Topological Data Analysis

Conceptual Introduction

Structure

- Motivation:
  - What can we do with it?
  - Why can we do with it?
- Conceptual Platform
- Reducing to the Nerve
- The persistence diagram as the final output

Primary Reference

Introductory Topological Data Analysis, Sheffar 2020

Motivation - What can we do with it?

- Medical signatures: Alzheimer's disease (Rieck et al. 2021), Neuropsychological Analysis (Robinson & Turner, 2016)
- Detection of 2D shapes (Hofer et al. 2017)

Motivation - Why do we?

- ML on novel spaces
  - We typically work in Euclidean space
  - Topological algorithms generalise immediately to more novel spaces, e.g. discrete spaces, disconnected spaces, rough spaces
- Topological algorithms are designed to be, in a sense, robust
- Topology is a natural language to describe a robust notion of shape
- Cheap global features

Motivation - Why do we?

- ML on novel spaces
    - We typically work in Euclidean space
    - Topological algorithms generalise immediately to more novel spaces, e.g. discrete spaces, disconnected spaces, rough spaces
- Topological algorithms are designed to be, in a sense, robust
- Topology is a natural language to describe a robust notion of shape
- Cheap global features
- Dimensionality Reduction

Conceptual Platform - Metric Spaces

- A **Metric** is a distance measure between two vectors $d(x, y)$
- Metrics are positive and respect the triangle inequality
- Examples
    - Euclidean Distance $d(x, y) = ||x - y||$
    - L1 Norm $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$
    - Discrete metric $d(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$

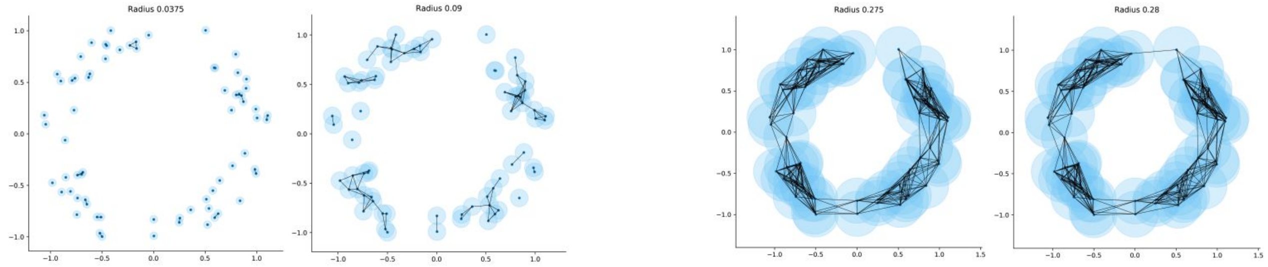
Conceptual Platform - Balls

- A **Ball (in metric space induced by d)** of radius ε centred at x is the set of vectors within ε of x, according to given metric d
- Examples
  - $\mathbf{B}\left(x, \frac{1}{2}\right)$ is a radius ½ circular area centred at x for the euclidean metric.
  - $\mathbf{B}\left(x, \frac{1}{2}\right)$ in

    Q: What would this be in the discrete (binary) metric

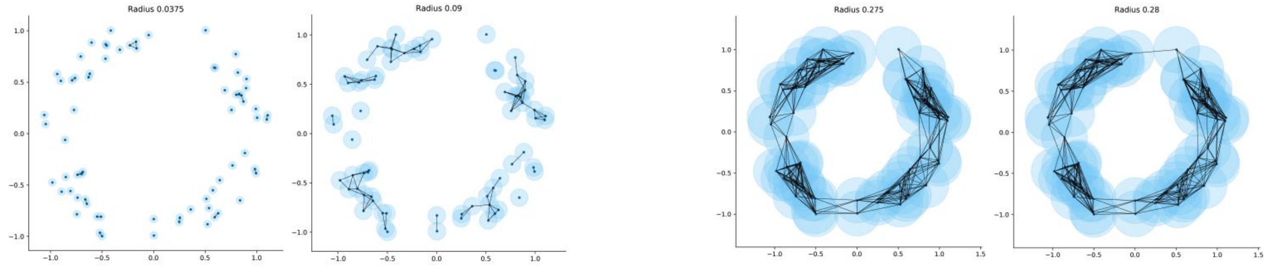
Clustering in a Metric

- This framework lends itself to clustering like DB Scan
- Clustering happens with a parameter ε



