LITE: A Framework for Lattice-Integrated Embedding of Topological Descriptors

Michael Etienne Van Huffel, Matteo Palo

13 March 2024

ETH Zurich, Department of Mathematics

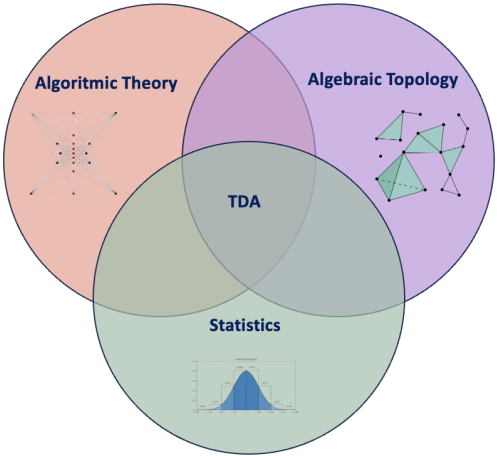
Unlocking Meaning through Shapes

"Data has a shape and shape has a meaning."
Prof. Gunnar Carlsson

- Topology focuses on properties invariant to changes in connectivity and other topological features, such as loops, holes, and more, offering powerful tools for characterizing and analyzing spaces and functions.
- In many practical applications, we do not directly observe the underlying manifold (space) of the data; instead, we observe the data as point clouds.

How can we harness the tools of combinatorial geometry, to unveil and analyze hidden continuous structures?

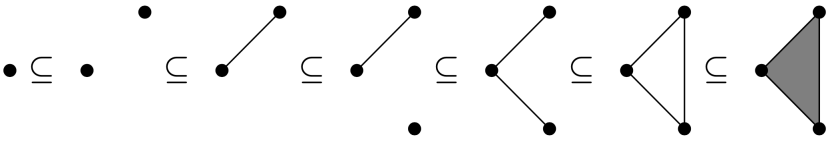
Topological Data Analysis: Bridging Interdisciplinary Boundaries



Filtration of simplicial complexes

Simplicial filtration: nested sequence of subcomplexes

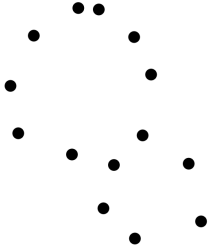
$$\mathcal{F} : \emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq \ldots \subseteq K_n = K$$



Example

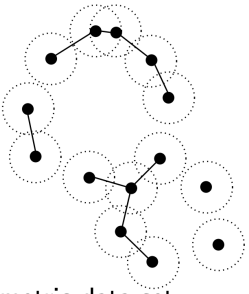
Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space. For $r \in \mathbb{R}^+$, the Vietoris-Rips filtration is the filtered simplicial complex that contains a simplex σ if $d_{\mathbb{X}}(p, q) \leq 2 \cdot r, \forall p, q \in \sigma$.

Applying Persistent Homology to Point Clouds



$\mathbb{X}$: metric data set

Applying Persistent Homology to Point Clouds

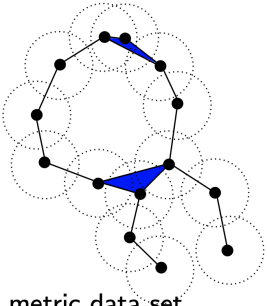


$\mathbb{X}$: metric data set

↳ $\mathrm{Filt}(\mathbb{X})$: filtered simplicial complex

Persistent homology encodes the evolution of the topology across scales.

