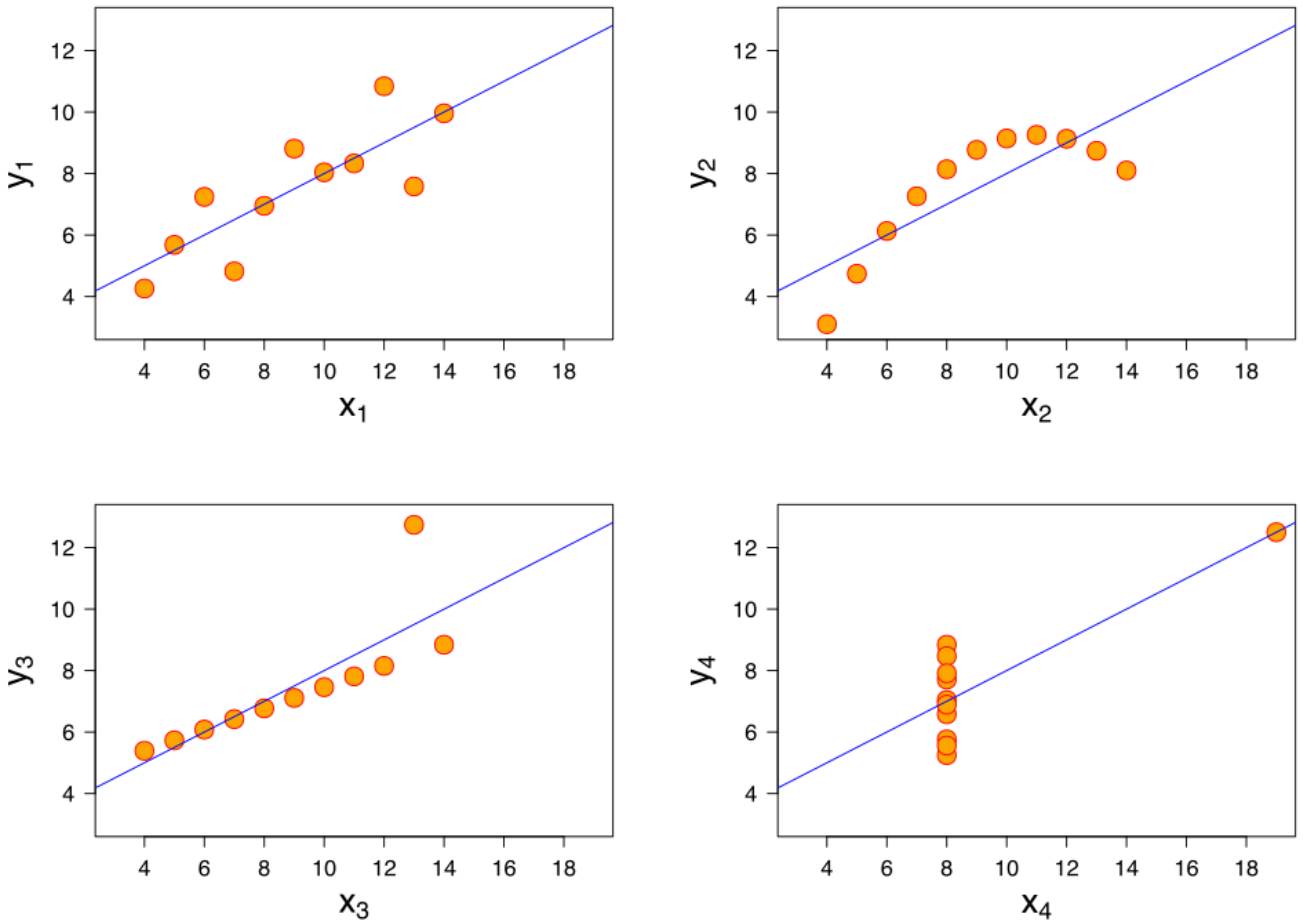
Visualizations and Features in Topological Data Analysis

Paweł Dłotko
Dioscuri Centre in TDA, Warsaw, Poland
Geometric realization of AATRN,
IMSI, Chicago, 22 Aug. 2025.

Motivation, Anscombe's Quartet



Same statistics, different shapes

MeanX 9.00, MeanY 7.50, VarianceX 11.00, VarianceY 4.12,
Correlation (x, y) 0.816, Linear Regression$y = 3 + 0.5x$

Key messages

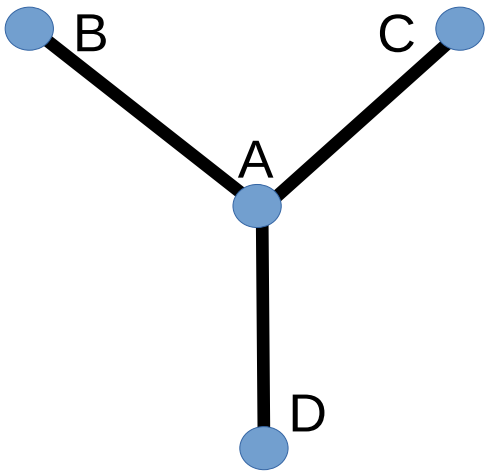
- Always visualize your data! But, how?
- Need for a better features / statistics of data.

Topological visualization

Competitors

- **PCA** aims to find linear subspace on which data has maximal variance
- **UMAP** non-linear dimension reduction attempting to preserve both local and global structure
- **T-SNE** non-linear dimension reduction attempting to preserve mostly local structure
- **PHATE** diffusion embedding technique to preserve local and global structure
- All methods except first assumes that the **data are sampled from a manifold**
- The **global layout** depend on a **seed of the method**
- All returns **embedding of the input space into Euclidean space**

Spaces that cannot be isometrically embedded

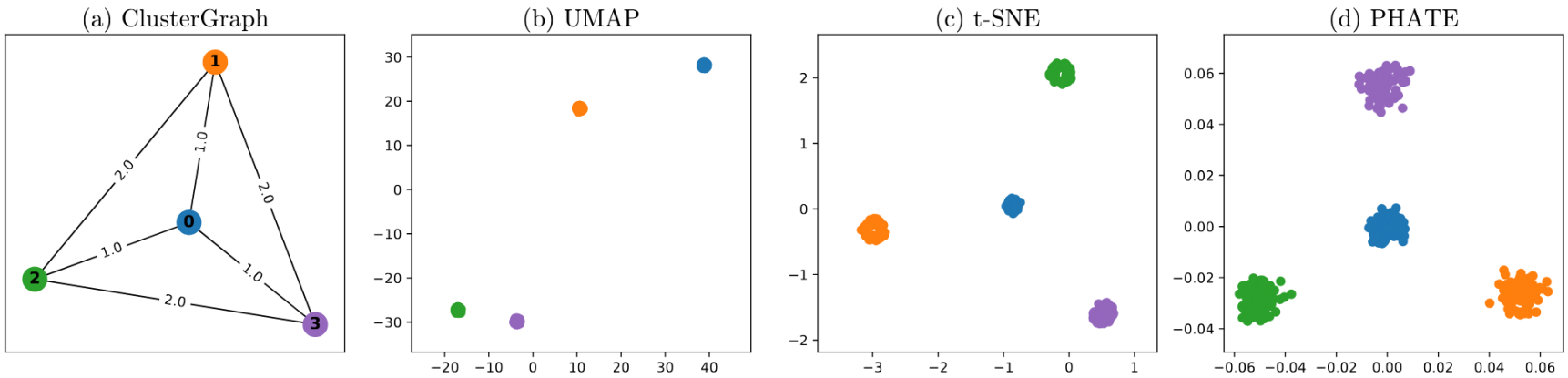


	A	B	C	D
A	0	1	1	1
B	1	0	2	2
C	1	2	0	2
D	1	2	2	0

An *isometric embedding* of a metric space (X, d_X) to $\mathbb{R}^n$ is a map $f : (X, d_X) \to \mathbb{R}^n$ such that for all $x_1, x_2 \in X$, the following condition holds:

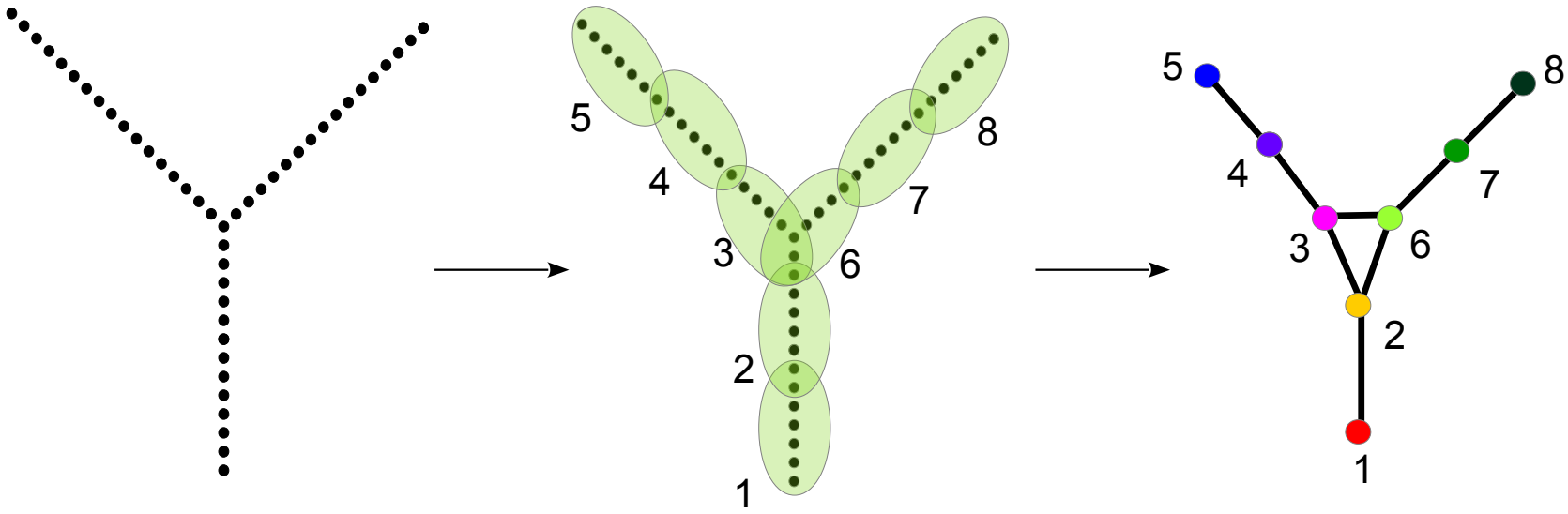
$$d_Y(f(x_1), f(x_2)) = d_{\mathbb{R}^n}(x_1, x_2)$$

Spaces with no isometric embedding to $\mathbb{R}^n$



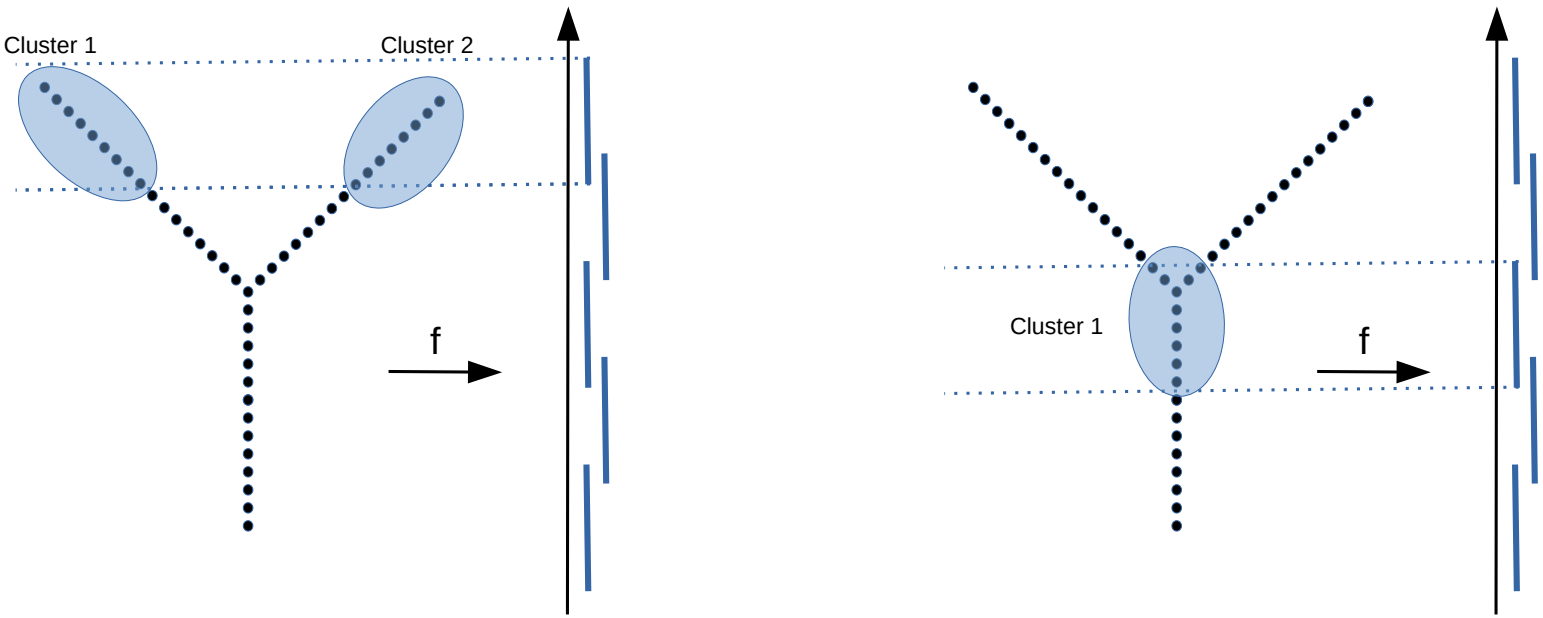
In such cases, topological visualization tools hold a distinct advantage as they produce a graph representation of the dataset rather than an embedding to $\mathbb{R}^n$ for some n.

