Topological Data Analysis with ϵ-net Induced Lazy Witness Complex

Naheed Anjum Arafat[1], Debabrota Basu[2], Stéphane Bressan[1]

[1]National University of Singapore
[2]Chalmers University of Technology

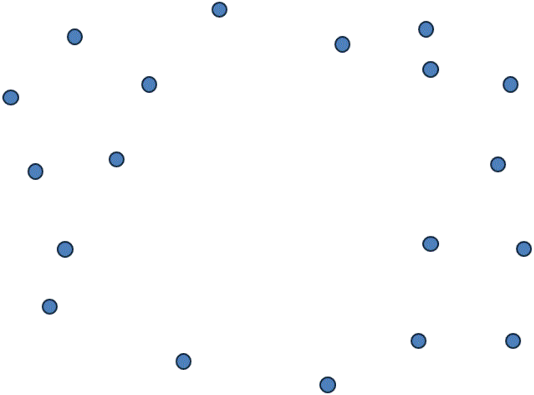
November 25, 2020

What is this Talk About?

Approximating persistent topological features
from point-clouds
via (better) sampling

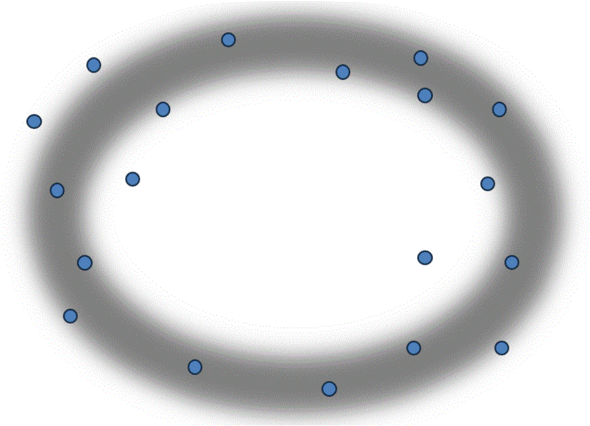
Topological features

A point-cloud



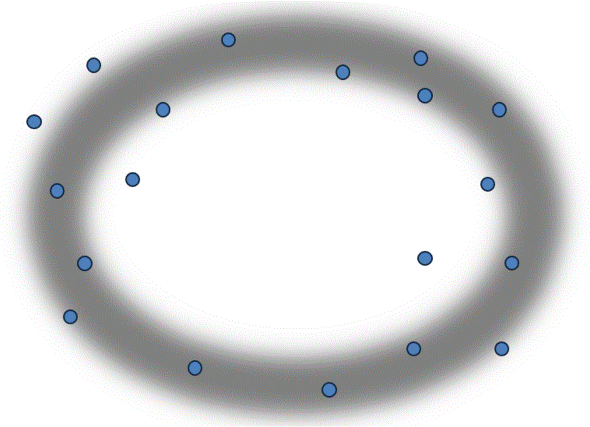
Topological features

The underlying space: A Ring in $\mathbb{R}^2$



Topological features

1 Connected component (Top. feat. at dim. 0)
1 cycle, Inner-cycle $\sim$ Outer-cycle (Top. feat. at dim. 1)
0 void (Top. feat. at dim. 2)

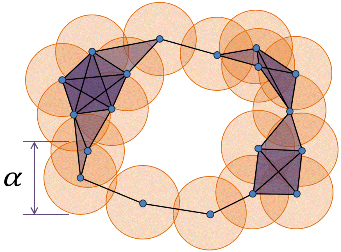


Formal Representation: Simplicial complex

- Simplicial Complex: A set of simplices (0-simplex: A point, 1-simplex: Edge, 2-simplex: Filled triangle)

Formal Representation: Simplicial complex

- Simplicial Complex: A set of simplices (0-simplex: A point, 1-simplex: Edge, 2-simplex: Filled triangle)
- Choose a threshold α.
  - Draw diameter α-balls around each **point**.
  - Connect two points with an **edge** if their corresponding balls intersect pairwise.
  - Connect three points with a **filled triangle** if their corresponding balls intersect pairwise. And so on.