Clustering in a Metric

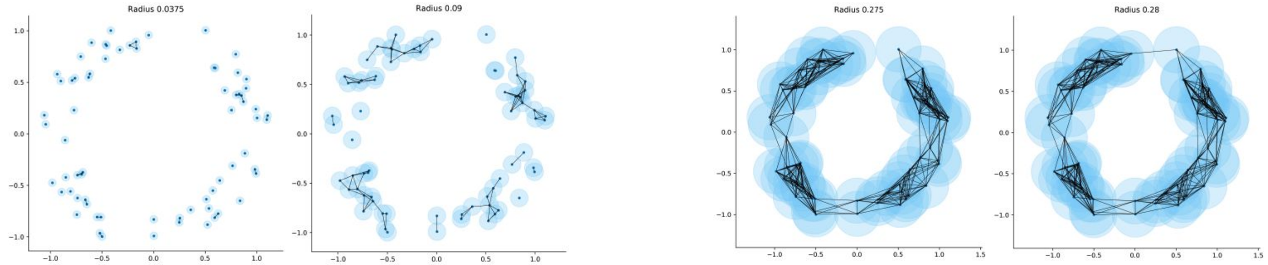
- Different structures form at different values of ε



Components 52 14 1 1

Clustering in a Metric

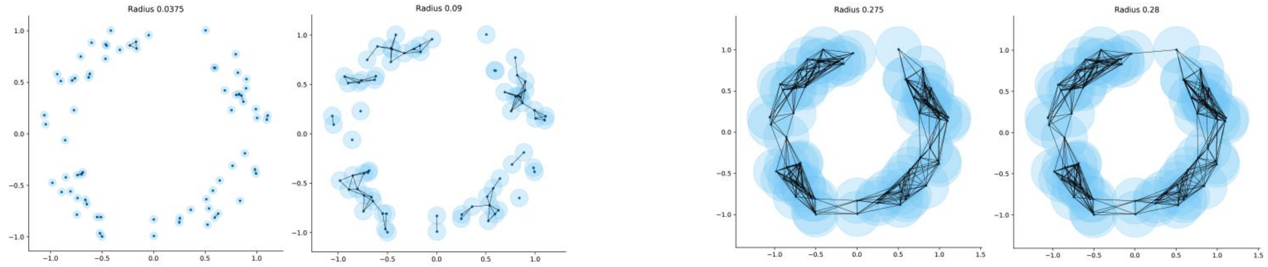
- Different structures form at different values of ε



Components	52	14	1	1
Loops	0	0	0	1

Clustering in a Metric

- Different structures form at different values of ε
- We get different levels of **fineness** for different ε



Components	52	14	1	1
Loops	0	0	0	1

Conceptual Platform - Topology (briefly)

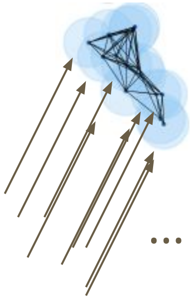
- A **Topology** is, formally, a set of “regions” in a space (e.g. Balls, squares, points) which is closed under arbitrary unions, finite intersections, and contains the whole space (and the empty set)
- Every metric space defines a topology (the regions are the open balls)
- Topologies can be *coarser* or *finer* than other topologies

Conceptual Platform - Topology (briefly)

- A **Topology** is, formally, a set of "regions" in a space (e.g. Balls, squares, points) which is closed under arbitrary unions, finite intersections, and contains the whole space (and the empty set)
- Every metric space defines a topology (the regions are the open balls)
- Topologies can be *coarser* or *finer* than other topologies
- The area covered by/imbued with the topology is the **Topological space**

Conceptual Platform - Topology (briefly)

- Every metric space defines a topology (the regions are the open balls)
- Topologies can be *coarser* or *finer* than other topologies
- Example:



Fine

Coarse

Applied Topology

Take the clusters as the regions. We could reduce dimensions by representing the data through the clusters/regions and the shapes they form.