Applying Persistent Homology to Point Clouds

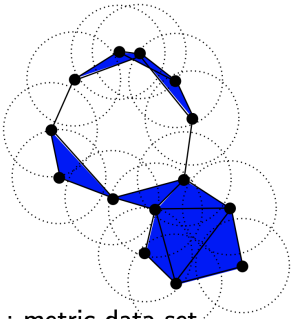


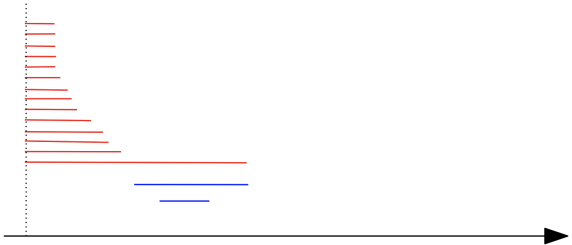
$\mathbb{X}$: metric data set

$\mathrm{Filt}(\mathbb{X})$: filtered simplicial complex

Persistent homology encodes the evolution of the topology across scales.

Applying Persistent Homology to Point Clouds



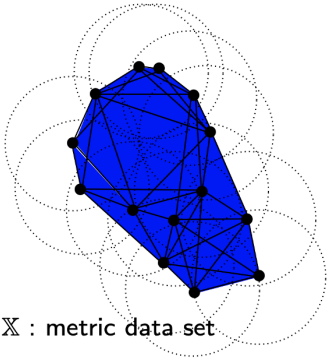


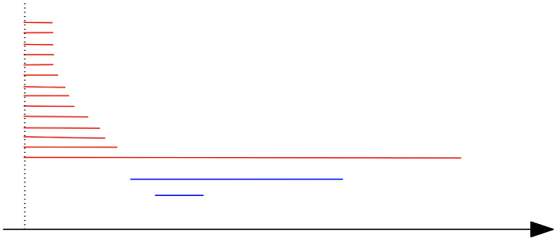
$\mathbb{X}$: metric data set

$\mathrm{Filt}(\mathbb{X})$: filtered simplicial complex

Persistent homology encodes the evolution of the topology across scales.

Applying Persistent Homology to Point Clouds



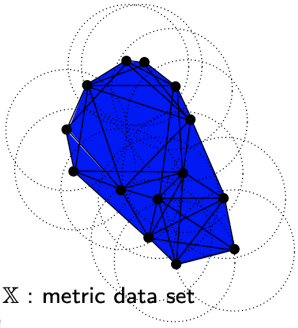


$\mathbb{X}$: metric data set

$\mathrm{Filt}(\mathbb{X})$: filtered simplicial complex

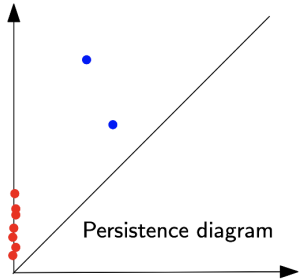
Persistent homology encodes the evolution of the topology across scales.

Applying Persistent Homology to Point Clouds



$\mathbb{X}$: metric data set

$\mathrm{Filt}(\mathbb{X})$: filtered simplicial complex



A Persistence Diagram (PD) D is a locally finite multiset of points in the half-plane $\Omega = \{(x, y) \in \mathbb{R}^2 \mid x < y\}$ together with points on the diagonal $\partial\Omega = \{(x, x) \in \mathbb{R}^2\}$ counted with infinite multiplicity.

Related Work and Contribution

- PDs are not a Hilbert space (*Bubenik et al.,(2018), Mileyko et al., (2011)*)!
- Two main methods have been proposed to address the unstructured nature of Persistence Diagrams spaces:
  - Vectorization methods: e.g., Persistence Images (*Adams et al. (2017)*), Persistence Landscapes (*Bubenik et al. (2015)*), and modern techniques like ATOL(*Royer et al. (2019)*) and PersLay (*Chazal et al. (2019)*).
  - Kernel-based approach: e.g., multi-scale (*Reininghaus et al. (2014)*), weighted Gaussian (*Kusano et al. (2016)*), and sliced Wasserstein kernels (*Carriere et al. (2017)*).
- We introduce **LITE (Lattice-Integrated Embedding of Topological Descriptors)**, a new vectorization framework that:
  - Embed PDs by first computing their induced Persistence Measures and then computing an integral transform.
  - Achieves, yet being very simple, results comparable to or surpassing those in the TDA literature on classical classification benchmark tasks (graphs/dynamical particles) .