Mapper–type algorithms



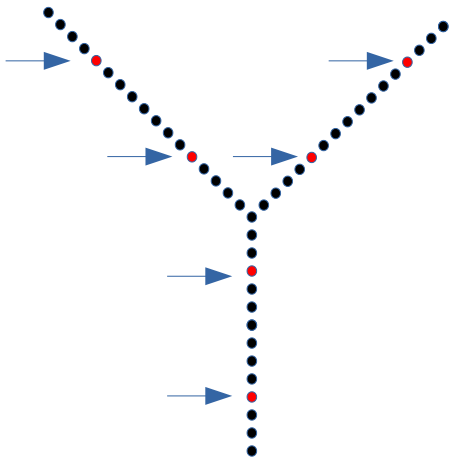
How to obtain an overlapping cover?

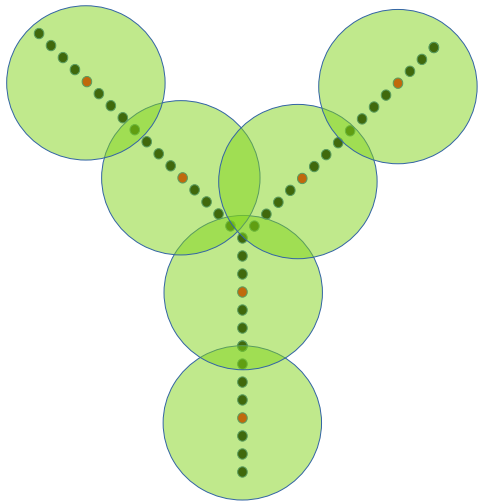
Conventional mapper



Hyperparameters: lens, resolution, gain, clustering method

Ball mapper, ϵ



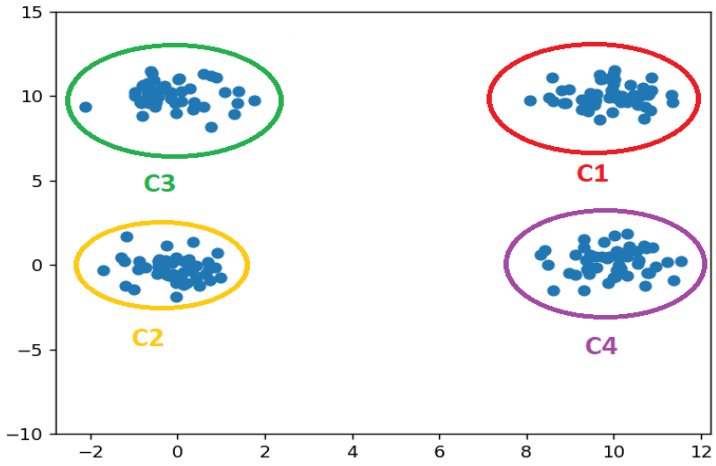


Hyperparameters: ϵ, metric

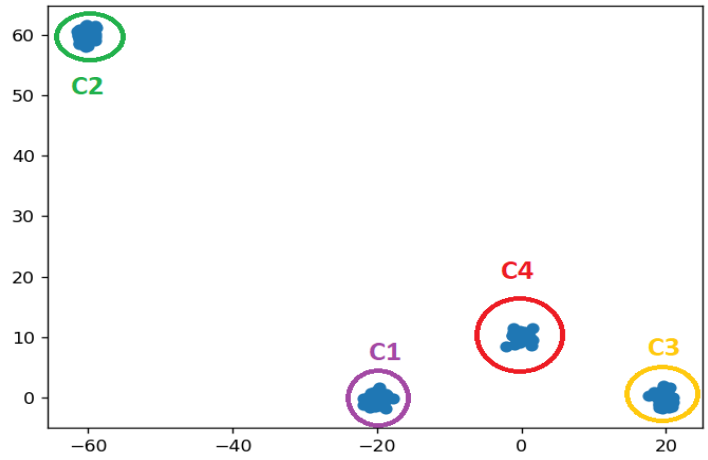
$Y \subset X$ such that for every $x \in X$ there exist $y \in Y$ such that $d(x, y) \leq \epsilon$. Course of dimensionality warning!

ClusterGraph

Same number of clusters, different layout



Four Gaussian distributions, square shape

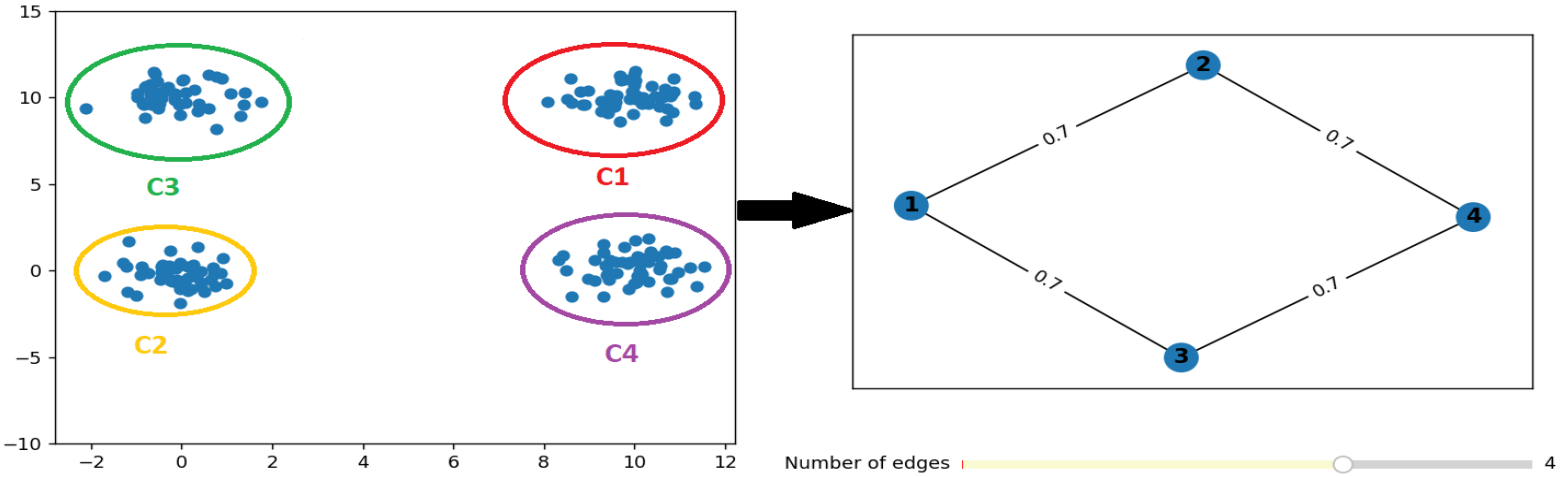


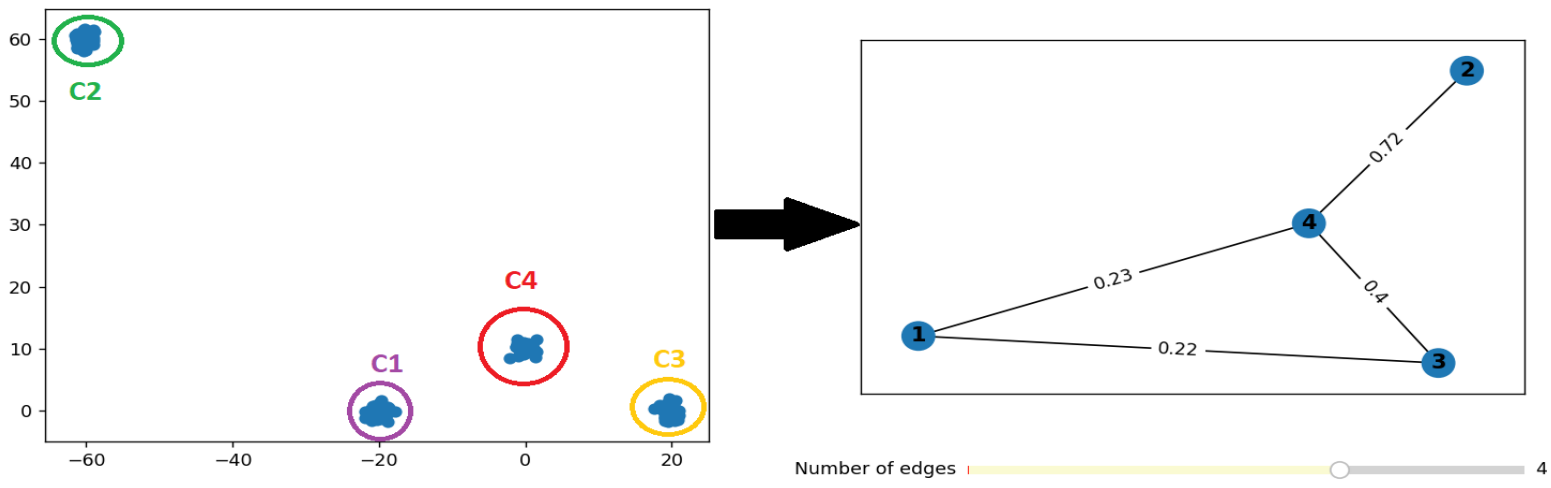
Square shape clusters

Clustering is the process of grouping a set of objects or data points into distinct groups (clusters) where points in the same group are more similar to each other than to those in other groups.

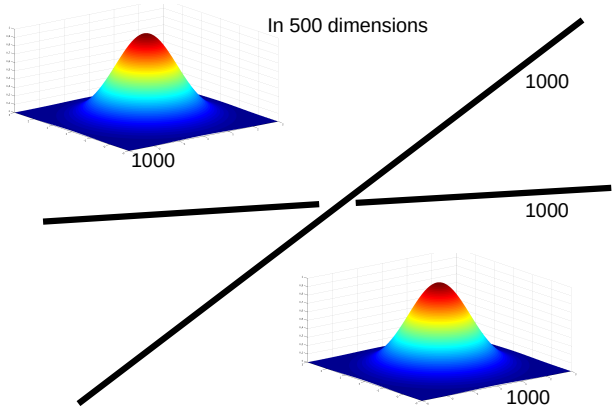
How to understand layout of clusters?

Geometric organization of clusters





ClusterGraph vs competitors, toy example



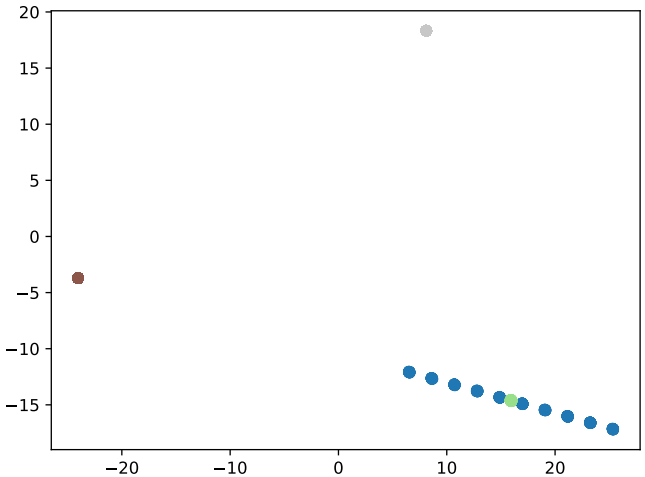


Figure: Gaussian lines PCA, 98% variance kept

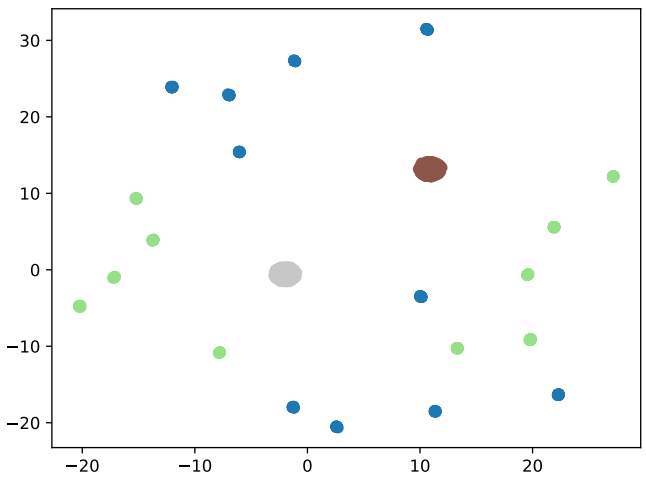


Figure: Gaussian lines Umap

ClusterGraph vs competitors, toy example

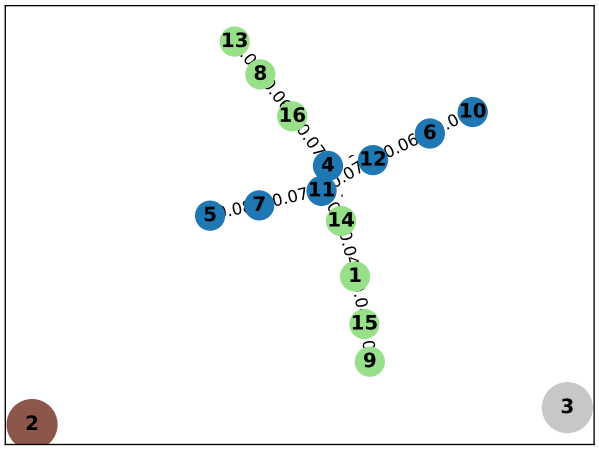


Figure: Gaussian lines ClusterGraph

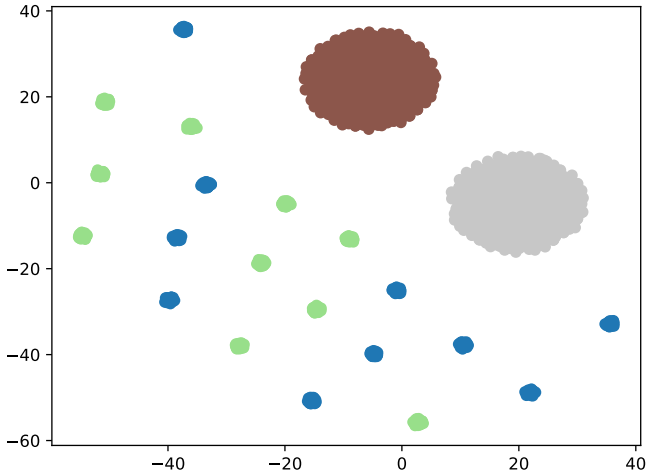


Figure: Gaussian lines T-SNE

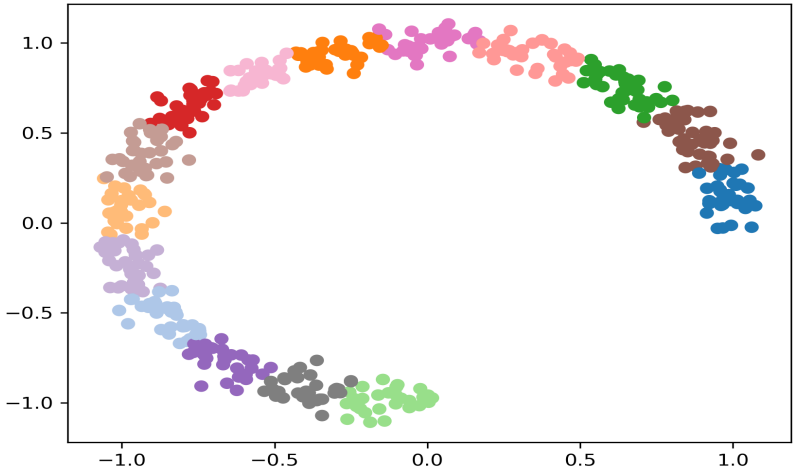
When we do not know the structure of dataset, how can we assess which visualization is better?