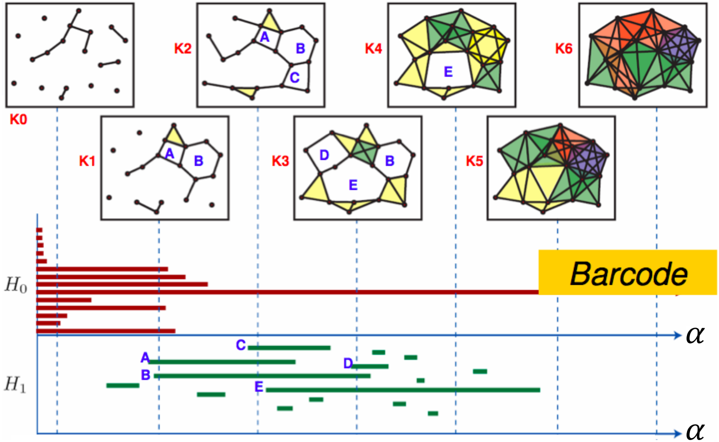
Vietoris-Rips complex at α (R_α).

Persistent Topological Features

- Issue: Choice of the right value for $\alpha \rightarrow$ Persistence.

Persistent Topological Features

- Issue: Choice of the right value for $\alpha \to$ Persistence.
- Construct simplicial complex at different scales i.e. α's $\to$ Filtration.
- Track appearance (birth) and merge (death) of topological features across scales of $\alpha \to$ Persistent homology.



Approximate Simplicial Representations

- **Čech complex** captures the actual topology of the underlying space of the point-cloud, but not feasible to compute $\rightarrow$ *at most* $(1+n)^k$ *simplices of dimension up to* k.

- **Vietoris-Rips Complex** is a 2-approximation of the Čech complex $\rightarrow$ *at most* $(1+n)^k$ *simplices of dimension up to* k.

Computational bottleneck: Enumerating large number of simplices.

Approximate Simplicial Representations

- **Čech complex** captures the actual topology of the underlying space of the point-cloud, but not feasible to compute $\rightarrow$ *at most* $(1+n)^k$ *simplices of dimension up to* k.
- **Vietoris-Rips Complex** is a 2-approximation of the Čech complex $\rightarrow$ *at most* $(1+n)^k$ *simplices of dimension up to* k.

Computational bottleneck: Enumerating large number of simplices.

The Central Computational Question of TDA

Can we have approximate simplicial representations which are computable in reasonable time, yet good approximations to Vietoris-Rips or Čech complex?

A Computationally Faster Approximation: Lazy Witness Complex and The Question to Solve

Lazy witness Complex $LW_\alpha(P, L, \nu)$

Lazy witness Complex $LW_\alpha(P, L, \nu)$ of a point-cloud P is a simplicial complex over a landmark set L that consists of k-simplices $[v_0 v_1 \cdots v_k]$ whose any two points v_i, v_j are in $\alpha + d_\nu$ proximity of some witness point w (d_ν is the distance from w to its ν-th nearest neighbour in L.)

- Lazy witness complex is Vietoris-Rips complex on landmarks L for $\nu = 0$.
- The size of the lazy witness complex is at most $(1 + |L|)^k$ where $|L| << n$.

A Computationally Faster Approximation: Lazy Witness Complex and The Question to Solve

Lazy witness Complex $LW_\alpha(P, L, \nu)$

Lazy witness Complex $LW_\alpha(P, L, \nu)$ of a point-cloud P is a simplicial complex over a landmark set L that consists of k-simplices $[v_0 v_1 \cdots v_k]$ whose any two points v_i, v_j are in $\alpha + d_\nu$ proximity of some witness point w (d_ν is the distance from w to its ν-th nearest neighbour in L.)

- Lazy witness complex is Vietoris-Rips complex on landmarks L for $\nu = 0$.
- The size of the lazy witness complex is at most $(1 + |L|)^k$ where $|L| << n$.

New Questions

How to select the landmarks?
How good are the landmarks selected by an algorithm?
Can we obtain any approximation guarantee for the lazy witness complex?

Our Contributions

The Central Concept

We respond to all these questions by reincarnating the idea of ϵ-net in TDA.

Q. Can we obtain any approximation guarantee for the lazy witness complex?

-> **Lazy witness complex induced by an ϵ-net is a 3-approximation to the Vietoris-Rips complex.**

Q. How good are the landmarks selected by an algorithm?

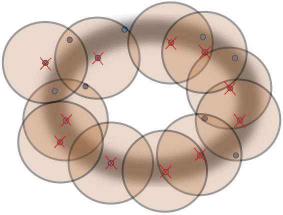
-> **ϵ-net is an ϵ-approximation of the point cloud.**

Q. How to select the landmarks?

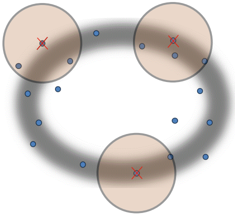
-> **We propose three algorithms to construct ϵ-net.**

Additionally, we validate these theoretical claims experimentally.

