DISMA Dipartimento di Scienze Matematiche G. L. Lagrange

ECCELLENZA 2018 • 2022

ENBIS2021 - Spring Conference Persistent Homology for Market Basket Analysis

Sara Scaramuccia (joint work with Roberto Fontana)

17/05/2021

POLITECNICO DI TORINO

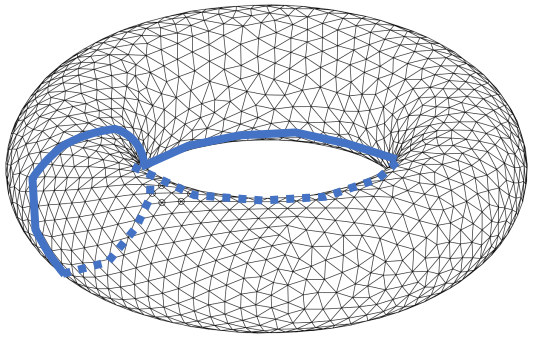
SmartData@PoliTO

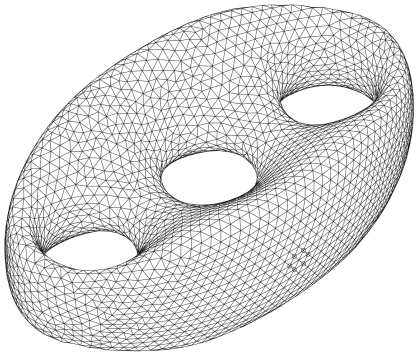
SmartData

A topological signature

Homology is the main feature invariant under topological transformations involved in TDA (topological data analysis). It counts the number of

0-cycles	connected components
1-cycles	independent loops
k-cycles	k-dimensional ind. cavities

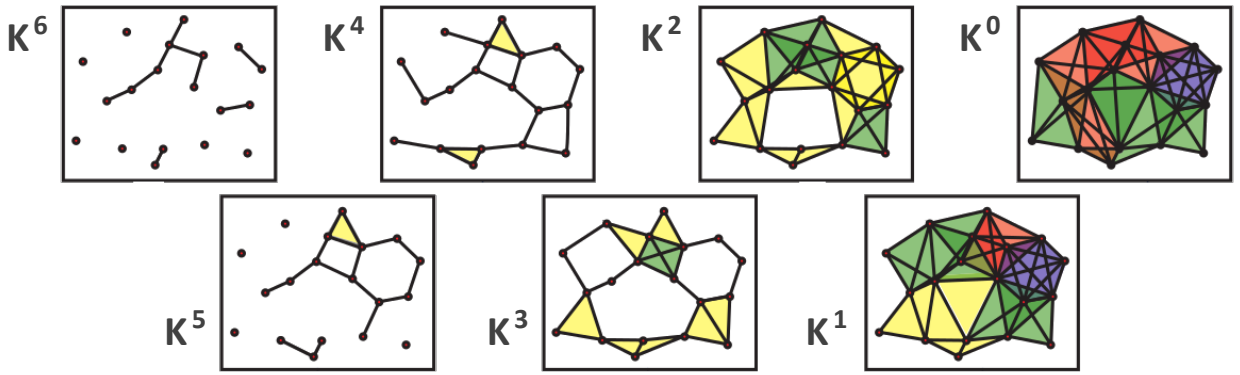




Persistent Homology (PH)

In a Nutshell:

Persistent homology tracks homology features along a nested sequence of discrete shapes



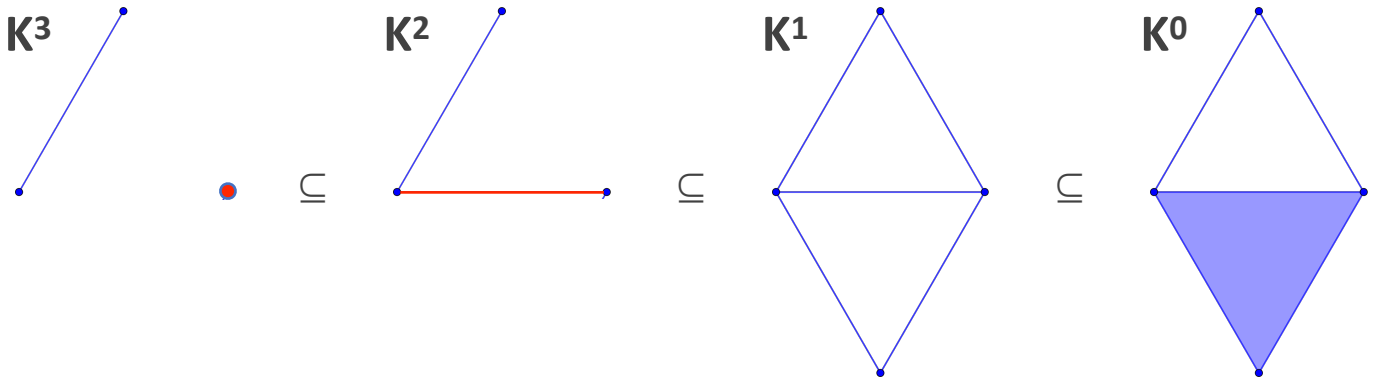
- filtering out information
- data-driven threshold selection

Image from [Ghrist 2008]

Persistence Pairs

The ***core information*** of persistent homology is given by the ***persistence pairs***