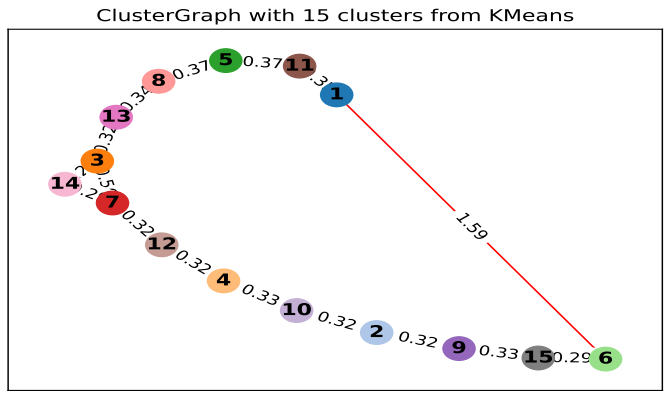
ClusterGraph Metric Distortion between nodes

- We define the Metric Distortion as

$$\delta_{i,j} = \frac{1}{|C_i||C_j|} \sum_{(x,y)\in(C_i,C_j)} |\log\left(\frac{d_{CG}(x,y)}{d_X^k(x,y)}\right)| \qquad (1)$$

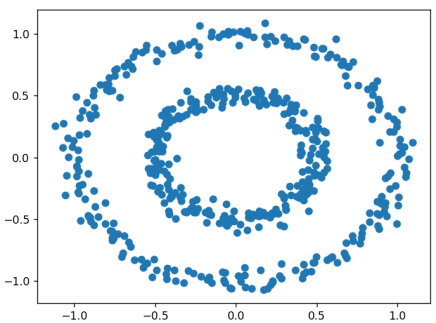
- with X the dataset, n the number of clusters, CG a ClusterGraph
- $d_{CG}(x,y)$ the distance between two points in the ClusterGraph
- $d_X^k(x,y)$ the shortest path between two points in the k-nearest neighbors graph



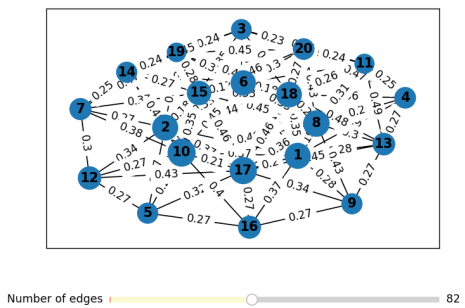


Score improvement

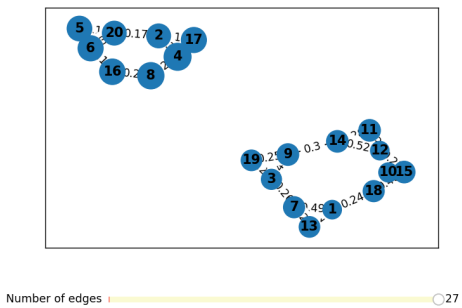
- Quality of the output visualization
- Allow to prune "shortcut" edges providing graph with a better score



(A) Noisy circles



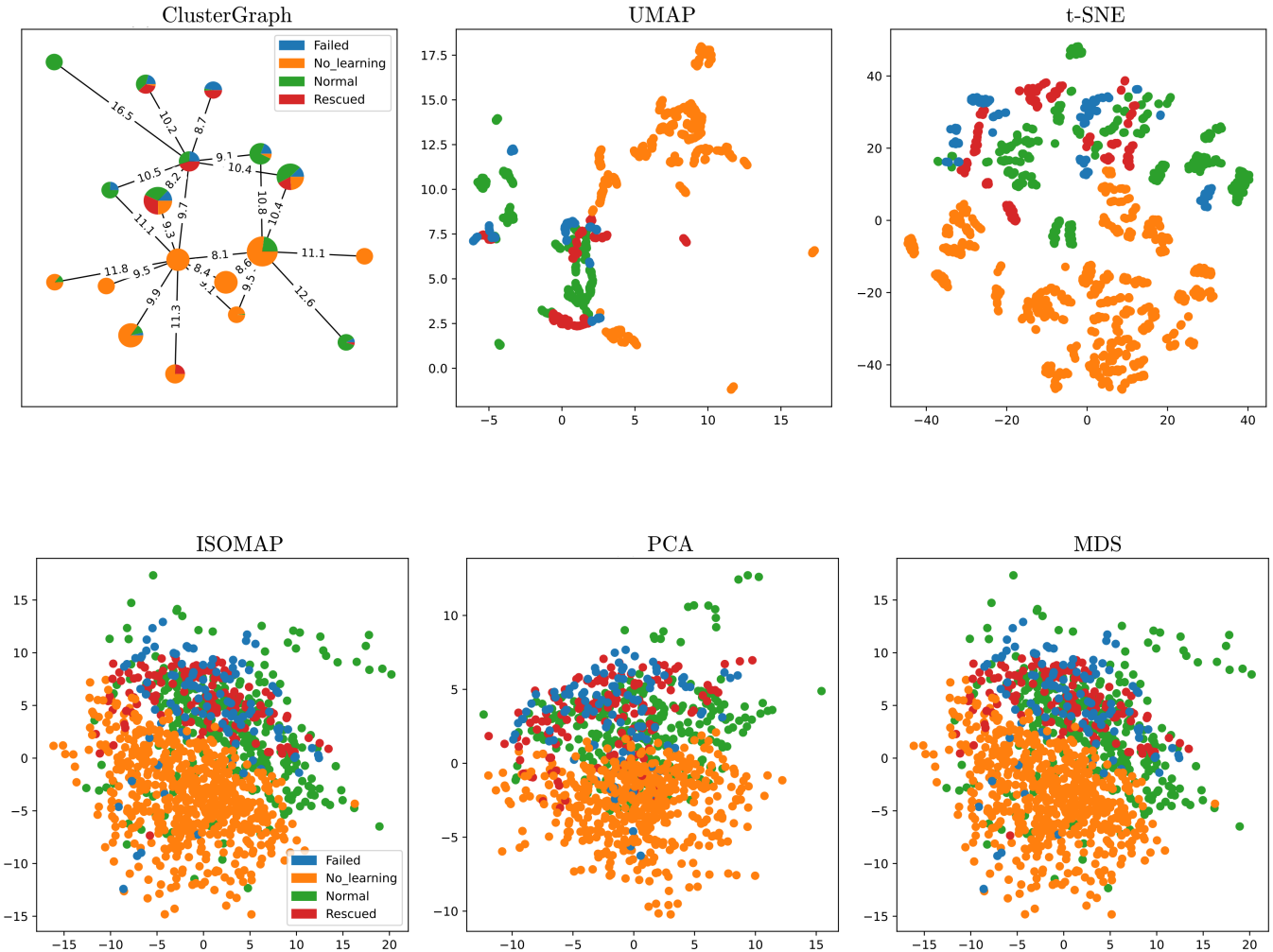
(B) ClusterGraph not pruned



(C) ClusterGraph pruned

Comparison before and after pruning for the noisy circles dataset

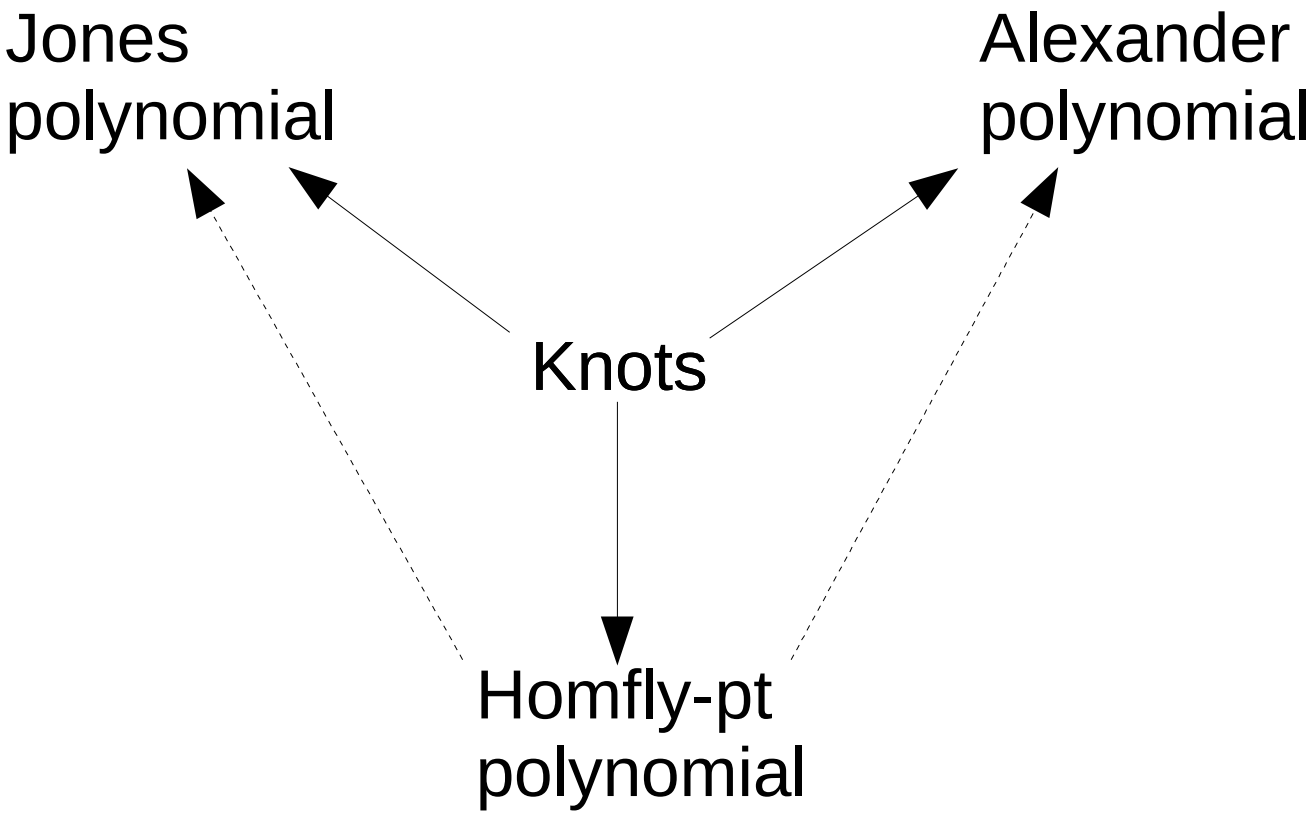
Cluster Graph and friends, trisomic mice dataset