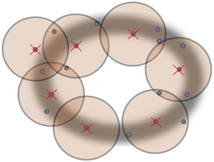
The Central Concept: ϵ-Net



ϵ-sample (Each blue point is within ϵ of some red point)



ϵ-sparse (Each pair of red points are ϵ-far from each other

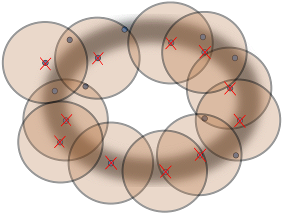


ϵ-net

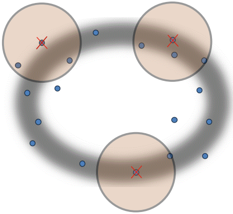
Definition (ϵ-sample)

A set $L \subseteq P$ is an ϵ-sample of P if the collection $\{B_\epsilon(x) : x \in L\}$ of ϵ-balls of radius ϵ-covers P , i.e. $P = \cup_{x \in L} B_\epsilon(x)$.

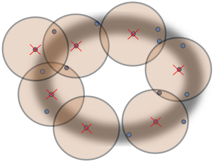
The Central Concept: ϵ-Net



ϵ-sample (Each blue point is within ϵ of some red point)



ϵ-sparse (Each pair of red points are ϵ-far from each other

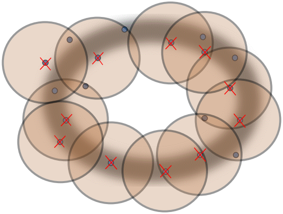


ϵ-net

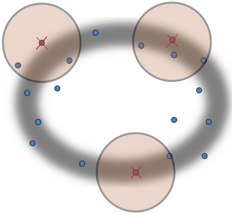
Definition (ϵ-sparse)

A set $L \subset P$ is ϵ-sparse if for all $x, y \in L$, $d(x, y) > \epsilon$.

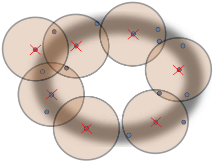
The Central Concept: ϵ-Net



ϵ-sample (Each blue point is within ϵ of some red point)



ϵ-sparse (Each pair of red points are ϵ-far from each other



ϵ-net

ϵ-net

A subset (L) of points which is ϵ-sparse and ϵ-sample of the point-cloud (P).

Approximation Guarantee: ϵ-Net Induced Lazy Witness Complex

Approximating the Vietoris-Rips Complex

If the landmark set L is an ϵ-net of the point cloud P, the lazy witness complex at α and $\nu = 1$ is 3-approximation of the Vietoris-Rips complex of L for $\alpha \geq 2\epsilon$. Mathematically,

$$R_{\alpha/3}(L) \subseteq LW_{\alpha}(P, L, 1) \subseteq R_{3\alpha}(L) \qquad \forall \alpha \geq 2\epsilon.$$

Approximation Guarantee: ϵ-Net Induced Lazy Witness Complex

Approximating the Vietoris-Rips Complex

If the landmark set L is an ϵ-net of the point cloud P, the lazy witness complex at α and $\nu = 1$ is 3-approximation of the Vietoris-Rips complex of L for $\alpha \geq 2\epsilon$. Mathematically,

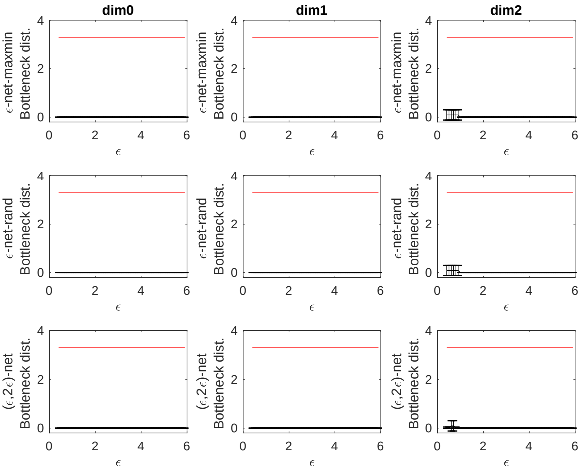
$$R_{\alpha/3}(L) \subseteq LW_{\alpha}(P, L, 1) \subseteq R_{3\alpha}(L) \qquad \forall \alpha \geq 2\epsilon.$$

Approximating the Persistent topological feature

If we compare the bars (in the barcode) appearing after 2ϵ, barcodes (log-scale) of the lazy witness filtration and the Vietoris-Rips filtration are $3 \log 3$-approximations of each other. (By weak-stability theorem[a])

[a] Chazal et al., "Proximity of persistence modules and their diagrams".