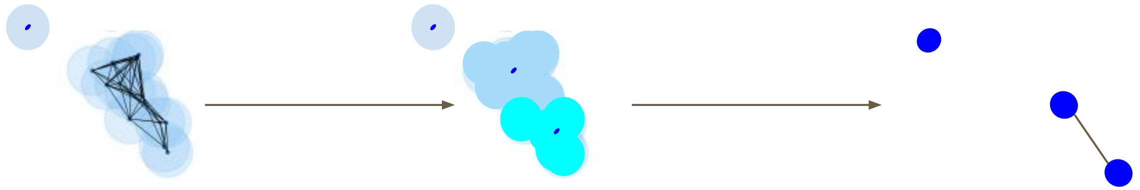
Example:

Reduction onto 0 dim complex

Applied Topology

Now picking the regions as more bespoke unions of balls, we end up retaining a different fineness of structure.

Example:



Reduction onto 1 dim complex
allowing clusters to overlap

Conceptual Platform - Simplices / Complexes

- A **simplex** of dimension k is a complete graph of k+1 pts
- A **complex** of dimension k is a graph made of simplices, each of dimension at most k
- The sense of dimension comes from the dimension of triangle/tetrahedron that the complex defines

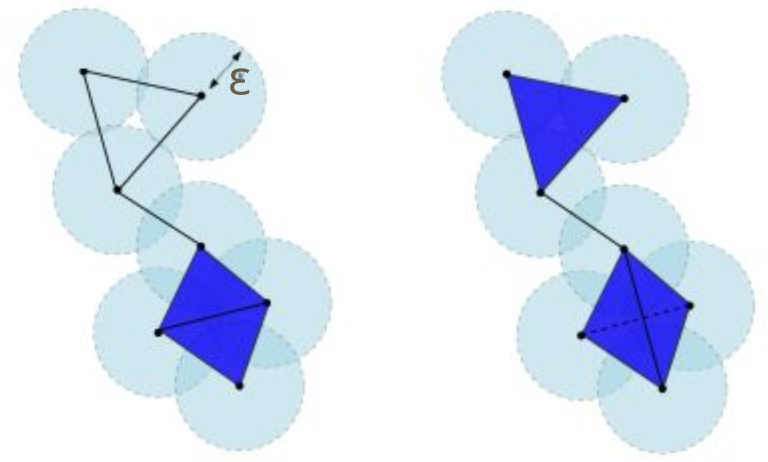


Figure 11: Čech *(left) and Rips (right) complexes on a toy dataset* $\mathbb{X}$ - figure via Chazal and Michel (2017)

Conceptual Platform - Simplices / Complexes

- A **simplex** of dimension k is a complete graph of k+1 pts
- A **complex** of dimension k is a graph made of simplices, each of dimension at most k
- The sense of dimension comes from the dimension of triangle/tetrahedron that the complex defines

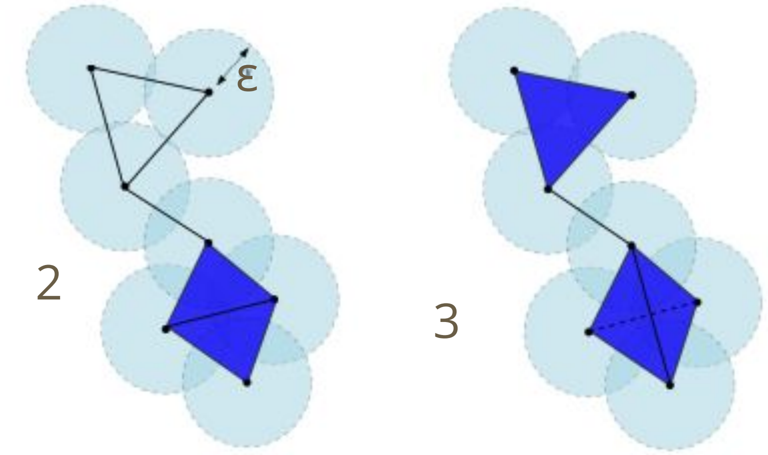


Figure 11: *Čech (left) and Rips (right) complexes on a toy dataset* $\mathbb{X}$ - figure via Chazal and Michel (2017)