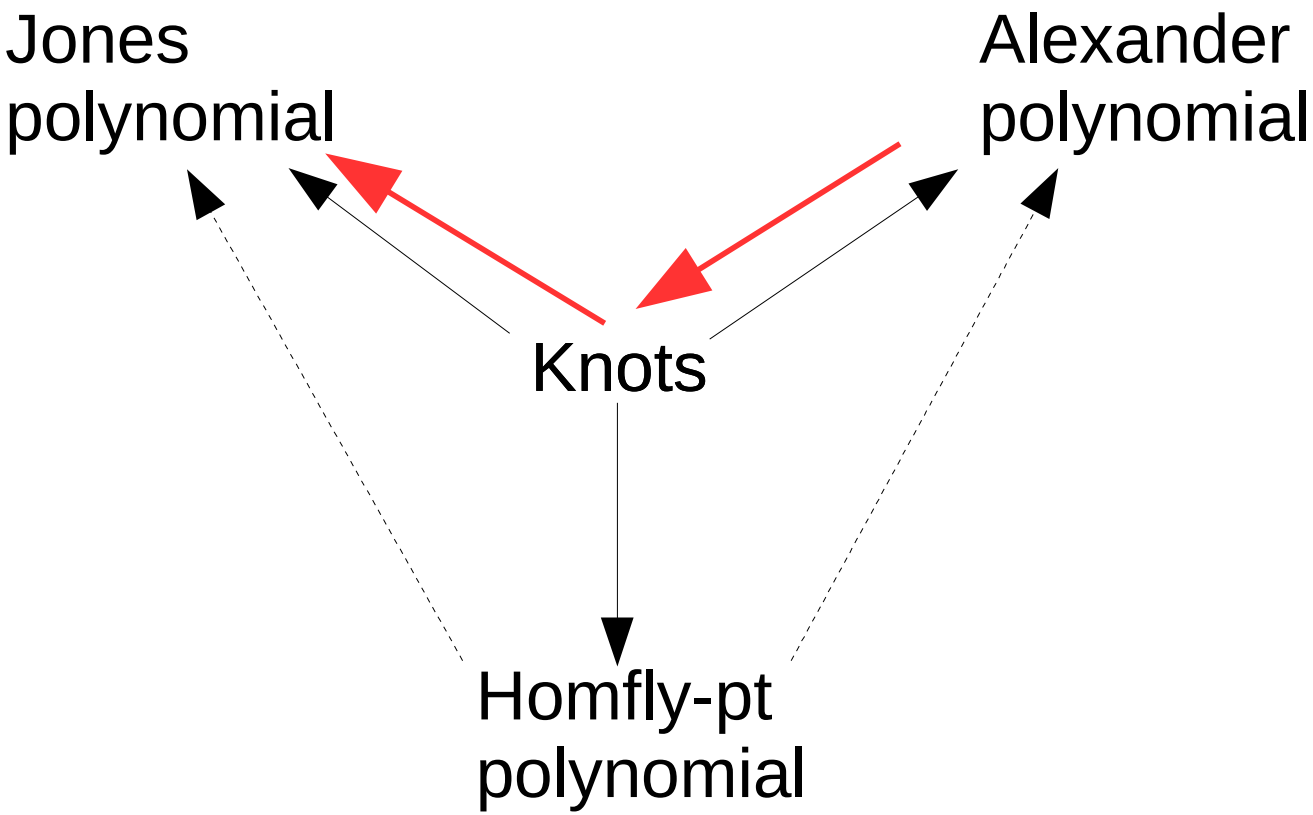
Let us dimensionalize our codomain

- So far mapper-type algorithms have mostly focused on visualizing a point cloud X and $f : X \to \mathbb{R}$
- What about two high dimensional point clouds X, Y and a function, or a relation $f : X \to Y$?

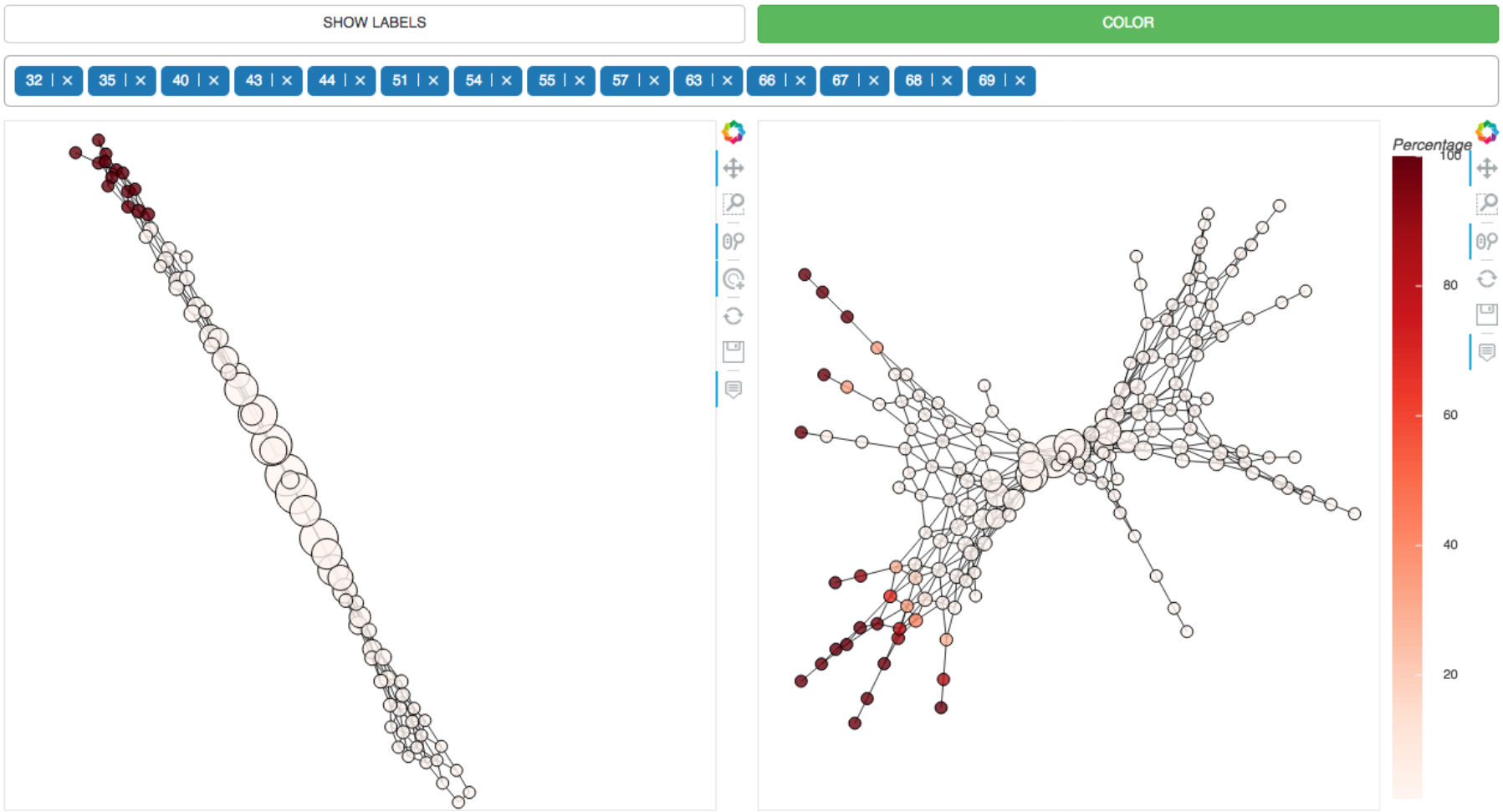
Different features of the same data



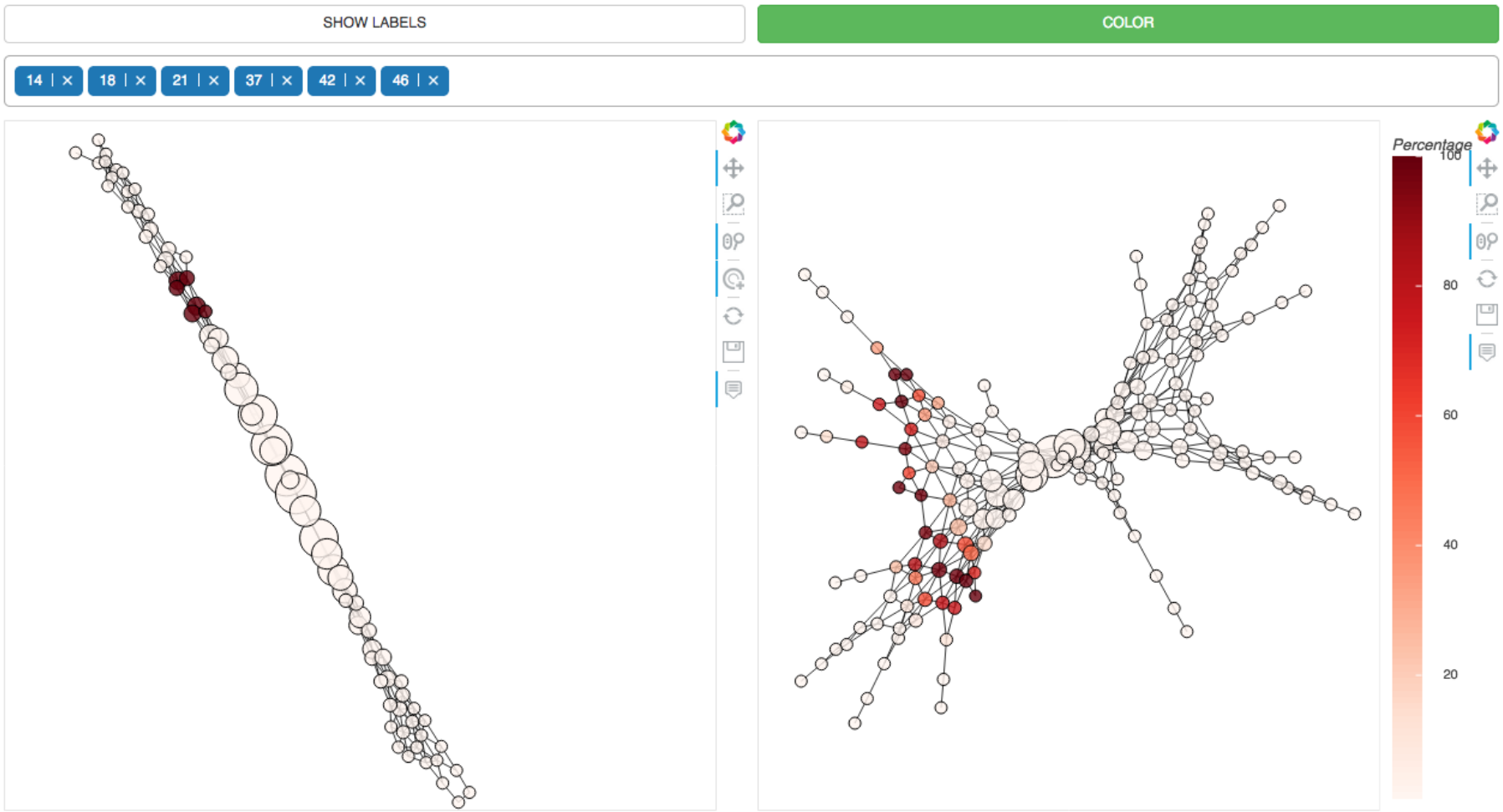
BALL MAPPER: $f : \mathbb{R}^n \supset X \to Y \subset \mathbb{R}^m$



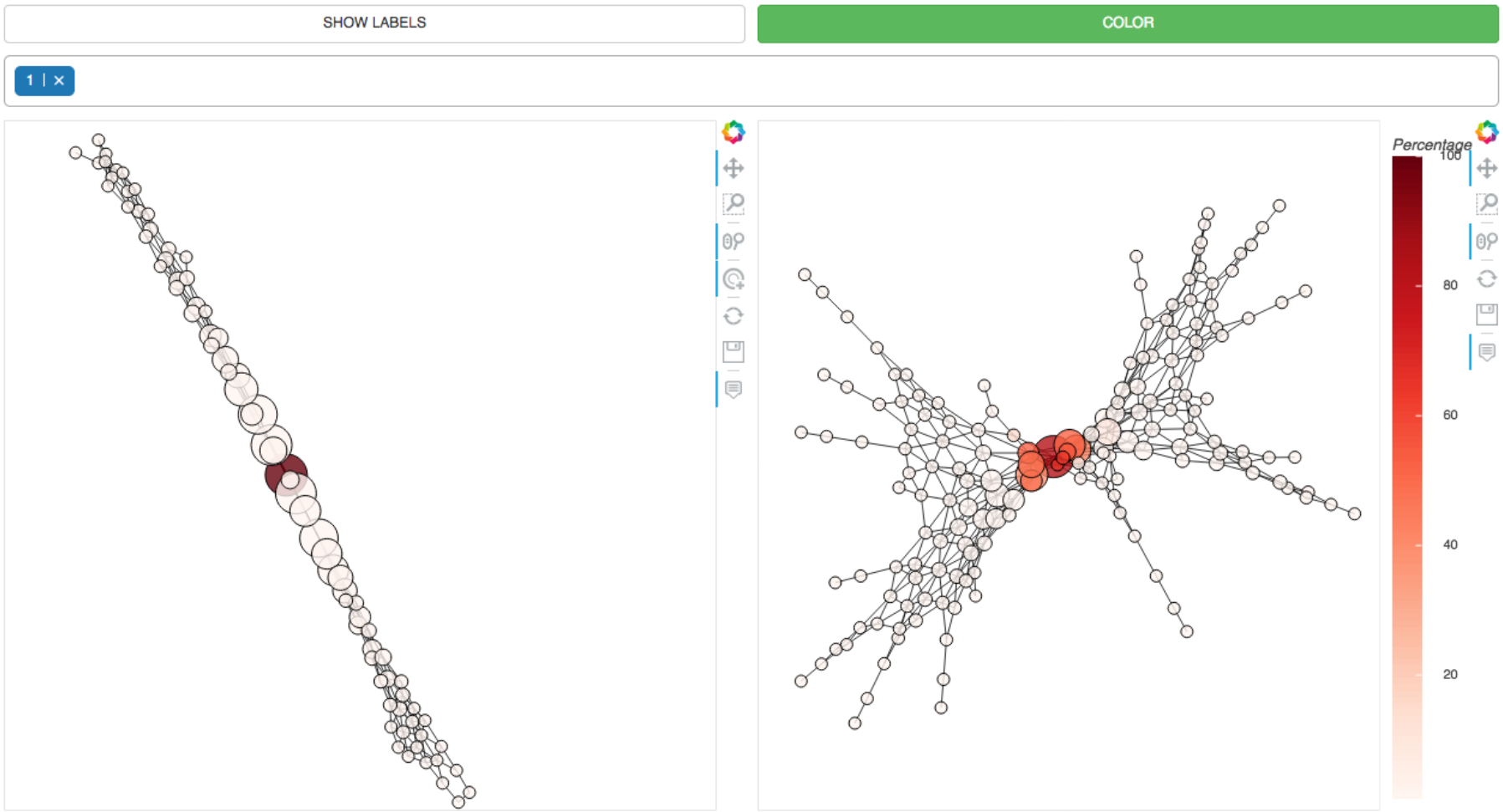
Space of Alexander vs Jones (15 crossings)