Experimental Validation of Approximation Guarantee



Validation of the bound on Point-cloud sampled from Torus.

Quality of Landmarks: Properties of ϵ-Net

Point-cloud Approximation Guarantee

The Hausdorff distance between the point cloud and its ϵ-net is at most ϵ.

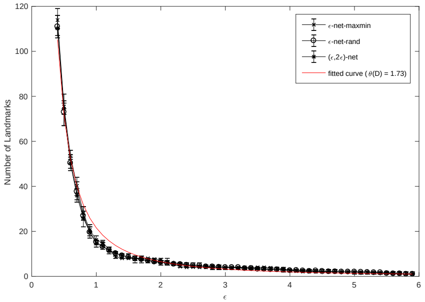
Quality of Landmarks: Properties of ϵ-Net

Point-cloud Approximation Guarantee

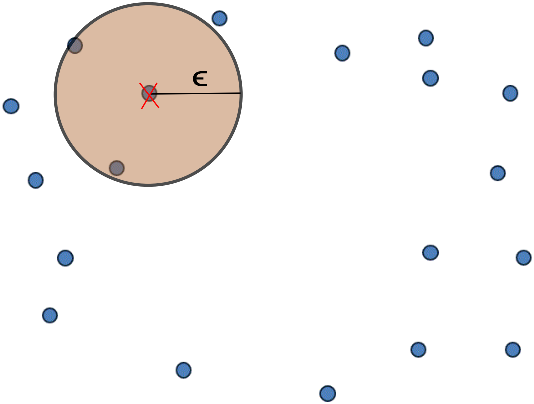
The Hausdorff distance between the point cloud and its ϵ-net is at most ϵ.

Size of an ϵ-Net

The number of points in an ϵ-net is at most $(\frac{\Delta}{\epsilon})^{\theta(D)}$ for $P \in \mathbb{R}^D$ of diameter Δ.

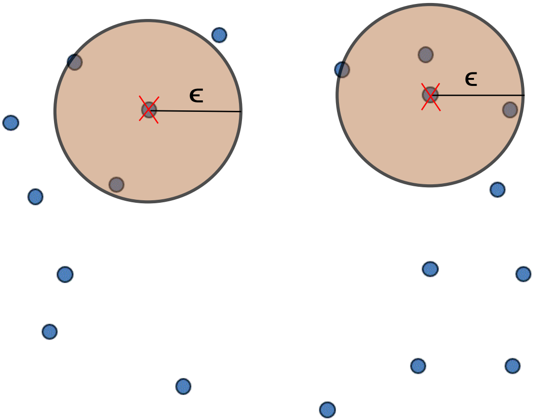


Algorithm: ϵ-net-rand



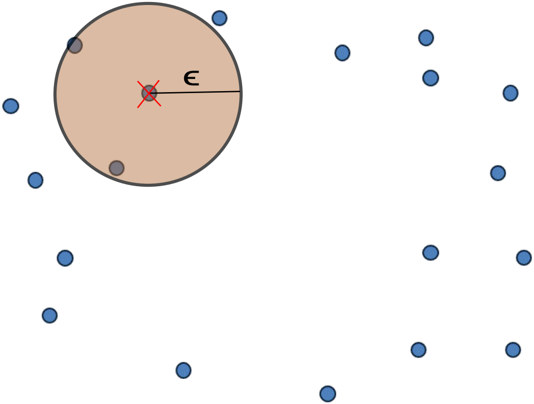
First landmark: Select uniformly at random from the point-cloud. Mark points in its ϵ-ball.