Getting Complexes (Footnote)

- Then Ĉech complex is computationally expensive but an exact representation.
- The Rips (Vietoris-Rips) complex is cheaper but more of an approximation.

The point is that these representations can be generated.

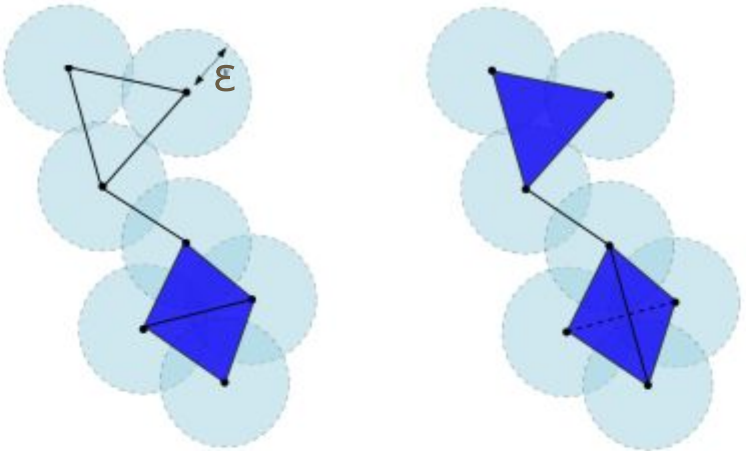
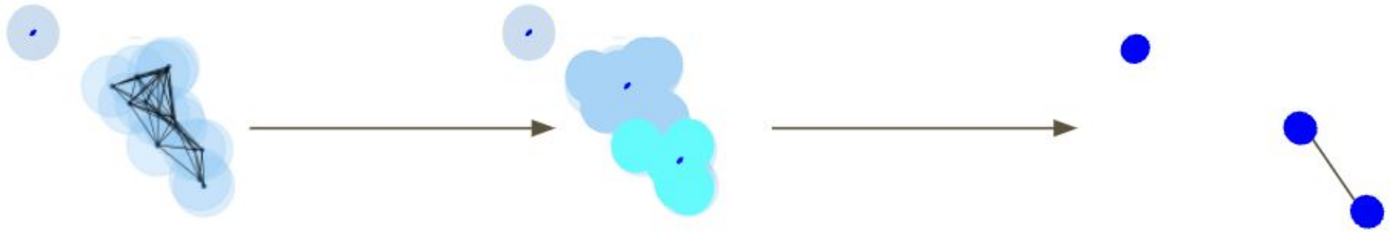


Figure 11: *Čech (left) and Rips (right) complexes on a toy dataset* $\mathbb{X}$ - figure via Chazal and Michel (2017)

Map data onto “nerves” (representation by complex)

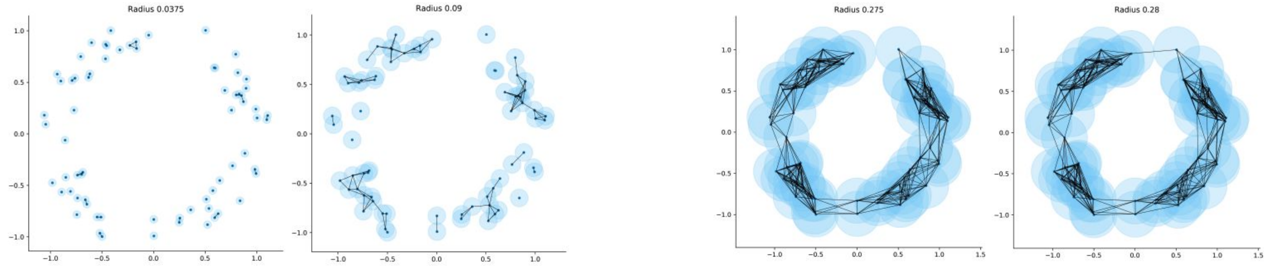


Nerve ≅ Complex

- Mapping onto nerves maintains to overall shape (components, loops, etc.)
- Nerve *homotopic to* Topological Space identical shape