Methodology I: Persistence Measures

- Given a PD, we apply the transform $\tau : (x, y) \mapsto (x, y - x)$ (*H. Adams et al., (2017)*).
- Following *Divol, Chazal (2019)*, we define PDs as measures on Ω by $\sum_{x \in D \cap \Omega} m_x \delta_x$ where x ranges all off-diagonal points in a persistence diagram D, m_x is the multiplicity of x, and δ_x is the Dirac measure at x.
- **Here:** instead of multiset of points of a PD D, take a lattice as support for the measure!

Methodology II: Integral transforms on Measures

- Given a Persistence Measure μ, we then take a integral transform for some function f:

$$\Psi_\mu(f) := \int_\Omega f(\mathbf{x}, \cdot) d\mu(\mathbf{x}) = \sum_{\mathbf{x} \in D} m(\mathbf{x}) f(\mathbf{x}, \cdot)$$

- **Here:** frequency based transforms: Fourier, Wavelet (db, coif) and Gabor.

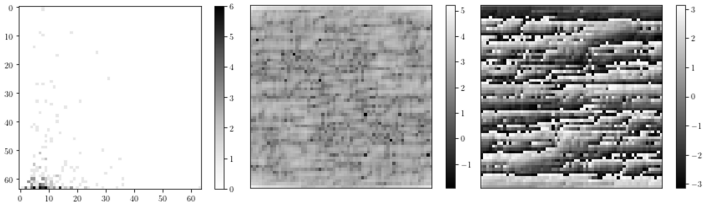


Figure 1: **Left:** Persistence Measure for H_1 Homology of a point cloud sampled from $\mathcal{S}^2$. **Middle:** Log Magnitude of the Fourier Transform of the Persistence Measure. **Right:** Angle of the Fourier Transform of the Persistence Measure.

Experimental Details

- Performance is assessed through ten 10-fold evaluations on each dataset, reporting both average and best 10-fold results.
- For graph datasets, we adopt the methodology from Atol (*Royer et al. (2021)*) for generating extended PDs, using two specific HKS diffusion times ($t_1 = 0.1$ and $t_2 = 10$).
  - For evaluation, we use three distinct square grid sizes (20×20, 32×32, and 50×50) across all proposed transforms for both Biomedical and Social Networking graph problems.
- For the `Orbit` learning task, we adopt a regular square grid of 64×64 and 128×128 for all transforms.

Graph classification Datasets

- Each vertex is assigned a filtration value using **Heat Kernel Signature**.
- Persistent features summarized using *Extended Persistence* (*Carriere et al.*).

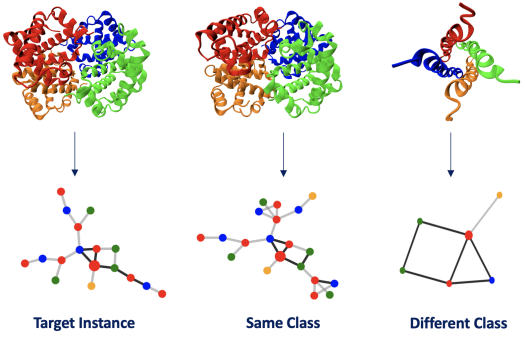


Figure 2: **Example:** Protein Classification Task.

Performance Evaluation on different Graph Datasets

- In all considered benchmark datasets, our method is competitive and some time outperforms current state-of-the-art "classic" vectorization methods.