Algorithm: ϵ-net-rand



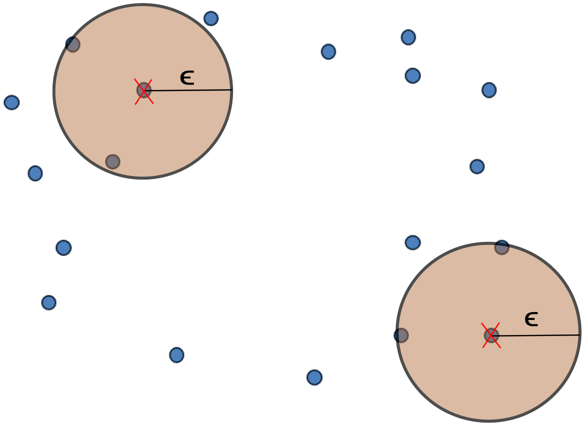
Next landmark: Select u.a.r from the set of unmarked points. Mark points in its ϵ-ball. And so on.

Algorithm: ϵ-net-maxmin



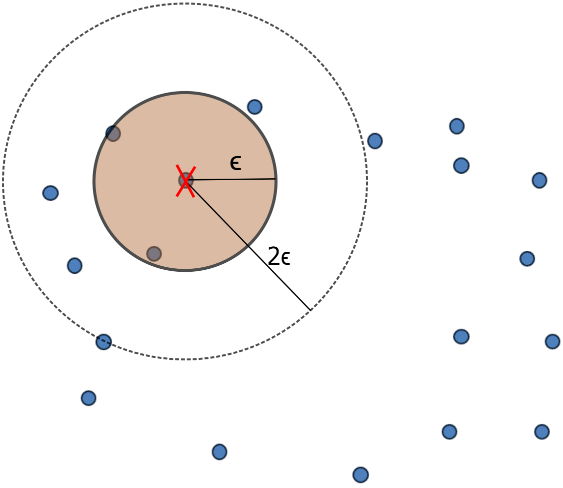
First landmark: Select u.a.r. from the point-cloud. Mark points in its ϵ-ball.

Algorithm: ϵ-net-maxmin



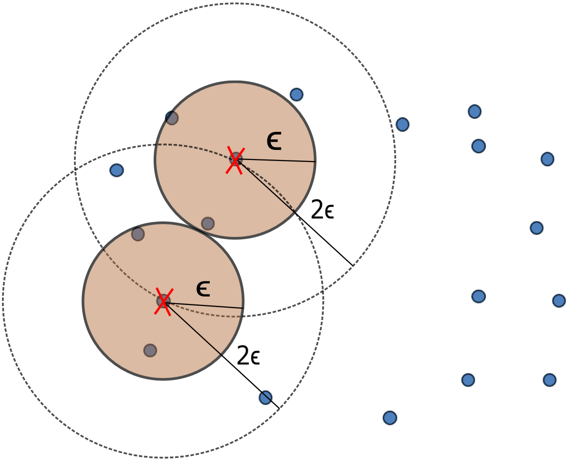
Next landmark: Select the point that is farthest from the current set of landmarks. And so on.

Algorithm: $(\epsilon, 2\epsilon)$-net



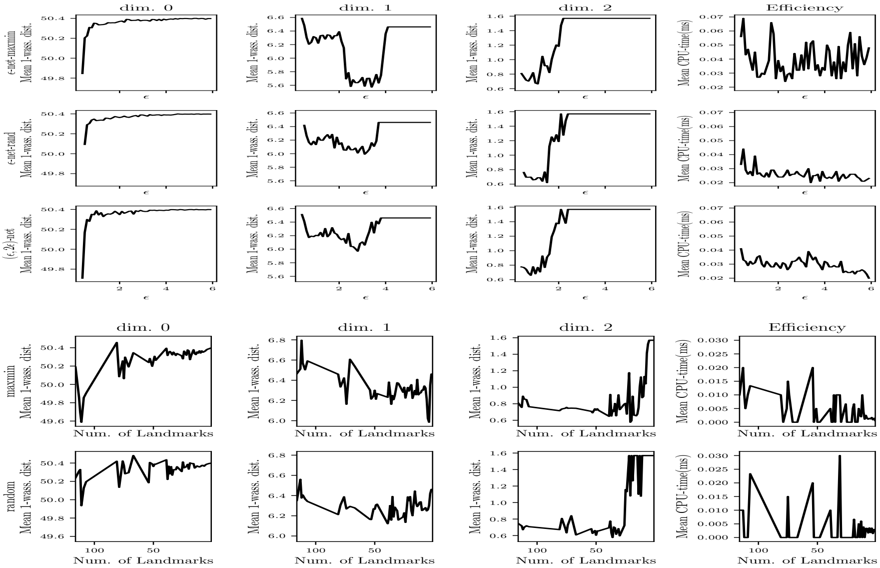
First landmark: Select u.a.r from the point-cloud. Mark points in its ϵ-ball.