Encoding shape

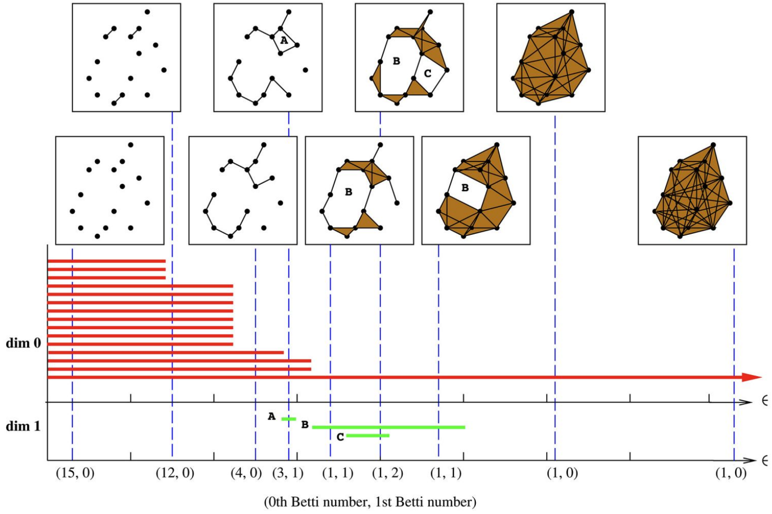
- For certain ε/fineness, we have a certain connected components, loops, and so on for higher dimensional equivalents
- These persist for some interval of ε