Given a filtration $K^m \subseteq K^{m-1} \subseteq \dots \subseteq K^0$,



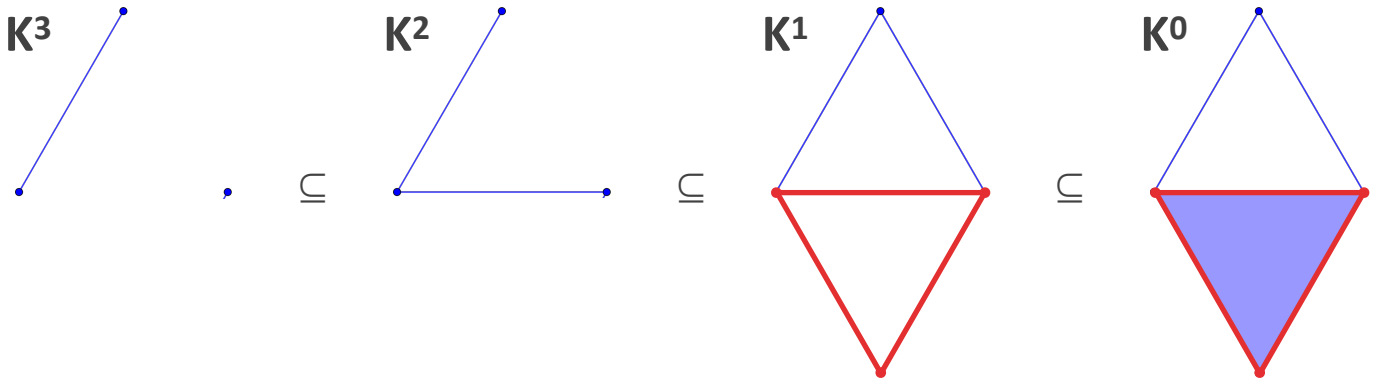
(3, 2) ⟷ ***(0-simplex,1-simplex)***

A **persistence pair *(p, q)*** is an element in $\{0, \dots, m\} \times (\{0, \dots, m\} \cup \{\infty\})$ such that $p < q$ representing a **homological class** that is **born at step *p*** and **dies at step *q***

Persistence Pairs

The ***core information*** of persistent homology is given by the ***persistence pairs***

Given a filtration $K^m \subseteq K^{m-1} \subseteq \dots \subseteq K^0$,



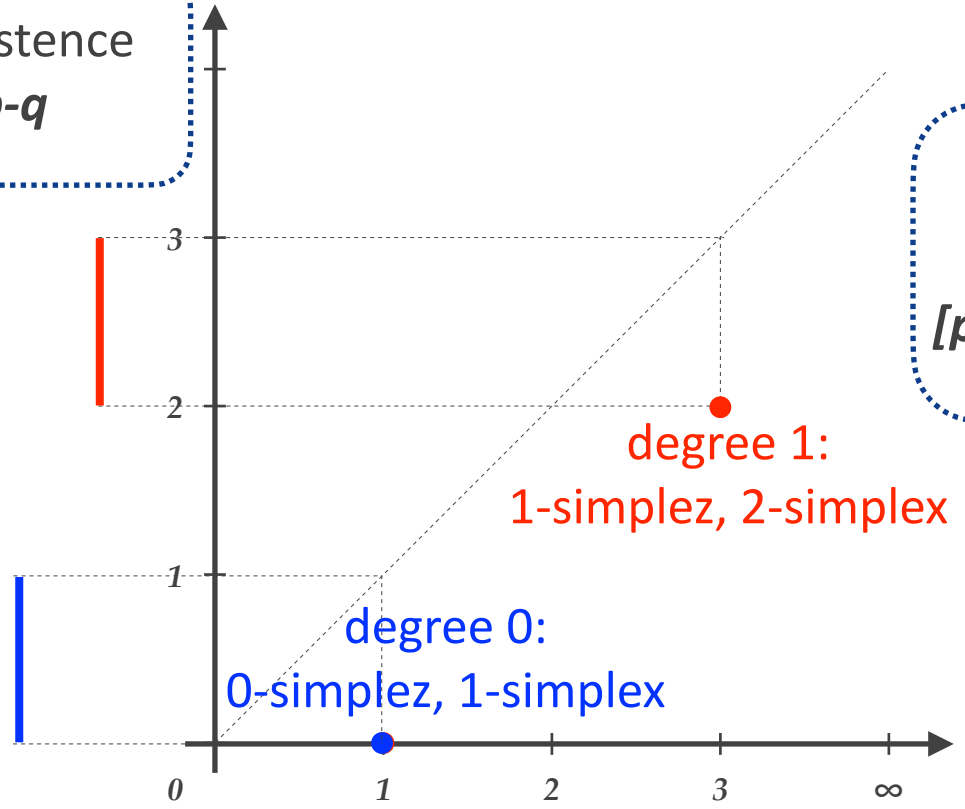
(1, 0) ⟷ ***(0-simplex,1-simplex)***

A **persistence pair *(p, q)*** is an element in $\{0, \dots, m\} \times (\{0, \dots, m\} \cup \{\infty\})$ such that $p < q$ representing a **homological class** that is **born at step *p*** and **dies at step *q***

Visualizing Persistence Pairs

Persistence Diagrams ***encode PH information***

persistence
p-q



persistence pair

[p, q] ⟷ ***(k-simplex,(k+1)-simplex)***

Market Basket Analysis (MBA)

aim: detecting association rules

A : itemset

B : itemset

$A \Rightarrow B$: association rule

$\mathbb{P}(A)$: purchase probability

$\mathbb{P}(B)$: purchase probability

$\mathbb{P}(B \mid A)$: conditional probability

conditional probability might not be enough:

support usually compared to the independent case

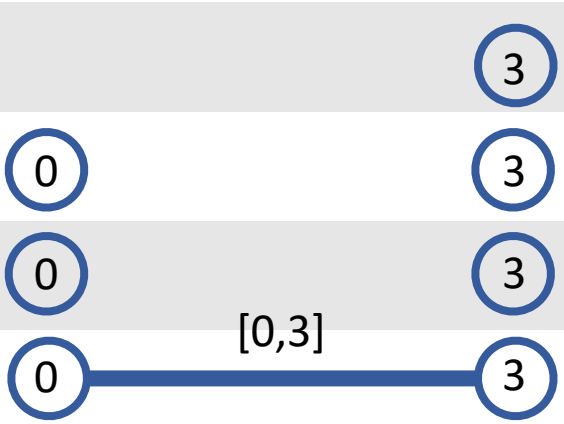
$$\text{lift}(A, B) := \frac{\mathbb{P}(B \mid A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)\mathbb{P}(B)}$$