Dataset	SV†	P†	MP†	Perslay⋆	ATOL⋆	BBA†	LITE (Our)		LITE-IdT (Our)	
							Mean⋆	Max†	Mean⋆	Max†
MUTAG	88.3	79.2	86.1	89.8	88.3	90.4	89.8	91.7	89.2	90.7
COX2	78.4	76.0	79.9	80.9	79.4	81.2	80.6	82.4	79.4	80.4
PROTEINS	72.6	65.4	67.5	74.8	71.4	74.7	72.8	73.6	72.2	73.2
DHFR	78.4	70.9	81.7	80.3	82.7	80.5	81.8	83.1	81.2	82.7
IMDB-B	72.9	54.0	68.7	71.2	74.8	69.4	68.4	69.8	67.2	68.3
IMDB-M	50.3	36.3	46.9	48.8	47.8	46.7	43.7	44.4	43.1	44.3

Table 1: Comparative Analysis of Classification Accuracy with topological methods on Benchmark Graph Datasets. Note: Symbol † compare with *Max* metric, while ⋆ with *Mean* due to different experimental setup.

The Orbit5K Dataset

- This dataset consists of subsets of size 1000 of the unit cube $[0,1]^2$ generated by a dynamical system that depends on a parameter $\rho > 0$.
- To generate a point cloud, a random initial point (x_0, y_0) is chosen uniformly in $[0,1]^2$ and a sequence of points (x_n, y_n) for $n = 0, 1, \ldots, 999$ is generated recursively by:

$$\begin{aligned} x_{n+1} &= x_n + \rho y_n (1 - y_n) && \mod 1 \\ y_{n+1} &= y_n + \rho x_{n+1} (1 - x_{n+1}) && \mod 1. \end{aligned}$$

- Given an orbit, we want to predict the value of ρ, that can take values in $\{2.5, 3.5, 4.0, 4.1, 4.3\}$.

Performance Evaluation on the Orbit5K Dataset

PSS-K	PWG-K	SW-K	PF-K	PI	Perslay	BBA	LITE (Our)	LITE-IdT (Our)
72.38	76.63	83.6	85.9	82.5	87.7	83.3	84.6	82.0

Table 2: Comparative Classification Accuracy on the Orbit5K Dataset.

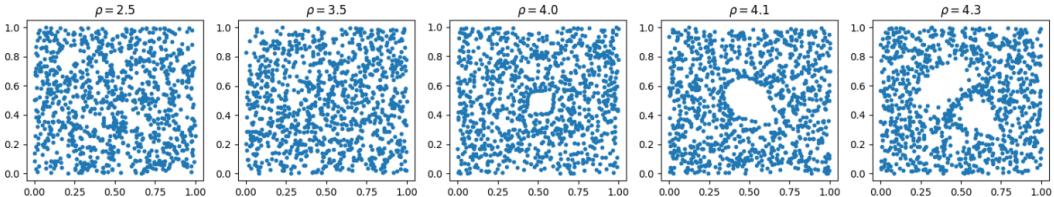


Figure 3: **Above:** Comparison of results on the Orbit5K Dataset. **Below:** Examples of point clouds from the Orbit5K dataset.

Further Work

- Optimize Choice of Grid Structure (Regular Grid): balance computational time against potential information loss.
- Interpretability of the embedding?
- Apply LITE to other supervised/unsupervised classification tasks.

Topological Descriptors on Graphs I

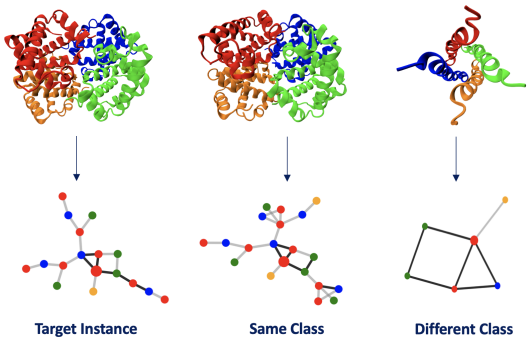
For $G = (V, E)$ with adjacency matrix A and degree matrix D, the normalized Laplacian is

$$L_w(G) = I - D^{-1/2} A D^{-1/2},$$

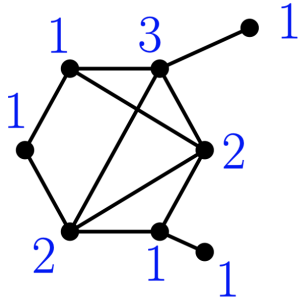
decomposing on orthonormal basis $\phi_1, \ldots, \phi_n$ with eigenvalues $0 \leq \lambda_1 \leq \cdots \leq \lambda_n \leq 2$.