Algorithm: $(\epsilon, 2\epsilon)$-net



Second landmark: Select u.a.r from the unmarked points in the $(\epsilon, 2\epsilon)$ envelope of the current set of landmarks. And so on.

Experimental Evaluation: Effectiveness-Efficiency



Effectiveness and Efficiency of the algorithms on Torus dataset.

Take Away

- If the landmarks is an ϵ-net, we know about the quality of the-
  - landmarks
  - lazy witness approximation
  - approximated persistent topological features.
- Use ϵ-net as landmarks.
- *You have a point-cloud dataset? -> Apply Topological Data Analysis!*

Thank You!

Supplementary Slides

Experiments: Datasets

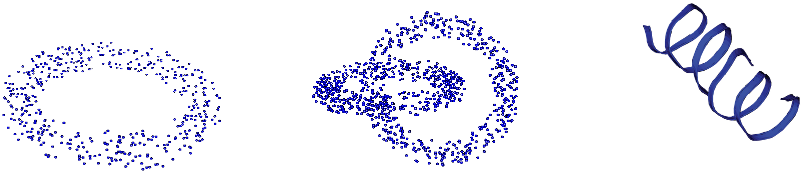


Figure 1: (left) Torus, (middle) Tangled-torus, and (right) 1grm Dataset

Relation to Maxmin and Random Landmark Selection Algorithms

- Given the number of landmarks $K > 1$, the set of landmarks selected by the algorithm random/maxmin is δ-sparse where δ is the minimum of the pairwise distances among the landmarks.
- The choice of K may not necessarily make the landmarks a δ-sample of the point cloud.

Algorithm complexity

- ϵ-net-rand: $O(\frac{1}{\epsilon^D})$
- ϵ-net-maxmin: $O(\frac{n}{\epsilon^D})$
- $(\epsilon, 2\epsilon)$-net: $O(\frac{1}{\epsilon^D}) \log(\frac{1}{\epsilon})$

Experimental Evaluation: Stability

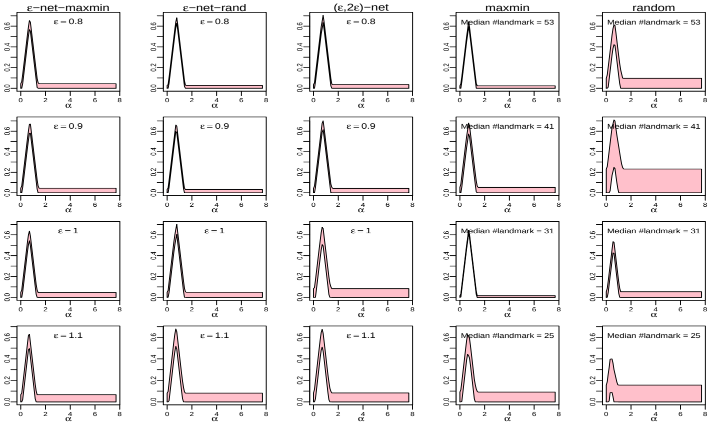


Figure 2: 95% confidence band of the rank one persistence landscape at dimension 1 of the lazy witness filtration induced by the landmark selection algorithms on Tangled-torus dataset.

Whats Next?

The topological approximation guarantee is

- with respect to the Vietoris-Rips complex on ϵ-**net landmarks** chosen from a **point-cloud input**.

Next up -

- Better guarantee w.r.t Vietoris-Rips complex on point-cloud:-
  - Improved the approximation guarantee from 3-approximation of $R_\alpha(L)$ to $\frac{3log(c)}{2}$-approximation of $R_\alpha(P)$ for $c \geq 2$.
- Graph data [1]:-
  - Defined ϵ-net for graphs.
  - Devised algorithms for computing ϵ-net of graphs.
  - Potential applications: Graph clustering, Graph visualization, Graph classification.
- Comparison with Sparse-Rips and Graph Induced filtration (A weakness!).

[1] To appear at ECML-PKDD'19 workshop on Applied Topological Data Analysis