Supermarket dataset:

https://www.kaggle.com/mittalvasu95/the-bread-basket

itemsets as simplices

bread	brownie	cake	coffee	cookies	hot chocolate	croissant	pastry	sandwich	tea	counter	ℙ
0	0	0	1	0	0	0	0	0	0	4529	0.5431
1	0	0	0	0	0	0	0	0	0	3097	0.3713
0	0	0	0	0	0	0	0	0	1	1352	0.1621
1	0	0	1	0	0	0	0	0	0	852	0.1021



0.5431

0.3713

0.1621

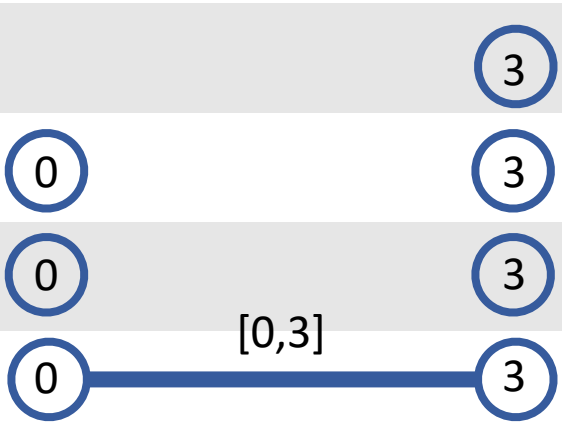
0.1021

Supermarket dataset:

https://www.kaggle.com/mittalvasu95/the-bread-basket

itemsets: support

bread	brownie	cake	coffee	cookies	hot chocolate	croissant	pastry	sandwich	tea	counter	support
0	0	0	1	0	0	0	0	0	0	4529	0.5431
1	0	0	0	0	0	0	0	0	0	3097	0.3713
0	0	0	0	0	0	0	0	0	1	1352	0.1621
1	0	0	1	0	0	0	0	0	0	852	0.1021



0.5431

0.3713

0.1621

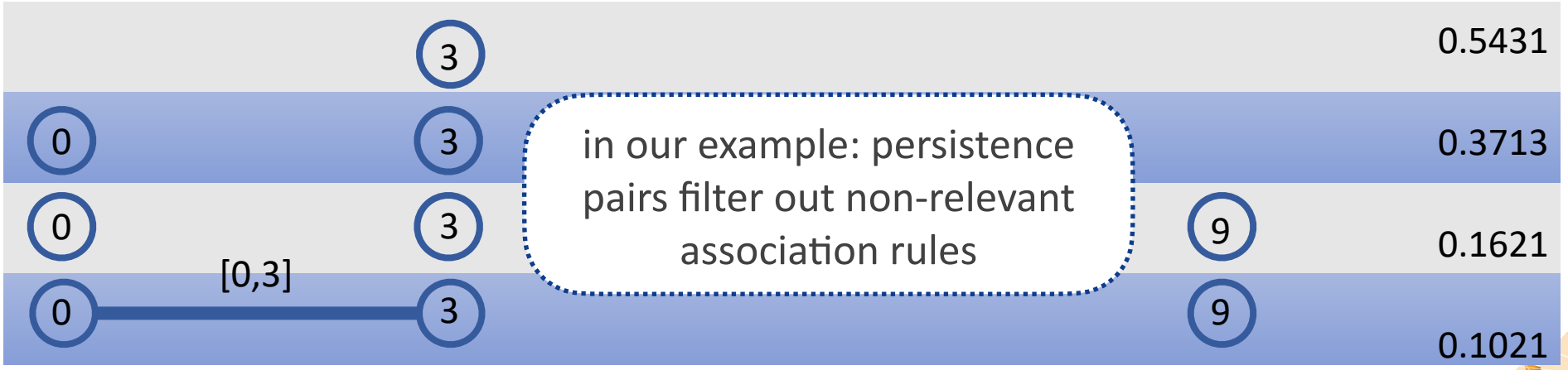
0.1021

Supermarket dataset:

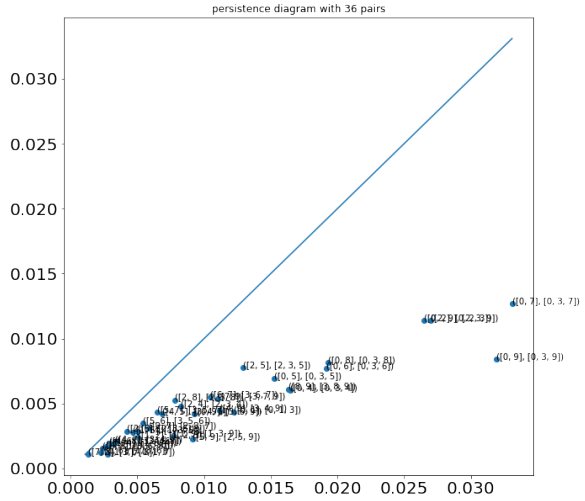
https://www.kaggle.com/mittalvasu95/the-bread-basket

itemsets: confidence - persistence

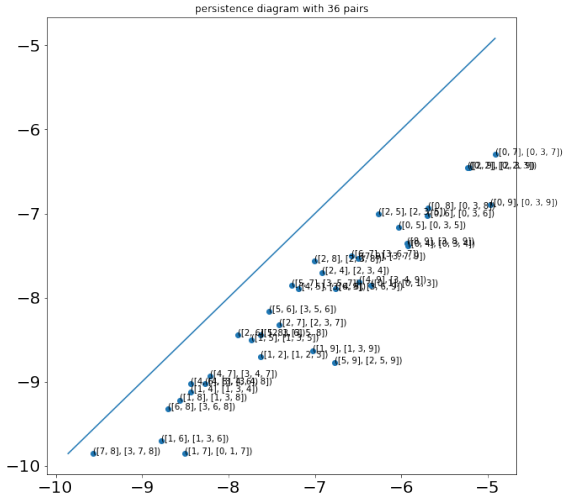
bread	brownie	cake	coffee	cookies	hot chocolate	croissant	pastry	sandwich	tea	counter	support
0	0	0	1	0	0	0	0	0	0	4529	0.5431
1	0	0	0	0	0	0	0	0	0	3097	0.3713
0	0	0	0	0	0	0	0	0	1	1352	0.1621
1	0	0	1	0	0	0	0	0	0	852	0.1021



