers problem and to the k-centers clustering problem.

Proof

- $W \subset \cup_{i=1}^{k-1} B(l_i, \lambda_k)$

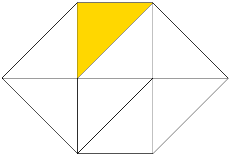
  $\Rightarrow$ Two points of L_{opt} lie in the same ball $B(l_i, \lambda_k)$, for some $i \leq k-1$

  $\Rightarrow \exists p, q \in L_{\mathrm{opt}}$ s.t. $\|p-q\| \leq 2\lambda_k$

- The distance between any two points of L_k is at least λ_k (Lemma 1).

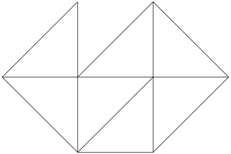
  $$\Rightarrow \quad \frac{1}{2}\,\mathrm{maxminPD}(L_{\mathrm{opt}}) \leq \lambda_k \leq \mathrm{maxminPD}(L_k)$$

Data structures to represent simplicial complexes



Atomic operations

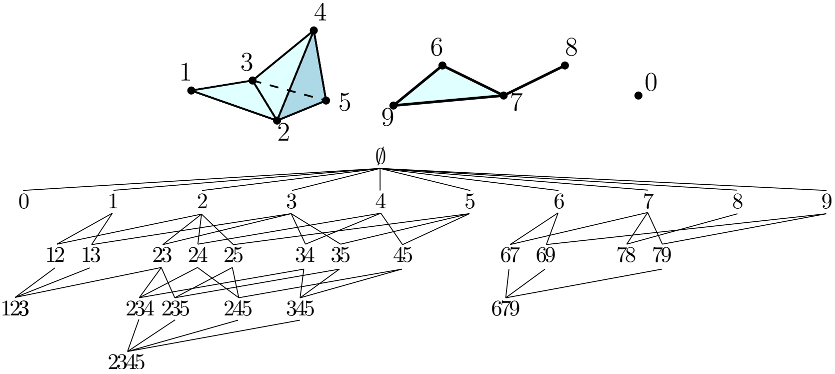
- **Look-up/Insertion/Deletion** of a simplex
- **Facets** and **subfaces** of a simplex
- **Cofaces**, **link** of a simplex
- **Topology preserving** operations
  - Edge contractions
  - Elementary collapses



Explicit representation of all simplices ? of all incidence relations ?

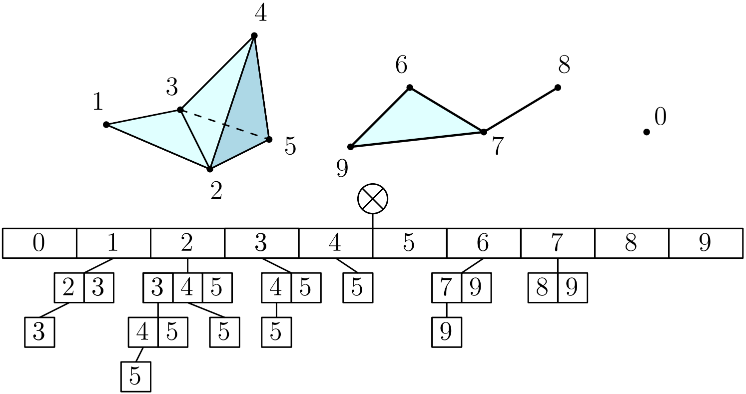
The Hasse diagram

$$
\begin{array}{llll}
G(V,E) & \sigma \in V & \Leftrightarrow & \sigma \in K \\
 & (\sigma,\tau) \in E & \Leftrightarrow & \sigma \subset \tau \ \wedge \ \dim(\sigma) = \dim(\tau) - 1
\end{array}
$$



The simplex tree is a prefix tree (trie)

1. index the vertices of K
2. associate to each simplex $\sigma \in K$, the sorted list of its vertices
3. store the simplices in a trie.