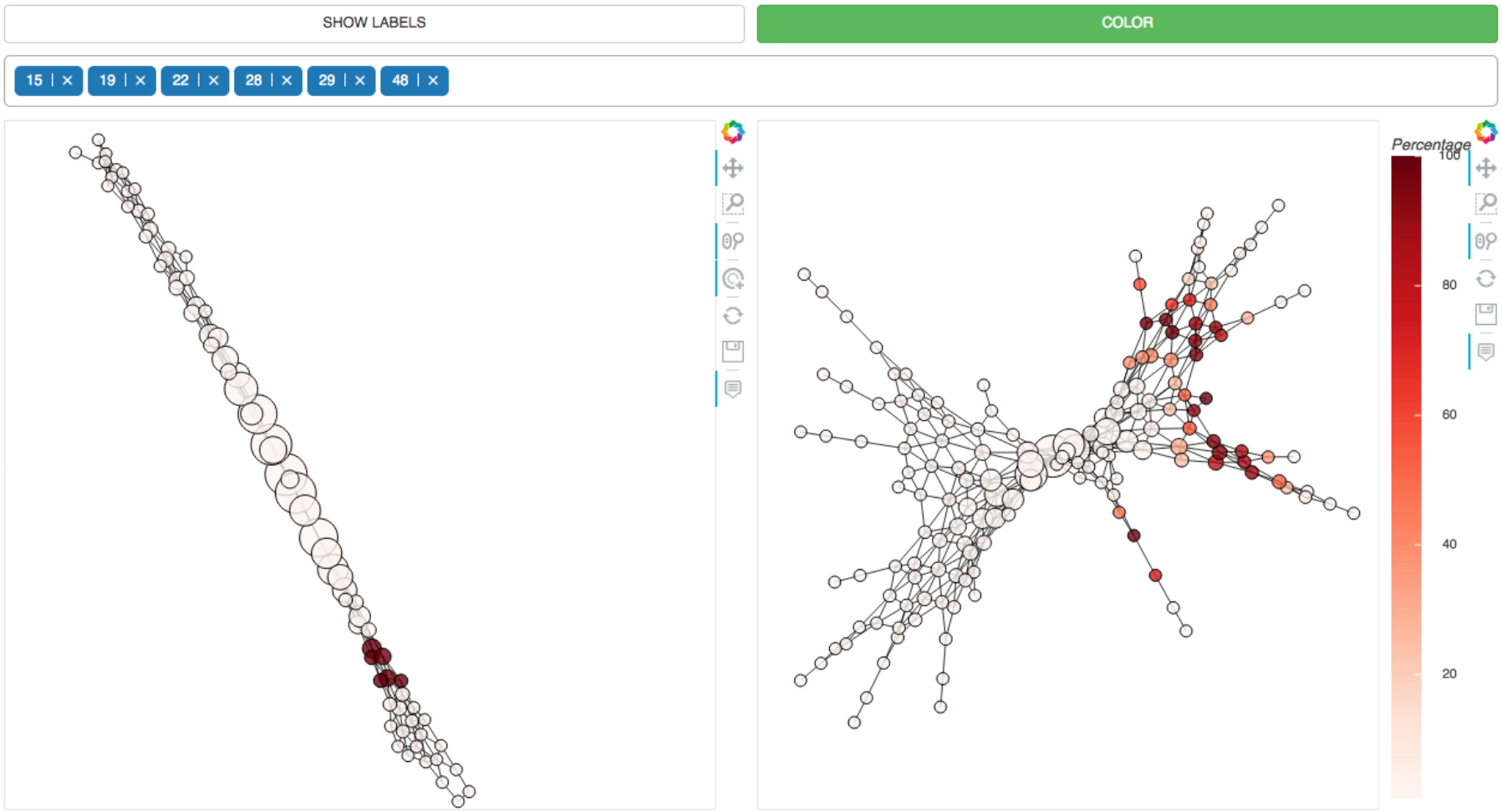
Space of Alexander vs Jones (15 crossings)



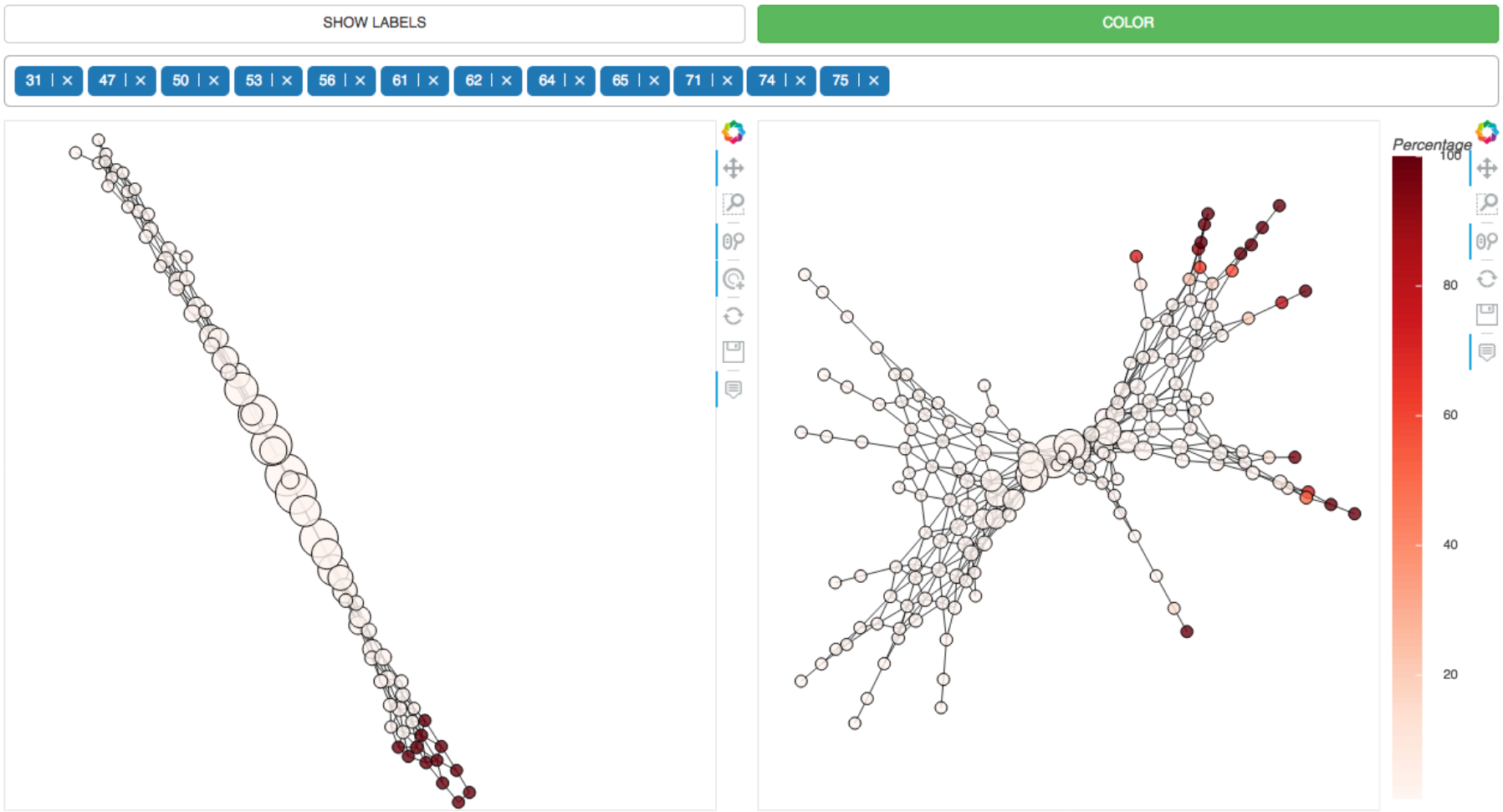
Space of Alexander vs Jones (15 crossings)



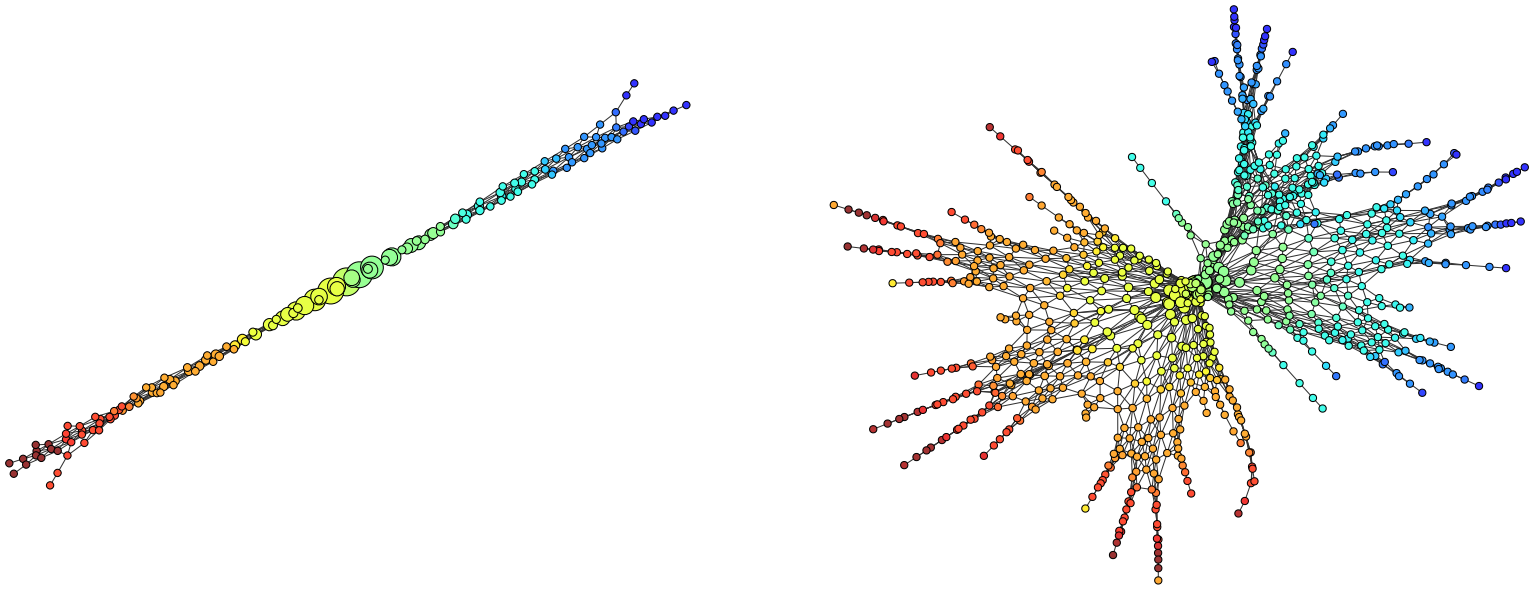
Space of Alexander vs Jones (15 crossings)



Space of Alexander vs Jones (15 crossings)



Map Alexander to Jones



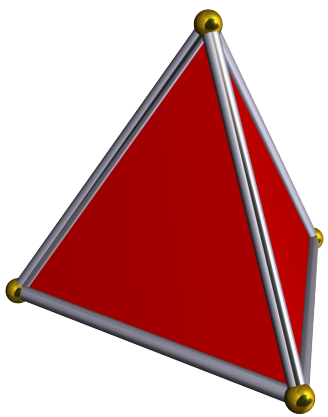
https://dioscuri-tda.org/BallMapperKnots.html

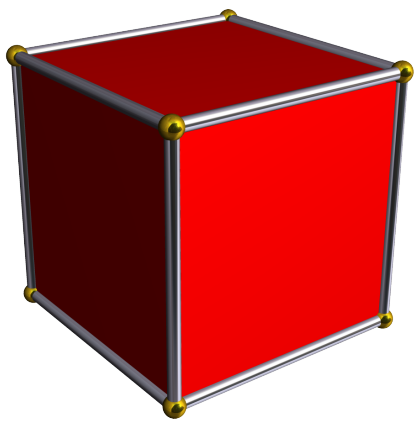
Topological visualization methods

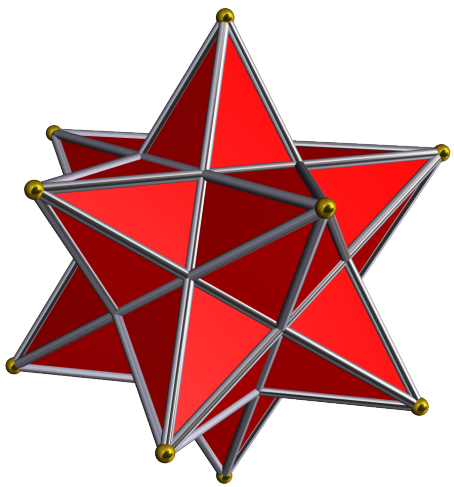
- Ball Mapper,
  - `cran.r-project.org/web/packages/BallMapper`
  - `github.com/dioscuri-tda/pyBallMapper`
  - pip install pyBallMapper
- Cluster graph,
  - `github.com/dioscuri-tda/ClusterGraph`
  - pip install clustergraph
- Quality scores of the representions.