Correspondence PH - MBA



axes represent support



logarithic axes: log(confidence)=diagonal shift

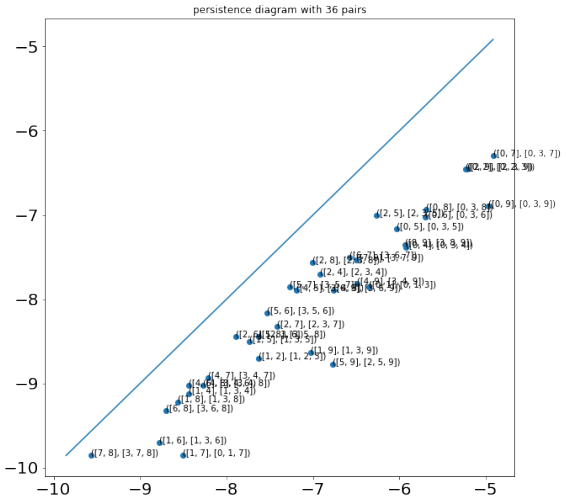
Supermarket dataset:

https://www.kaggle.com/mittalvasu95/the-bread-basket

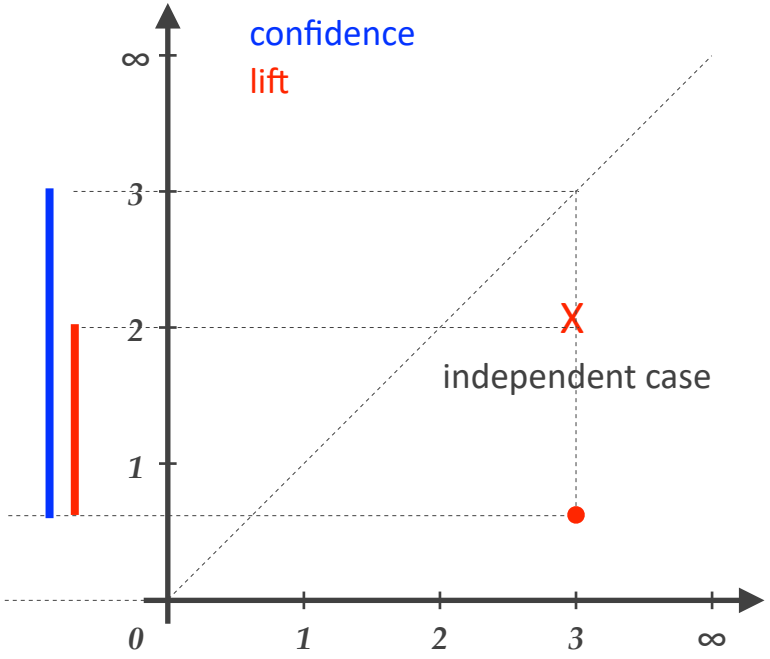
itemsets: lift

bread	brownie	cake	coffee	cookies	hot chocolate	croissant	pastry	sandwich	tea	counter	support
0	0	0	1	0	0	0	0	0	0	4529	0.5431
1	0	0	0	0	0	0	0	0	0	3097	0.3713
0	0	0	0	0	0	0	0	0	1	1352	0.1621
1	0	0	1	0	0	0	0	0	0	852	0.1021

Correspondence PH - MBA

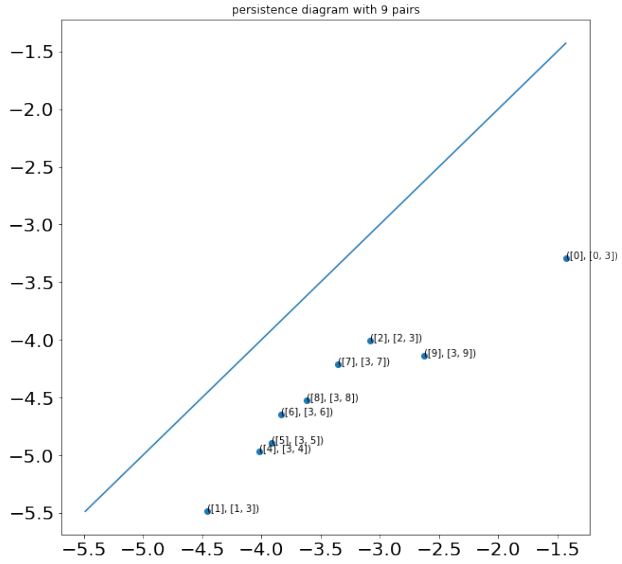


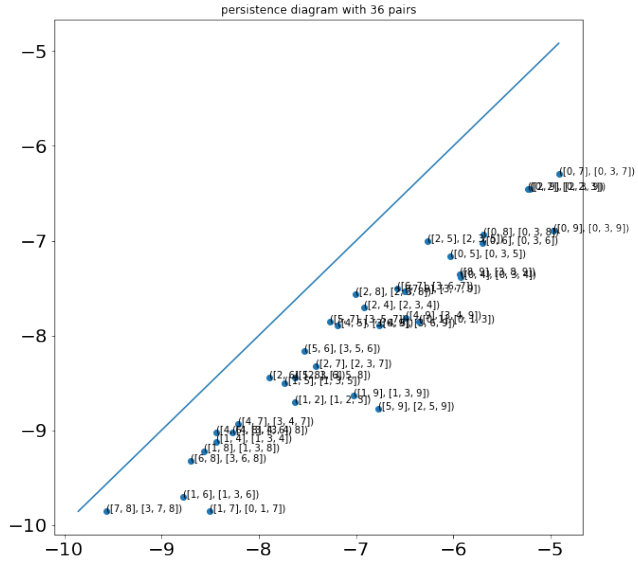
logarithic axes:
log(confidence)=diagonal shift