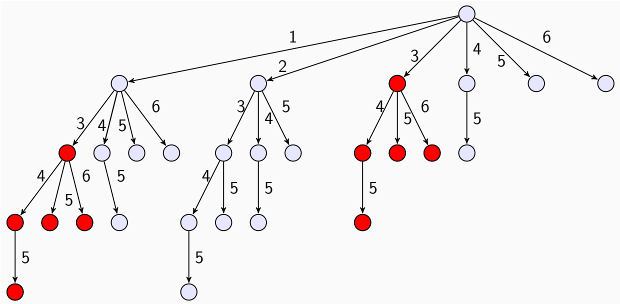


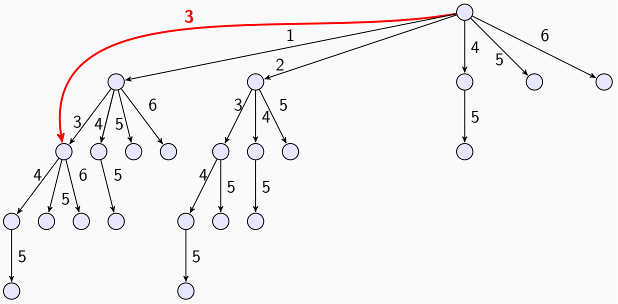
Performance of the simplex tree

- A subgraph of the Hasse diagram
- Explicit representation of all simplices
- $\#$nodes $=$ $\#\mathcal{K}$
- depth $=$ $\dim(\mathcal{K}) + 1$
- $\#$children(σ) $\leq$ $\#$cofaces(σ) $\leq$ deg(last(σ))
- Memory complexity: $O(1)$ per simplex
- Basic operations
  - Membership (σ) : $O(d_\sigma \log n)$
  - Insertion (σ) : $O(2^{d_\sigma} d_\sigma \log n)$

Implemented in the GUDHI library

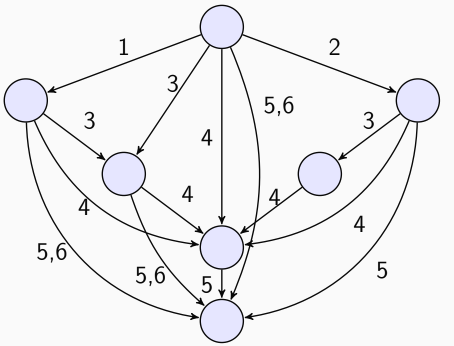
Redundancy in the Simplex Tree





Minimal simplex automaton

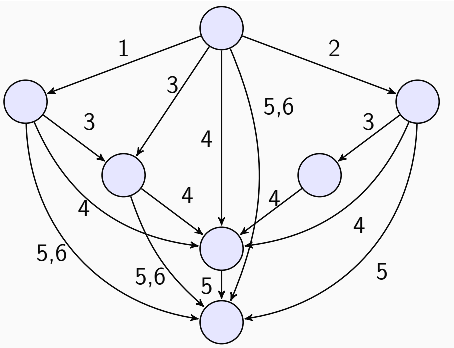
B., Karthik, Tavenas 2016



- Compression time : $O(m \log m \log n)$ [Hopcroft 1971]
- Static queries: unchanged
- Dynamic queries: more complex

Minimal simplex automaton

B., Karthik, Tavenas 2016



- Compression time : $O(m \log m \log n)$ [Hopcroft 1971]
- Static queries: unchanged
- Dynamic queries: more complex

- The size of the automaton depends on the labelling of the vertices

  Finding an optimal labelling is NP-complete

Experiments

Data Set 1: Rips Complex from sampling of Klein bottle in $\mathbb{R}^5$.

n	α	d	k	m	Size After Compression	Compression Ratio
10,000	0.15	10	24,970	604,573	218,452	2.77
10,000	0.16	13	25,410	1,387,023	292,974	4.73
10,000	0.17	15	27,086	3,543,583	400,426	8.85
10,000	0.18	17	27,286	10,508,486	524,730	20.03

Experiments

Data Set 1: Rips Complex from sampling of Klein bottle in $\mathbb{R}^5$.

n	α	d	k	m	Size After Compression	Compression Ratio
10,000	0.15	10	24,970	604,573	218,452	2.77
10,000	0.16	13	25,410	1,387,023	292,974	4.73
10,000	0.17	15	27,086	3,543,583	400,426	8.85
10,000	0.18	17	27,286	10,508,486	524,730	20.03