Euler characteristic curves and profiles

$$\chi = V - E + F$$







1758, by Euler

Euler Characteristic - Definition

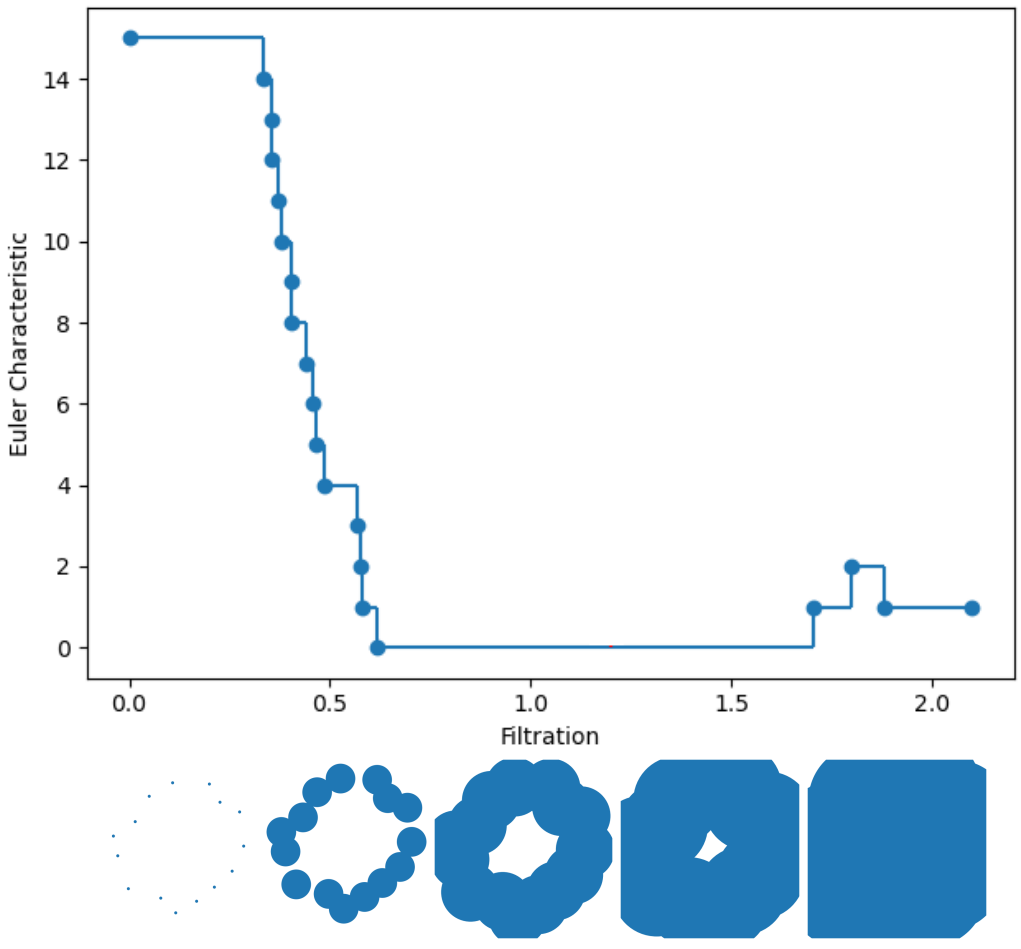
We are interested in computing the **Euler characteristic** of the simplicial complex K defined as

$$\begin{aligned}\chi(K) &= \sum_{n\geq 0}(-1)^n|K_n| \\ &= \sum_{n\geq 0}(-1)^n\beta_n(K)\end{aligned}$$

Euler-Poincare formula

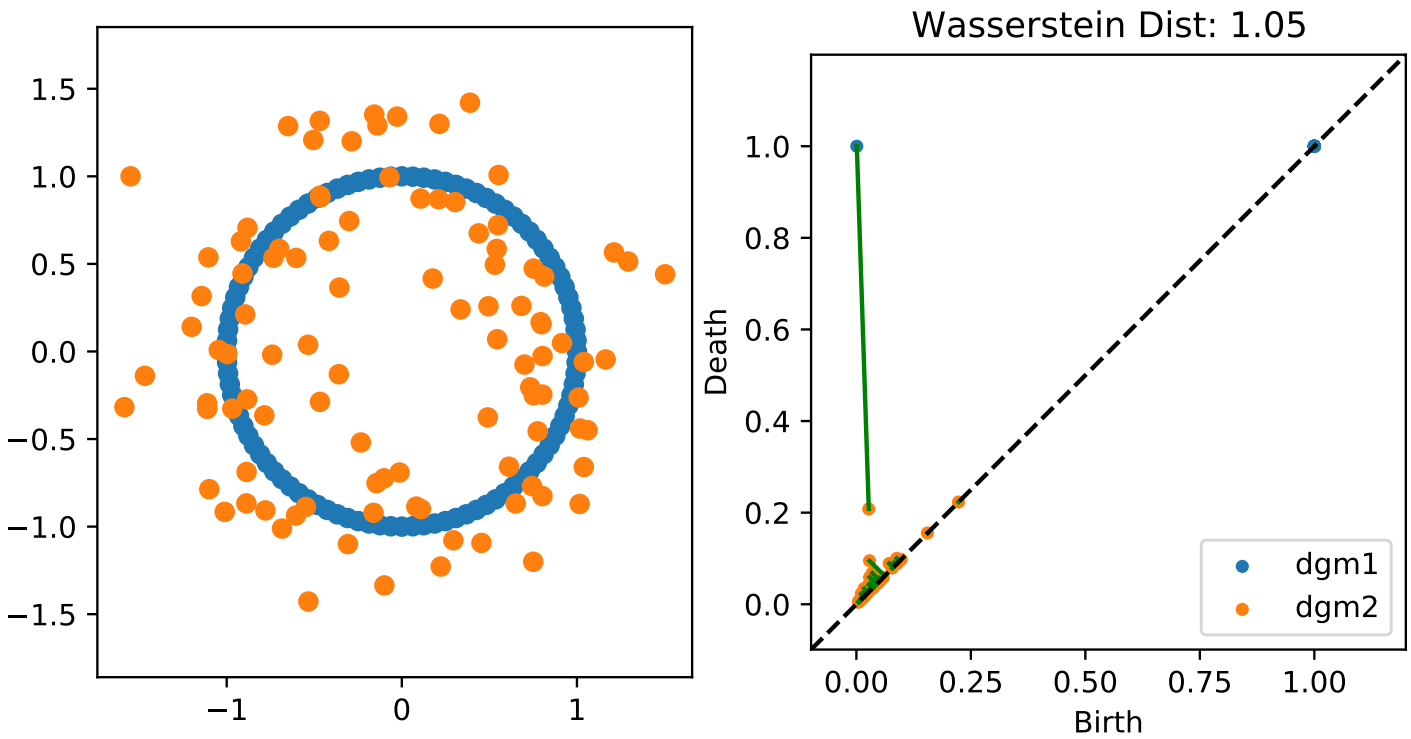
Euler Characteristic Curve

- Let us consider a filtered simplicial complex K with filtration function $f : \mathcal{K} \to \mathbb{R}$.
- It induce $\emptyset \subset \mathcal{K}_1 \subset \mathcal{K}_2 \subset \ldots \subset \mathcal{K}_n = \mathcal{K}$



Robustness, p-Wasserstein stability

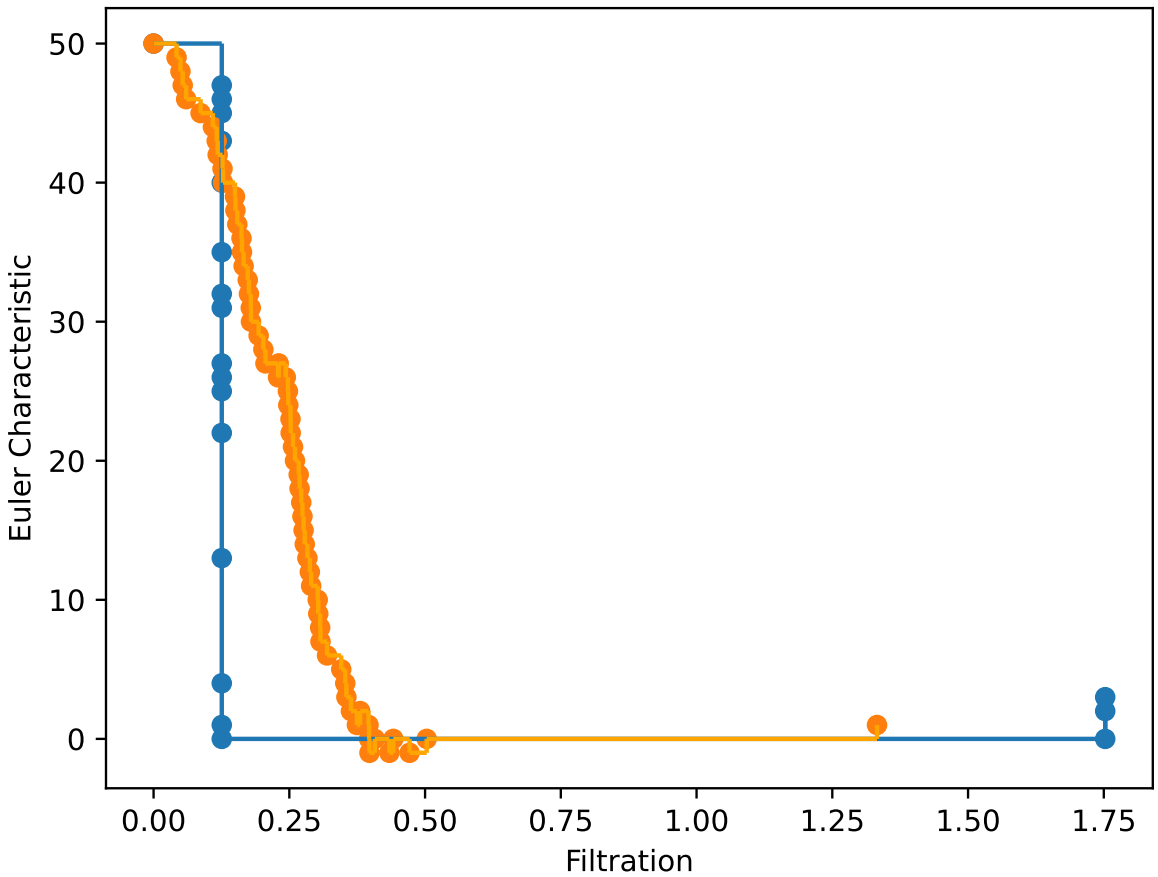
$$W_p(C, D) = \left[\inf_{\eta: C \to D} \sum_{(b,d) \in C} \|(b, d) - \eta(b, d)\|_\infty^p \right]^{1/p}$$



ECC enjoys similar stability property

ECC are stable with respect to the 1-Wasserstein distance:

$$\|ECC(X) - ECC(Y)\|_1 \leq \sum_k 2W_1(Dgm_k(X), Dgm_k(Y))$$

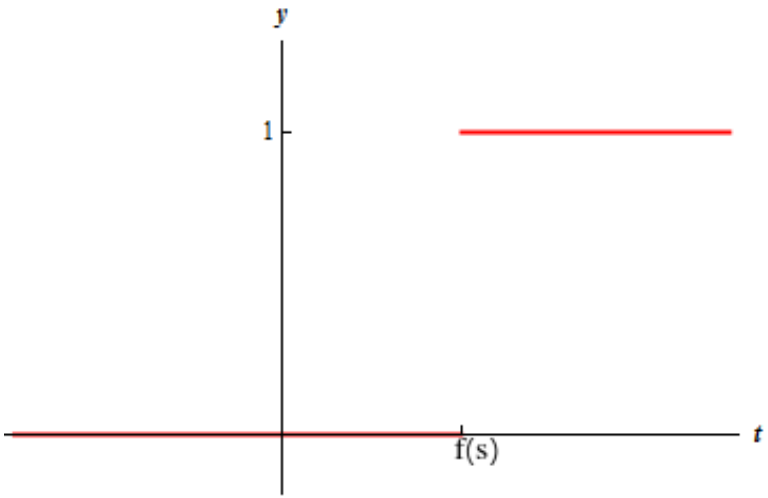


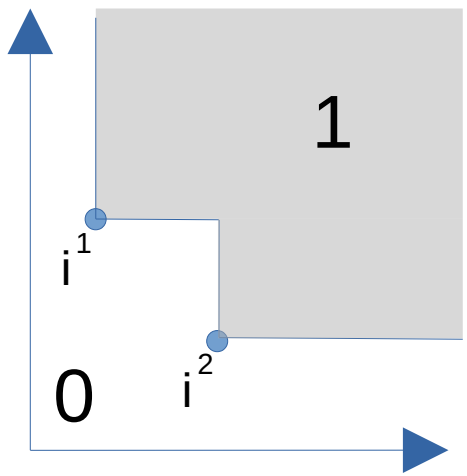
Why to use ECC?

- Easy to compute, pure combinatorics
- Embarrassingly parallel
- Stable
- Present in many theorems from various branches of mathematics
- What about multifiltrations?