Results for supermarket dataset

dataset : 8339 transactions, 10 items, 3 max items per itemset



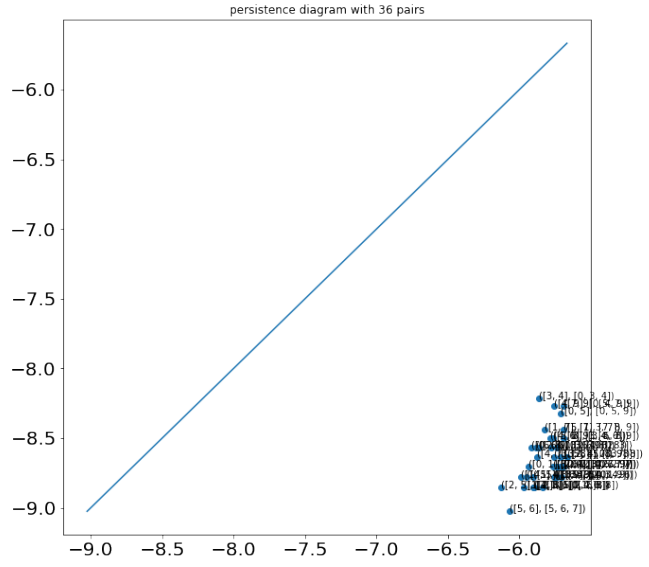


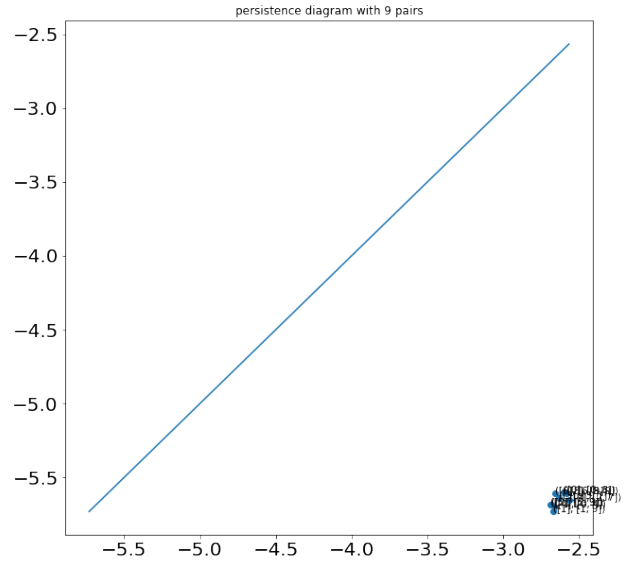
0-degree

1-degree

Results for supermarket dataset

Random : same number of transactions with random items



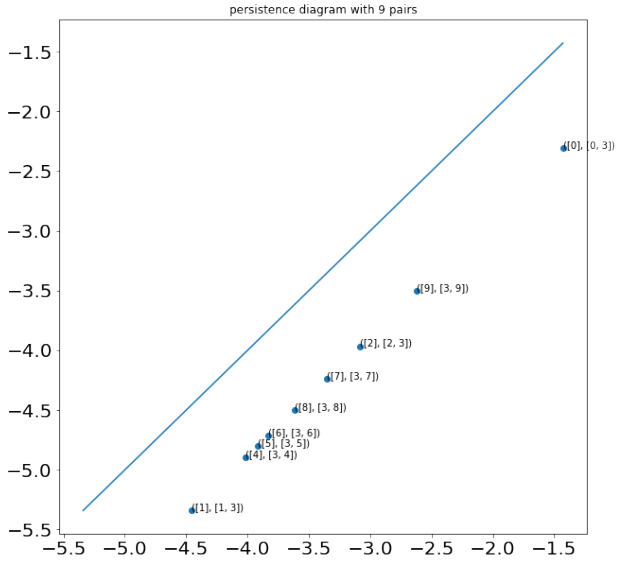


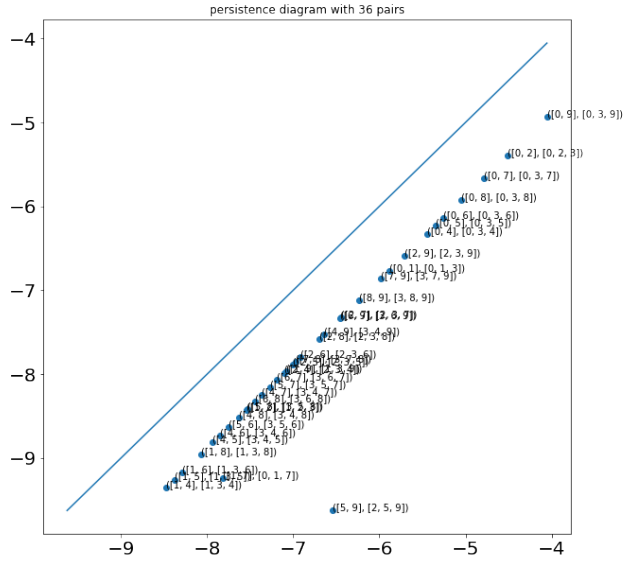
0-degree

1-degree

Results for supermarket dataset

independent : itemset probability = item probability multiplied





0-degree

1-degree

Conclusions

- promising PH properties:
  - filtering rules
  - determining data-driven thresholds
- correspondences PH - MBA detected:
  - support, confidence, lift
- other application investigated:
  - museum's visitors behavior
- to be investigated
  - non-incident persistence pairs and confidence
  - defining global summaries based on PH