Components	52	14	1	1
Loops	0	0	0	1

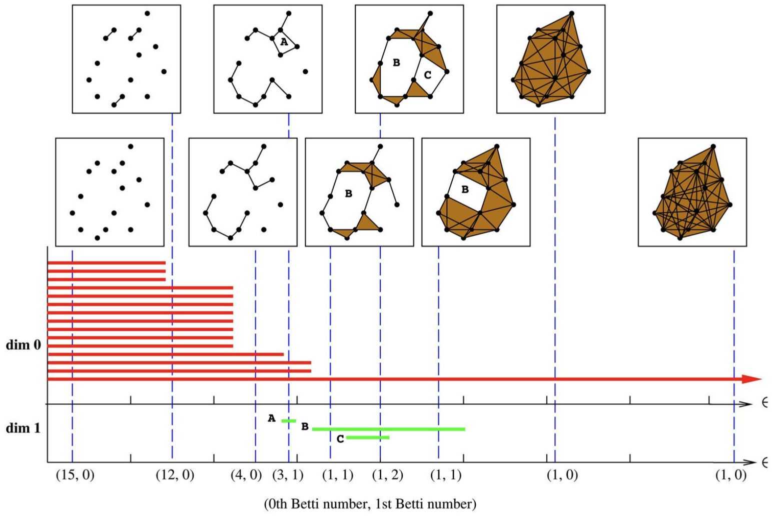
Encoding shape

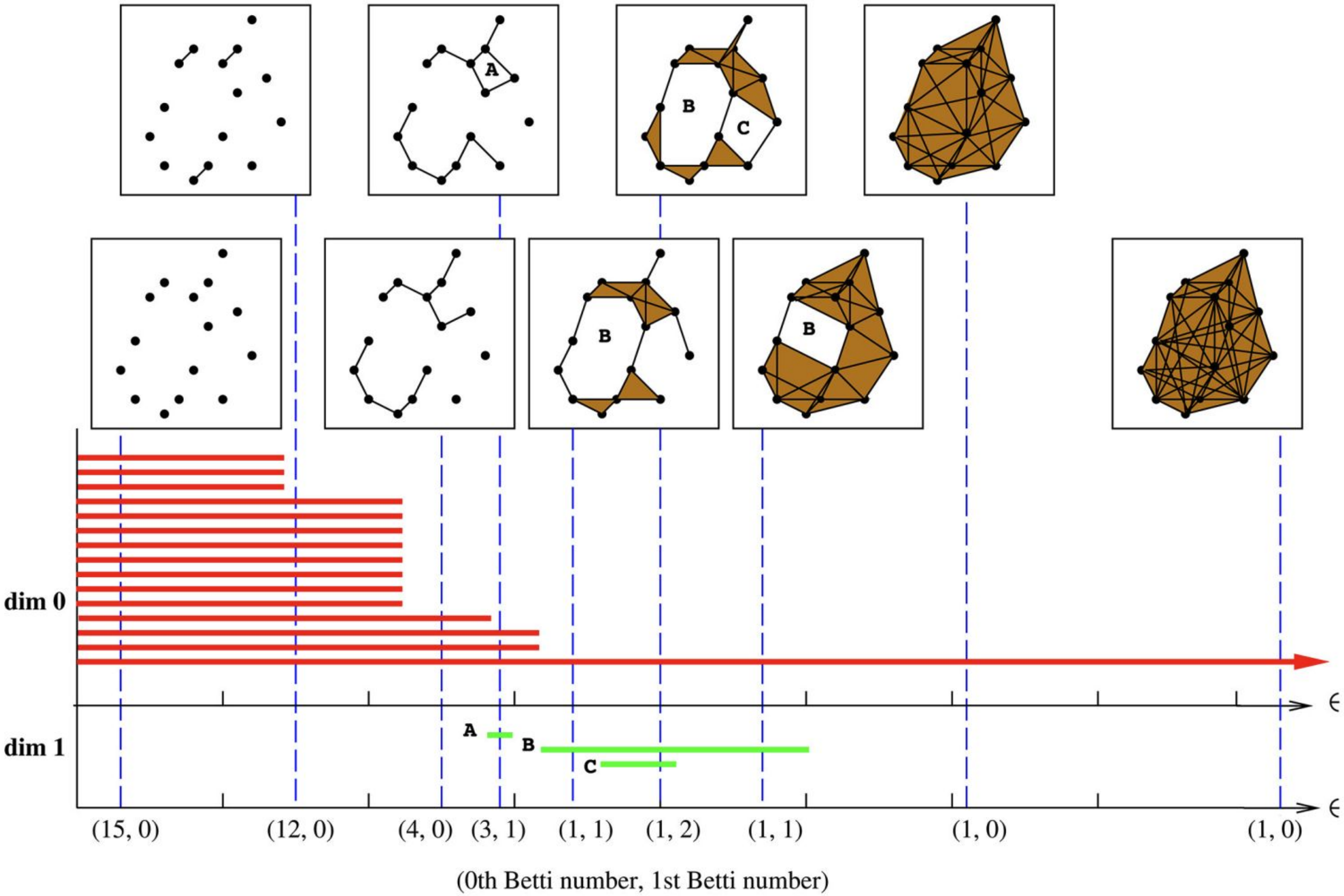
- We encode components, loops, on further voids as counts, **Betti Numbers** β_k
- E.g. β_1 is number of loops
- The set of betti numbers for the generated complex is a function of ε
- Betti numbers (and the counted features) persist for an interval of ε values



Encoding shape

- The output of the TDA pipeline is a **persistence diagram**
- This diagram is the typical output
- Derived from the diagram are relevant features - see here




A
B
C
B
B
dim 0
dim 1
A
B
C
ϵ
ϵ
(15, 0)
(12, 0)
(4, 0)
(3, 1)
(1, 1)
(1, 2)
(1, 1)
(1, 0)
(1, 0)
(0th Betti number, 1st Betti number)