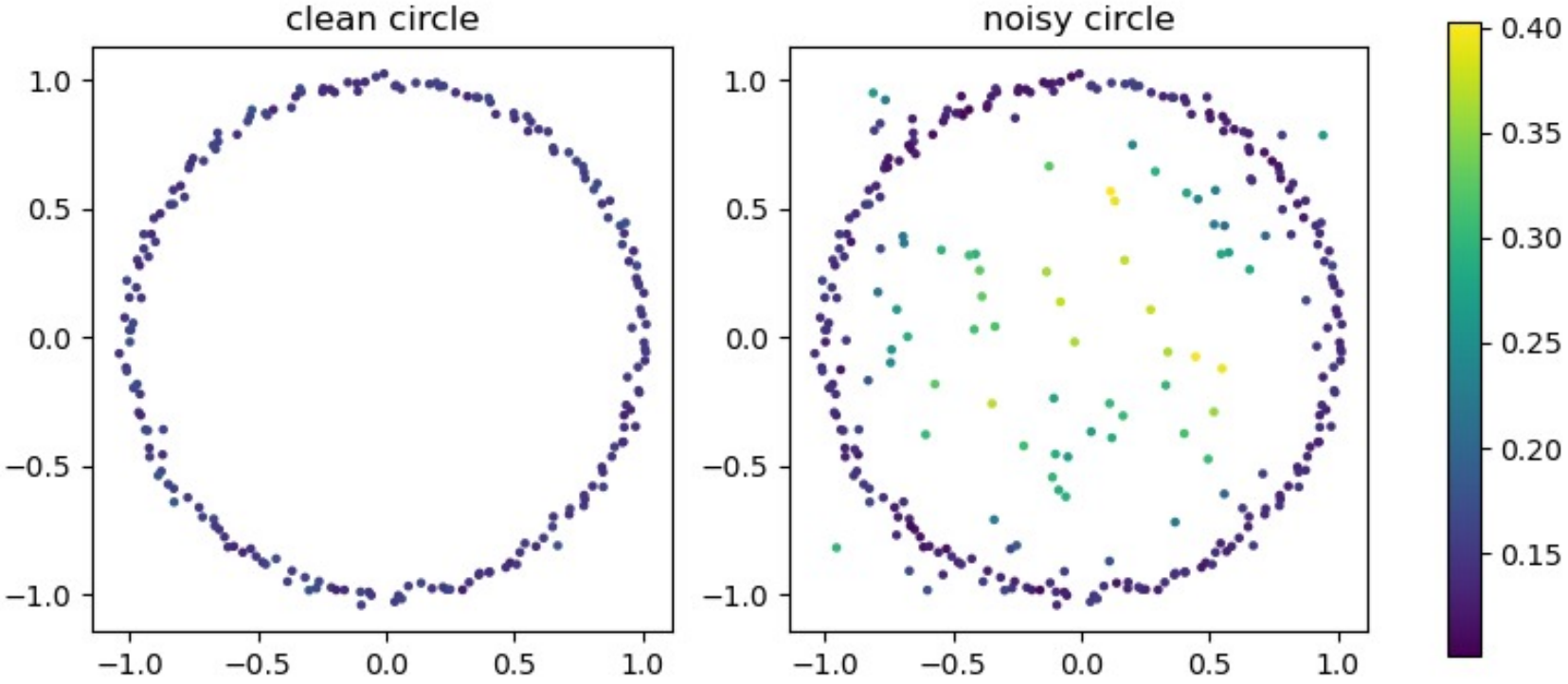
Euler profiles

- ECC is an alternating sum of indicator function on simplices.
- In one dimension it is a Heaviside function,
- For many, its suitable generalization.



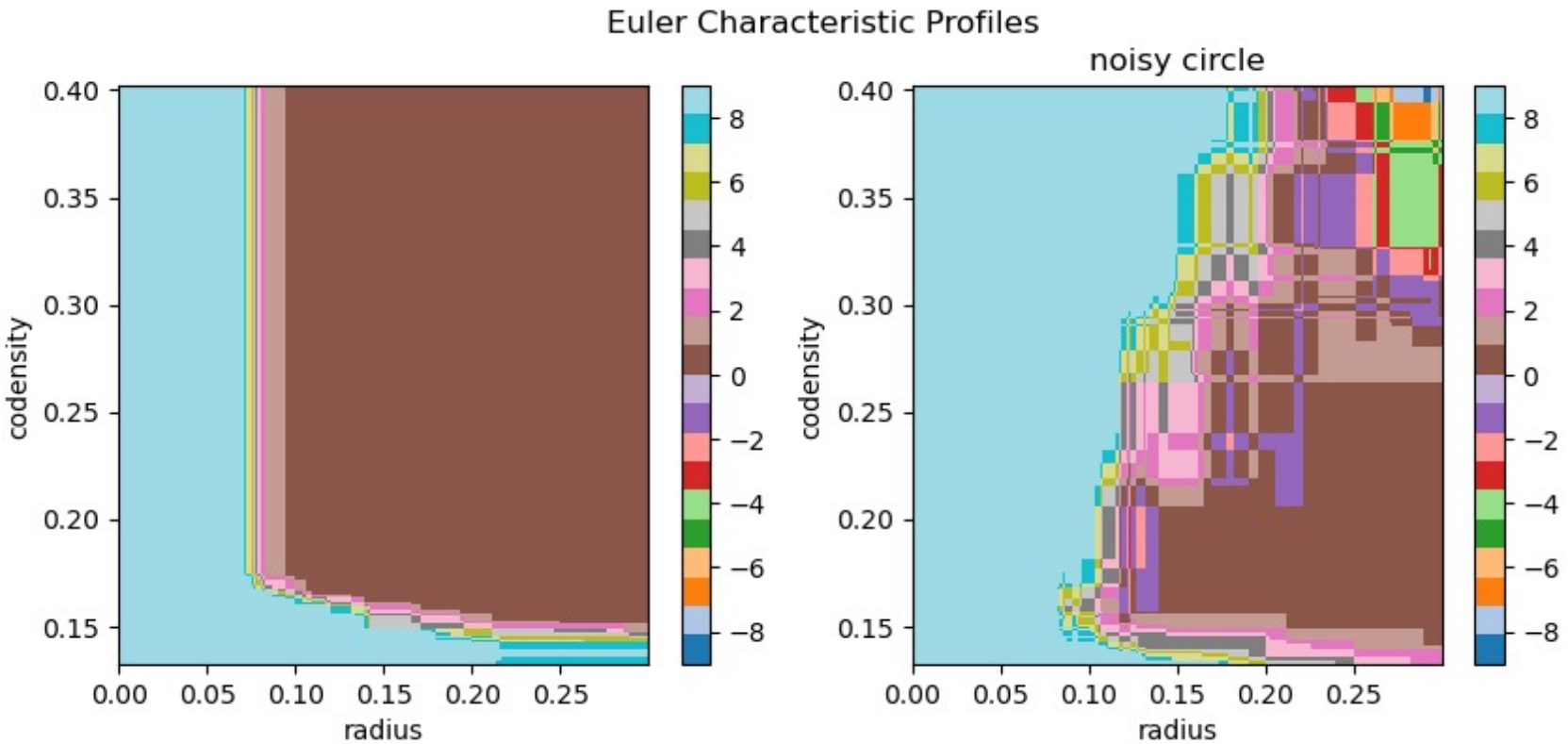


Euler profiles, robustness

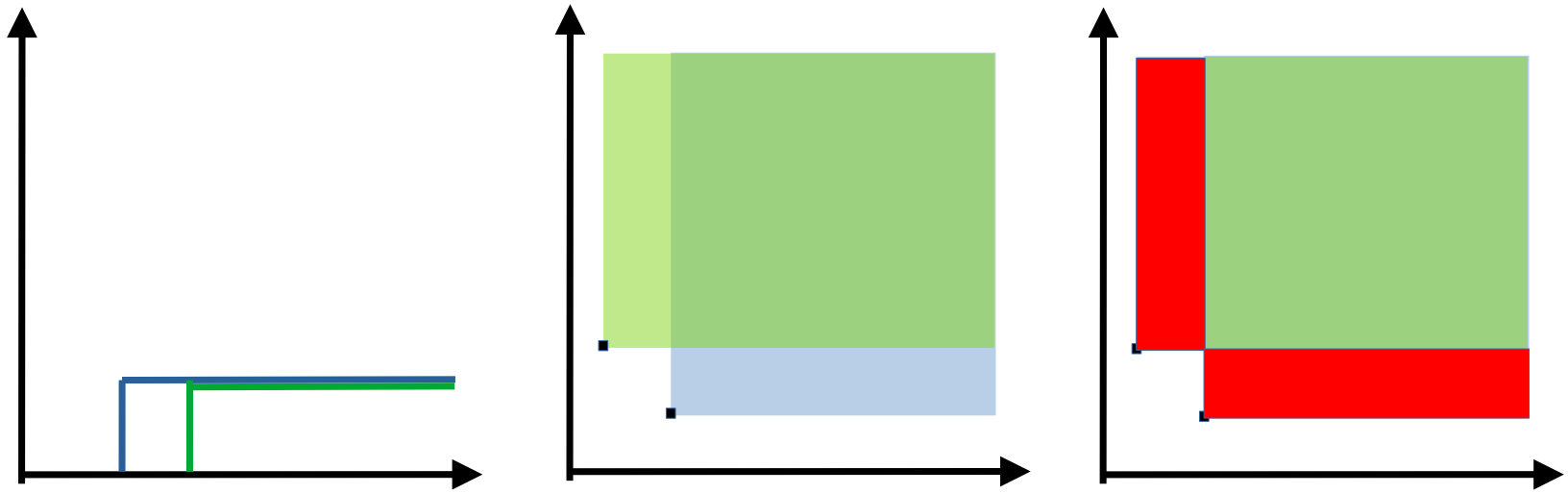


Sampling from unit circle (left) with added salt and pepper noise (colored by distance to kth nearest neighbor, on the right)