Ackowledgements

The authors acknowledge the connection of this work with ELBA, an ERASMUS+ project aiming at the establishment of training and research centers on Big Data Analysis in Central Asia (https://elba.famnit.upr.si)

Erasmus+

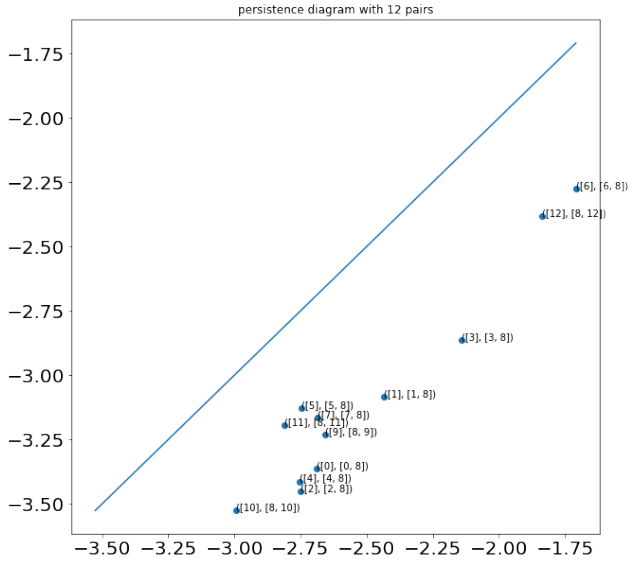
Bibliography

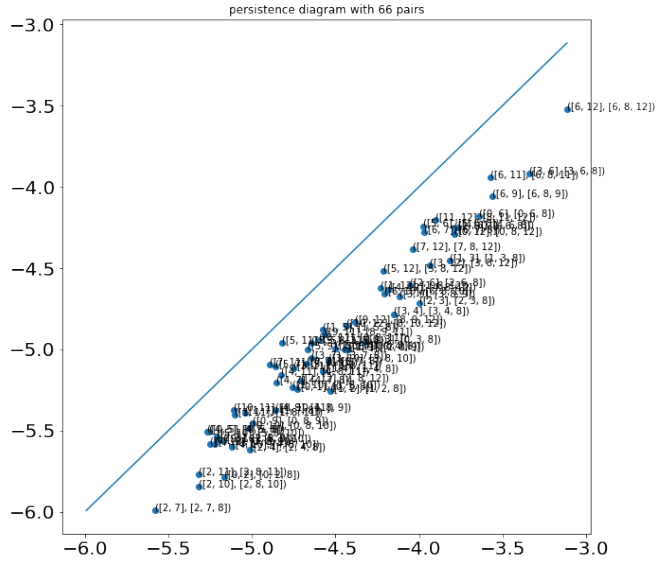
- ***TDA:***
  - H. Edelsbrunner, J. Harer. ***Computational topology: an introduction***. American Mathematical Soc., 2010.
  - R. W. Ghrist. ***Elementary applied topology***. Seattle: Createspace, 2014.
  - G. Carlsson. ***Topology and data***. Bulletin of the American Mathematical Society 46.2, pages 255-308, 2009.
- ***Persistent Homology:***
  - U. Fugacci, S. Scaramuccia, F. Iuricich, L. De Floriani. ***Persistent homology: a step-by-step introduction for newcomers***. Eurographics Italian Chapter Conference, pages 1-10, 2016.
- ***PH Stability Results:***
  - D. Cohen-Steiner, H. Edelsbrunner, J. Harer. ***Stability of persistence diagrams***. Discrete & Computational Geometry 37.1, pages 103-120, 2007.