Heat Kernel Signature ($t \geq 0$):

$$\mathrm{hks}_{G,t} : v \mapsto \sum_{k=1}^{n} \exp(-\lambda_k t) \phi_k(v)^2 .$$



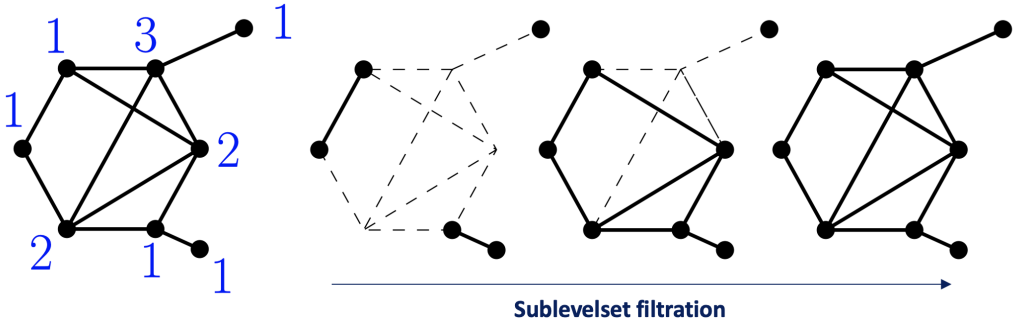
Topological Descriptors on Graphs II



Key Aspect

Monitor the birth and death of each topological event (e.g., creation/merging of connected components, formation of loops, ...)

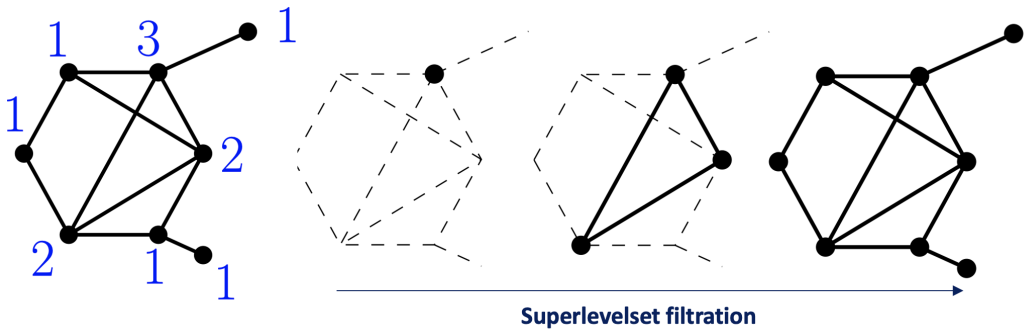
Topological Descriptors on Graphs II



Key Aspect

Monitor the birth and death of each topological event (e.g., creation/merging of connected components, formation of loops, ...)

Topological Descriptors on Graphs II



Key Aspect

Monitor the birth and death of each topological event (e.g., creation/merging of connected components, formation of loops, ...)

Graph Datasets Miscellaneous

Datasets used for the experiment are very popular for graph classification, including also 'Proteins' and 'IMDB'.

Figure 4: IMDB

IMDB, movie collaboration dataset, networks of 1,000 actors who played roles in movies in IMDB. In each graph, nodes represent actors/actress, and there is an edge between them if they appear in the same movie.

Figure 5: PROTEINS

PROTEINS is a dataset of proteins that are classified as enzymes or non-enzymes. Nodes represent the amino acids and two nodes are connected by an edge if they are less than 6 Angstroms apart.