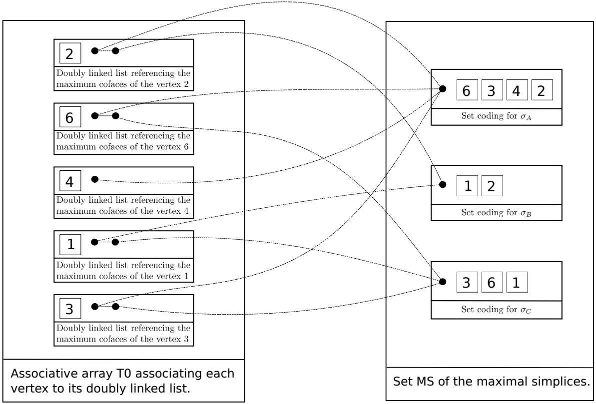
Data Set 2: Flag complexes generated from random graph $G_{n,p}$.

n	p	d	k	m	Size After Compression	Compression Ratio
25	0.8	17	77	315,370	467	537.3
30	0.75	18	83	4,438,559	627	7,079.0
35	0.7	17	181	3,841,591	779	4,931.4
40	0.6	19	204	9,471,220	896	10,570.6
50	0.5	20	306	25,784,504	1,163	22,170.7

Simplex Array List

[B., Karthik C.S., Tavenas 2017]

Store only the maximal simplices



Memory storage : $O\left(\sum_{\sigma \in K} d_\sigma\right) = O(kd)$ Optimal

Proof of optimality

Theorem

Consider the class of all simplicial complexes $\mathcal{K}(n, k, d)$ where $d \geq 2$ and $k \geq n + 1$.

Any data structure that can represent the simplicial complexes of this class requires $\log \binom{\binom{n/2}{d+1}}{k-n}$ bits to be stored,

which is $\Omega(kd \log n)$ for any constant $\varepsilon \in (0, 1)$ and for $\frac{2}{\varepsilon} n \leq k \leq n^{(1-\varepsilon)d}$ and $d \leq n^{\varepsilon/3}$.

Proof $\mathcal{P} = |\mathrm{vert}(K)|$, $\mathcal{P}' \subset \mathcal{P}$, $|\mathcal{P}'| = n/2$

Consider the set S of all simplicial complexes with vertex set $\subset \mathcal{P}'$, of dimension d and having $k - n$ maximal simplices (all of dimension d) and observe that $|S| = \binom{\binom{n/2}{d+1}}{k-n}$

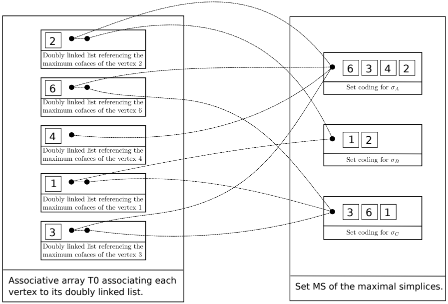
Let $K_1, ..., K_{|S|}$ be those complexes with vertex sets $\mathcal{P}_1, ..., \mathcal{P}_{|S|}$

Complete each K_i with vertices in $\mathcal{P} \setminus \mathcal{P}_i$ and edges spanning those vertices so that K_i^+ has n vertices and k maximal simplices (of dimension 1 or h)

We have $|S|$ complexes of $\mathcal{K}(n, k, d, m)$

Basic operations

Complexity depends on a local parameter



$\Gamma_i(\sigma)$ = number of maximal cofaces of σ of dimension i

$\Gamma_i = \max_{\sigma \in K} \Gamma_i(\sigma)$

Membership (σ): $O\left(\sum_{i=0}^{d_\sigma - 1} \Gamma_i(\sigma) \log n\right) = O(\Gamma_0 d \log n)$ ST : $O(d \log n)$

Insertion (σ) : $O(\Gamma_0(\sigma) d_\sigma^2 \log n) = O(\Gamma_0 d^2)$ ST : $O(d_\sigma 2^{d_\sigma} \log$

Experimental results

Data Set 1 (Rips complex on a Klein bottle in $\mathbb{R}^5$)

No	n	α	d	k	m	Γ_0	Γ_1	Γ_2	Γ_3	$\|SAL\|$
1	10,000	0.15	10	24,970	604,573	62	53	47	37	424,440
2	10,000	0.16	13	25,410	1,387,023	71	61	55	48	623,238
3	10,000	0.17	15	27,086	3,543,583	90	67	61	51	968,766
4	10,000	0.18	17	27,286	10,508,486	115	91	68	54	1,412,310

To be released in the GUDHI library (F. Godi)

Conclusions

Next lectures

- Other types of simplicial complexes
- Triangulation of manifolds

Open questions

- Bound on Γ_0 for interesting simplicial complexes
- Lower bounds on query time assuming optimal storage $O(kd \log n)$