POLITECNICO DI TORINO

Results for museum dataset

dataset : 8339 transactions, 10 items, 3 max items per itemset



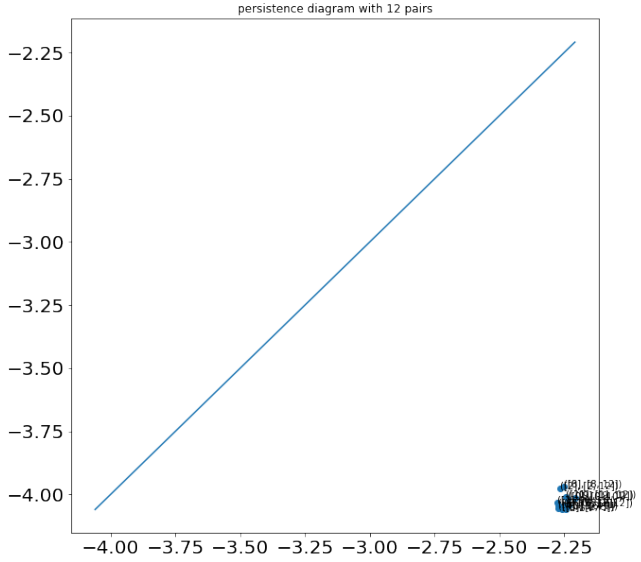


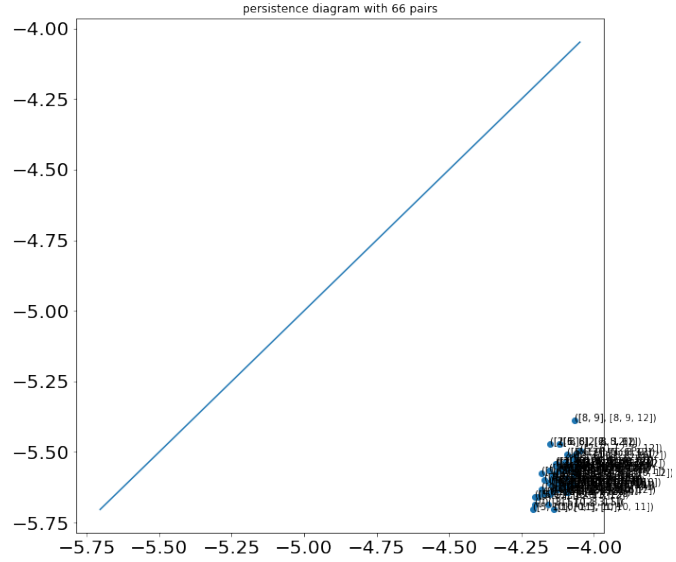
0-degree

1-degree

Results for museum dataset

Random : same number of transactions with random items





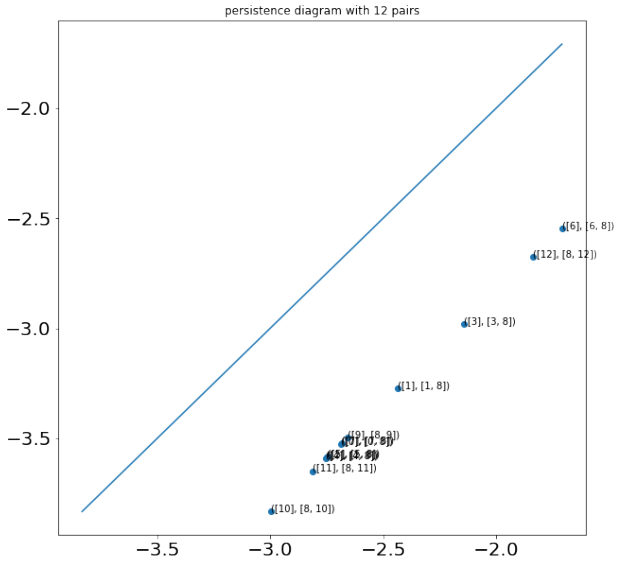
0-degree

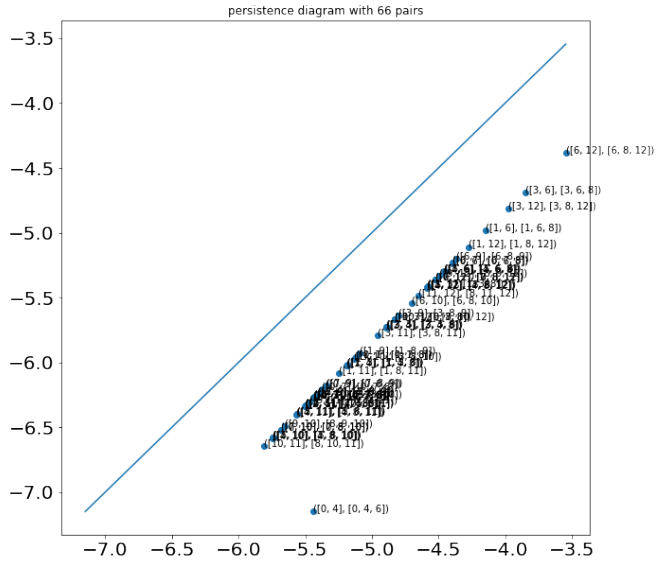
1-degree

POLITECNICO DI TORINO

Results for museum dataset

independent : itemset probability = item probability multiplied





0-degree

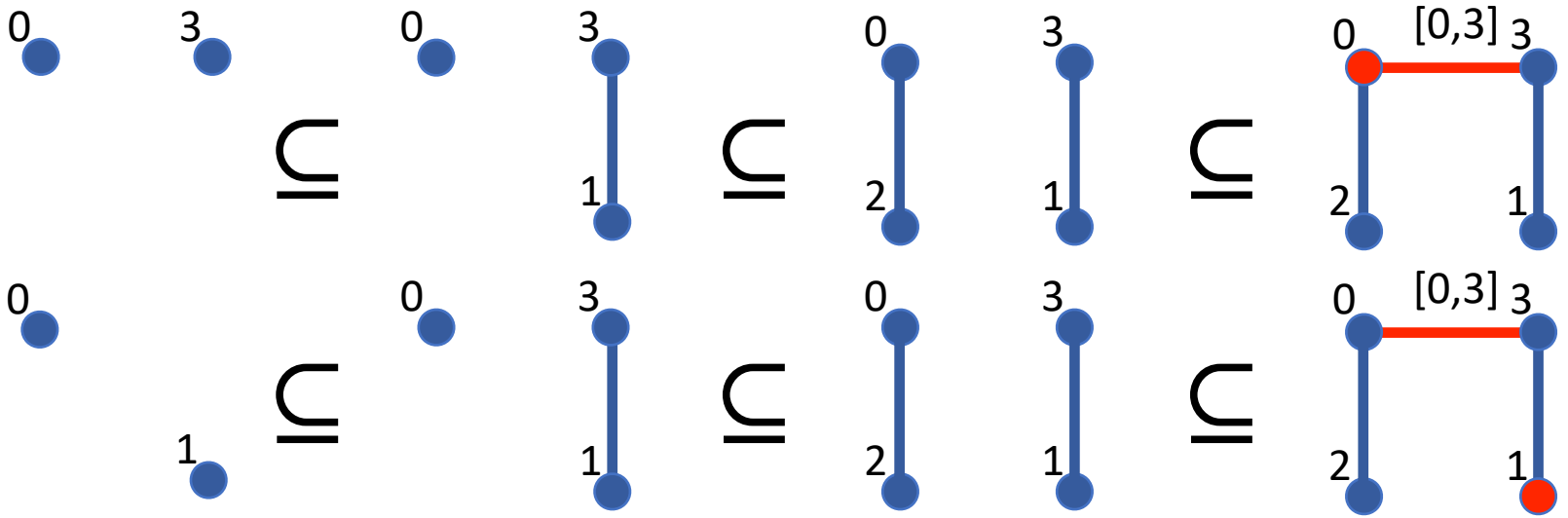
1-degree

DISMA Dipartimento di Scienze Matematiche G. L. Lagrange
ECCELLENZA 2018 · 2022

Correspondence PH - MBA

all our persistence pairs are incident
e.g., (0,[0,3])

in general it might be
(1,[0,3])

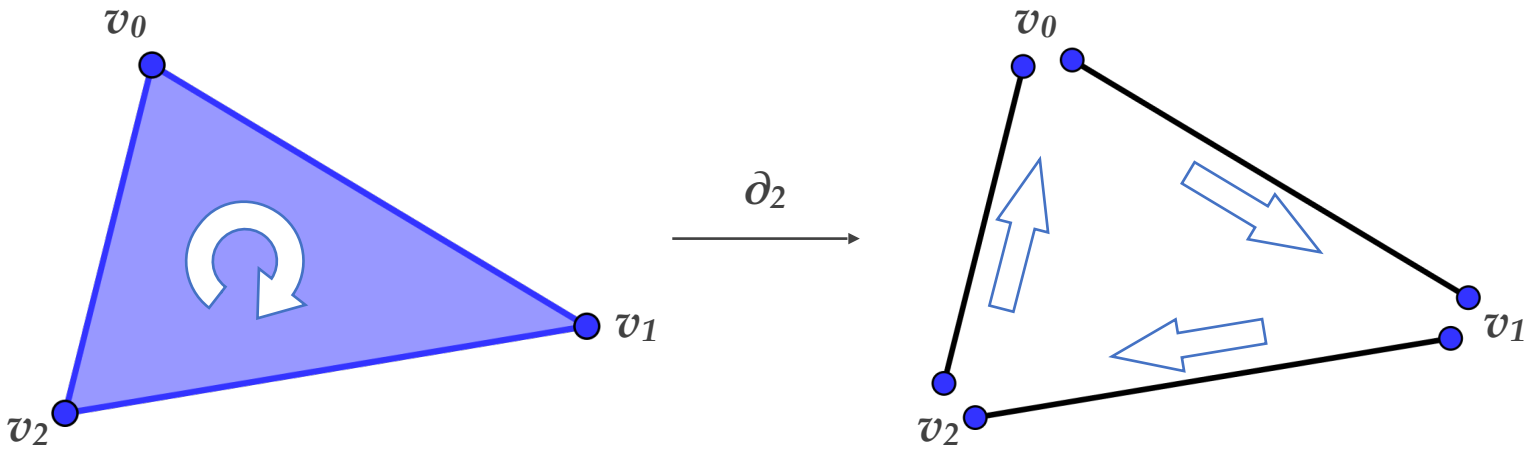


Homology

Given a (finite) simplicial complex K,

- a ***k-chain*** is a formal sum (with $\mathbb{Z}$ coefficients) of (oriented) k-simplices of K
- **$C_k(K)$** is the ***group of the k-chains of K***
- the ***boundary map*** **$\partial_k : C_k(K) \longrightarrow C_{k-1}(K)$** is defined as

$$\partial_k([v_0, \ldots, v_k]) := \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

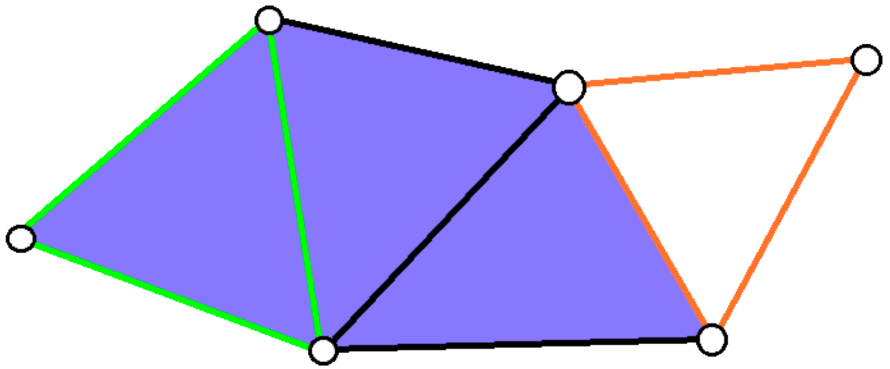


Homology

Given a (finite) simplicial complex K,

- a ***k-chain*** is a formal sum (with $\mathbb{Z}$ coefficients) of (oriented) k-simplices of K
- **$C_k(K)$** is the ***group of the k-chains of K***
- the ***boundary map*** **$\partial_k : C_k(K) \longrightarrow C_{k-1}(K)$** is defined as

$$\partial_k([v_0, \ldots, v_k]) := \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$



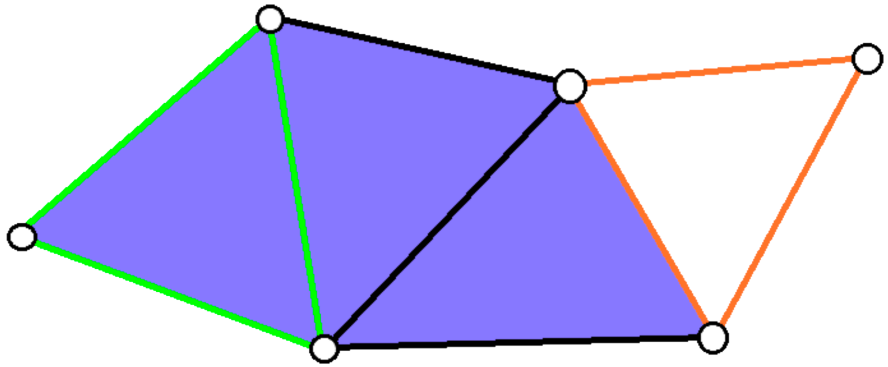
A k-simplex σ is called:

- ***k-cycle*** if $\sigma \in \mathrm{Ker}(\partial_k)$
- ***k-boundary*** if $\sigma \in \mathrm{Im}(\partial_{k+1})$

Homology

Given a (finite) simplicial complex K, the ***k-homology group $H_k(K)$*** of K is defined as

$$H_k(K) := Z_k(K)/B_k(K)$$



where:

- ***$Z_k(K)$*** is the ***group of k-cycles*** of K
- ***$B_k(K)$*** is the ***group of k-boundaries*** of K