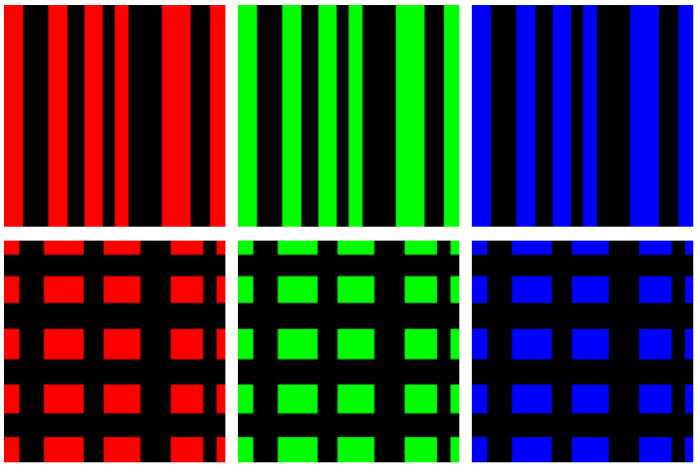
Euler profiles, robustness

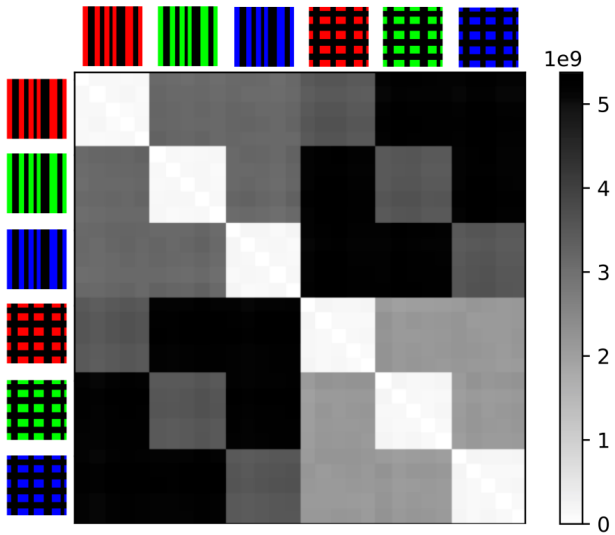


Euler profiles, stabity



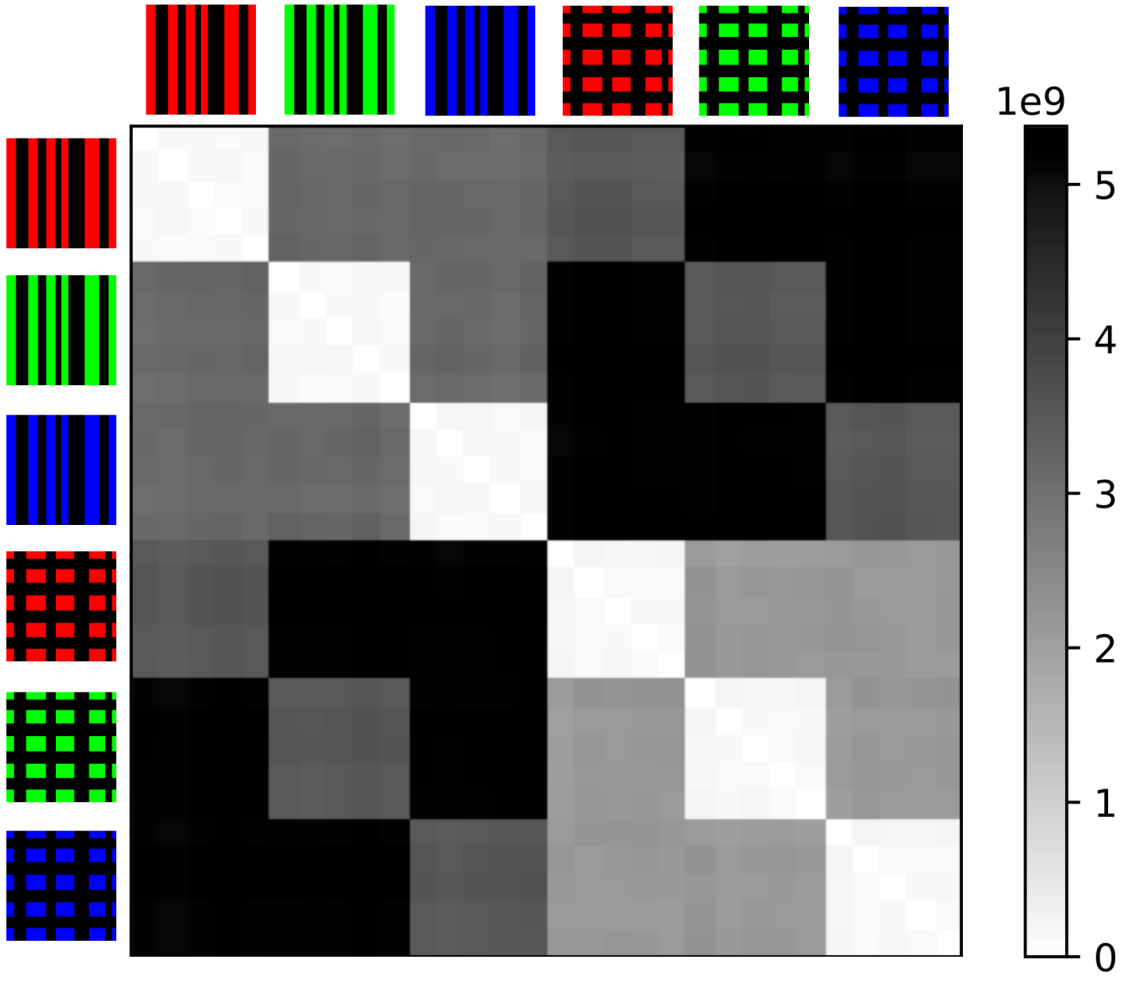
Euler profiles, toy example





Stripes and tartans

Euler profiles, toy example

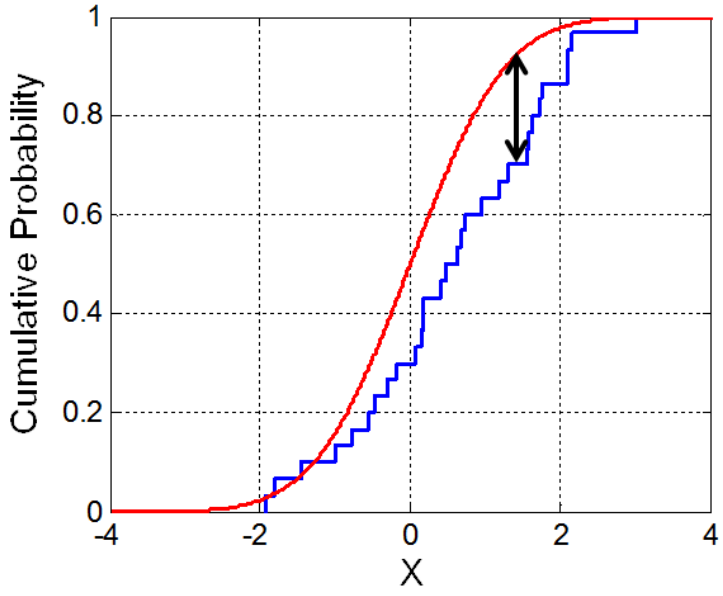


- Fast, parallel and distributed algorithms exists
- No problem with multifiltrations,
- Stable,
-  Code available on `github.com/dioscuri-tda/pyEulerCurves`
- pip install pyEulerCurves

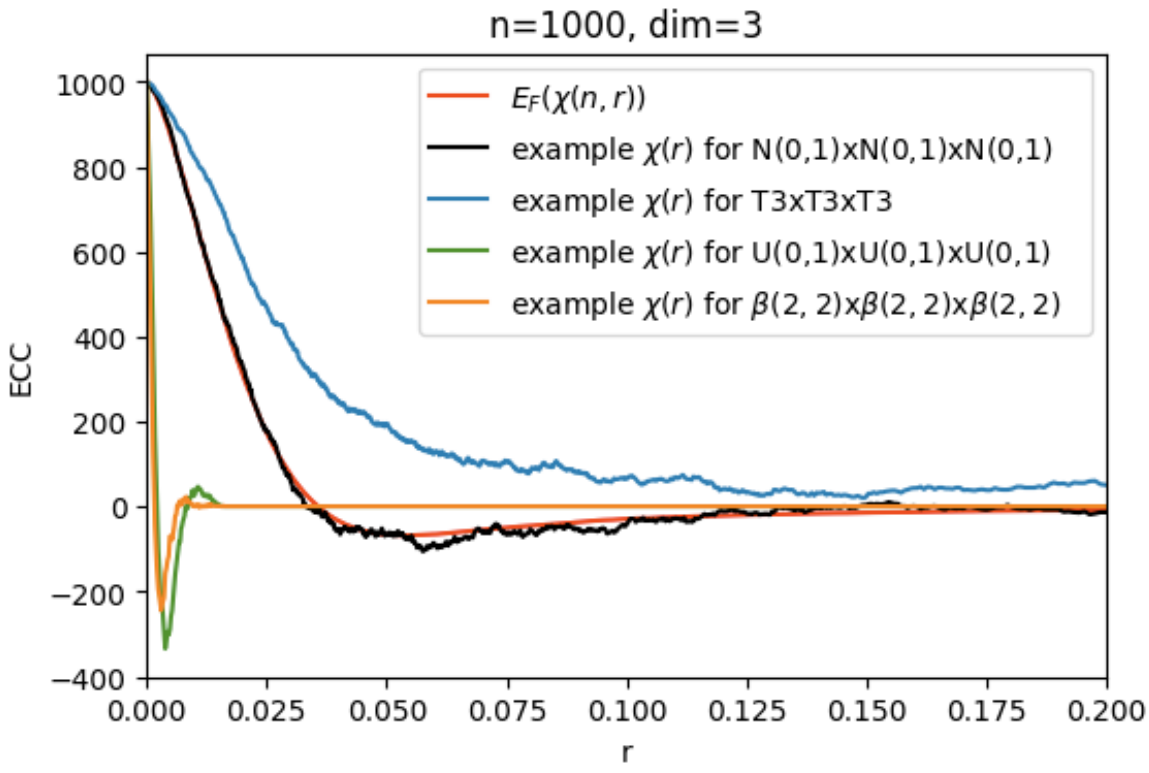
Topotests

- **One-sample problem**: We are given a data sample $X = \{x_1, x_2, \ldots, x_n\}, x_i \in R^d$ and cumulative distribution function $F : R^d \to [0, 1]$. Does the data X follow the distribution F: $X \sim F$?

$$H_0 : X \sim F \quad \text{vs.} \quad H_1 : X \nsim F$$

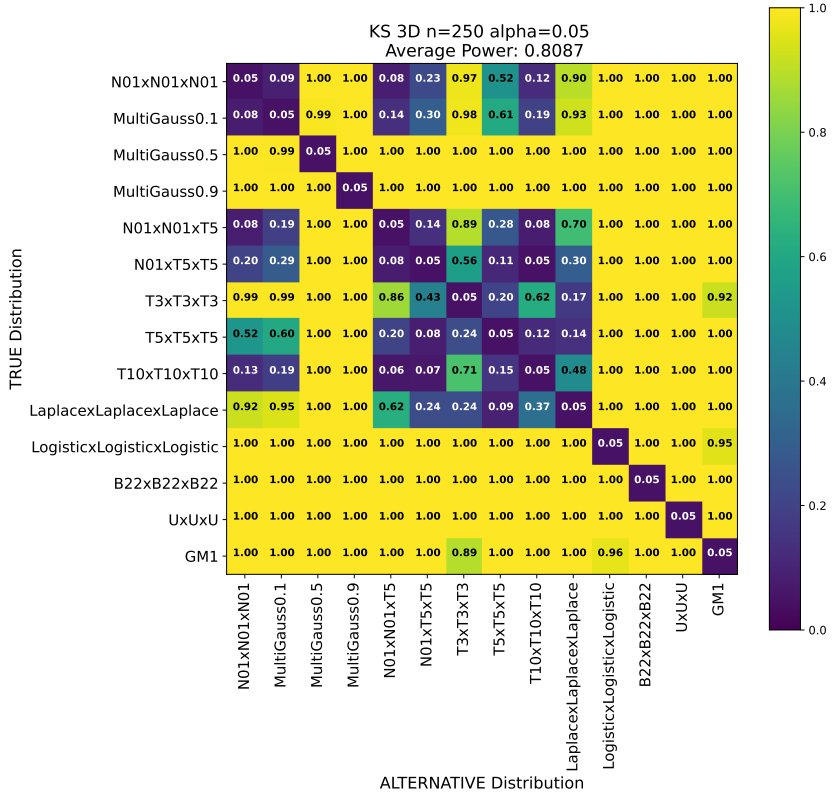


- We use KS as benchmark
- One-sample: $D_n = \sup_x |F_n(x) - F(x)|$
- Compare to tabulated values of the statistics.



- Compute expected ECC for point clouds of size n sampled from F,
- Compute critical values,
- Compute ECC for your sample and its distance to the expected ECC,
- Check if it is closer than the critical value or not.

Simulation results (one-sample)



average power at $\alpha = 0.05$:

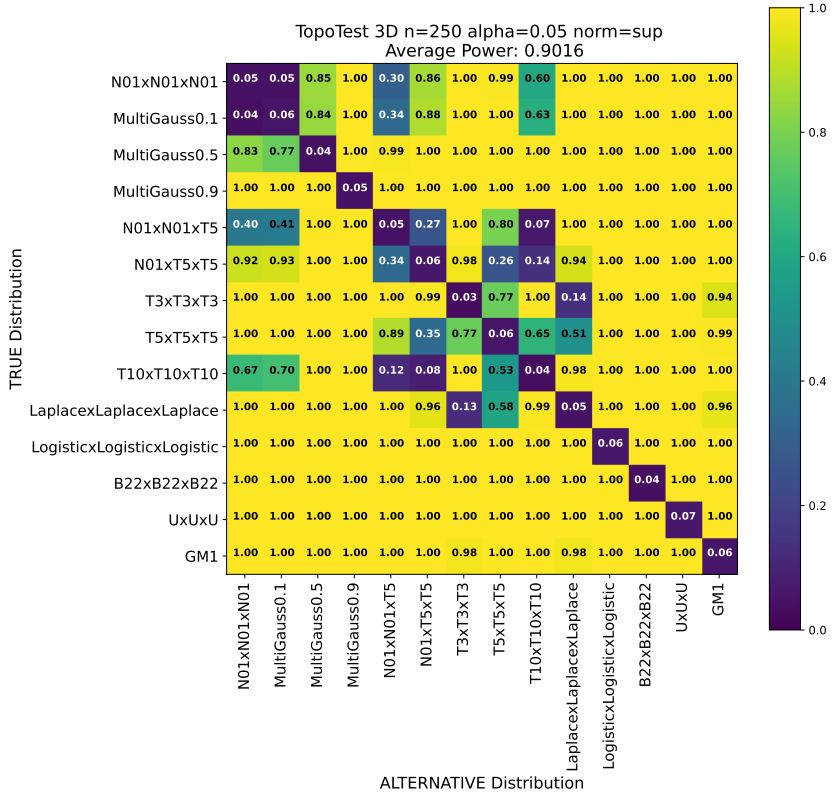
$d = 3, n = 250$ TT:0.9016, KS : 0.8087

$d = 5, n = 500$ TT:0.8465, KS : $- - -$

Test Power: probability that H_0 is correctly rejected when H_1 is true

- Samples sizes 100–5000 data points
- test power estimated using 1000 MC replications
- power compared with KS ($d \leq 3$)
- α on diagonal is expected
- TopoTests yielded higher power than KS in most of the cases

Simulation results (one-sample)



average power at $\alpha = 0.05$:

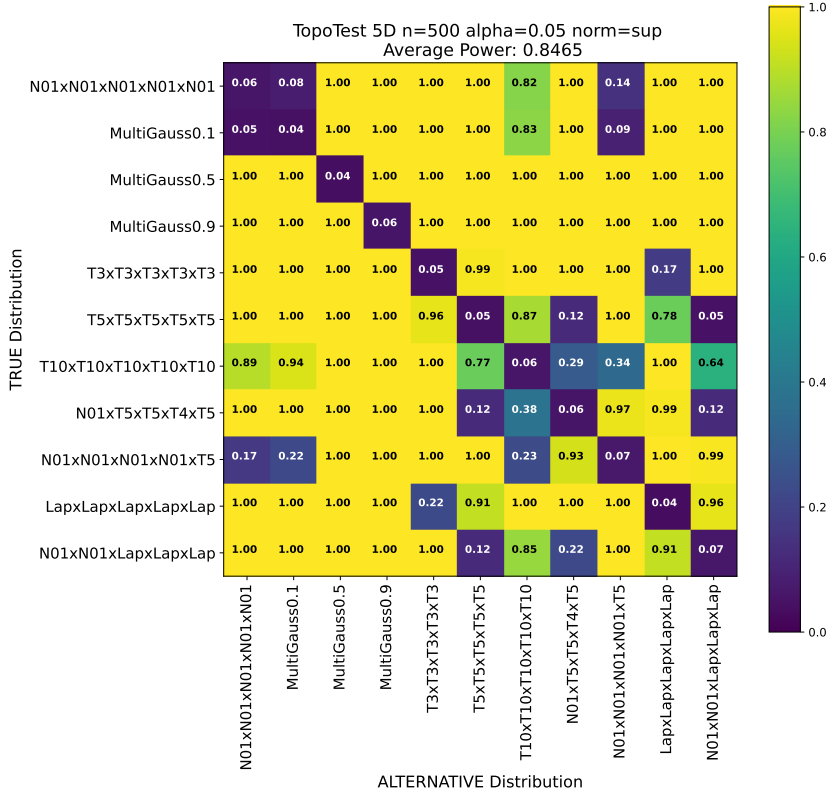
$d = 3, n = 250$ TT:0.9016, KS : 0.8087

$d = 5, n = 500$ TT:0.8465, KS : $- - -$

Test Power: probability that H_0 is correctly rejected when H_1 is true

- Samples sizes 100–5000 data points
- test power estimated using 1000 MC replications
- power compared with KS ($d \leq 3$)
- α on diagonal is expected
- TopoTests yielded higher power than KS in most of the cases

Simulation results (one-sample)



average power at $\alpha = 0.05$:

$d = 3, n = 250$ TT:0.9016, KS : 0.8087

$d = 5, n = 500$ TT:0.8465, KS : $- - -$

Test Power: probability that H_0 is correctly rejected when H_1 is true

- Samples sizes 100–5000 data points
- test power estimated using 1000 MC replications
- power compared with KS ($d \leq 3$)
- α on diagonal is expected
- TopoTests yielded higher power than KS in most of the cases

TopoTests: highlights

- Statistical tests based on topological descriptors of the data constructed
-  Code available on `github.com/dioscuri-tda/topotests`
- pip install topotests
- We have a theory that justifies the approach
- Performance of the method is higher, while computational effort is lower, than for Kolmogorov-Smirnov

Philosophical remark

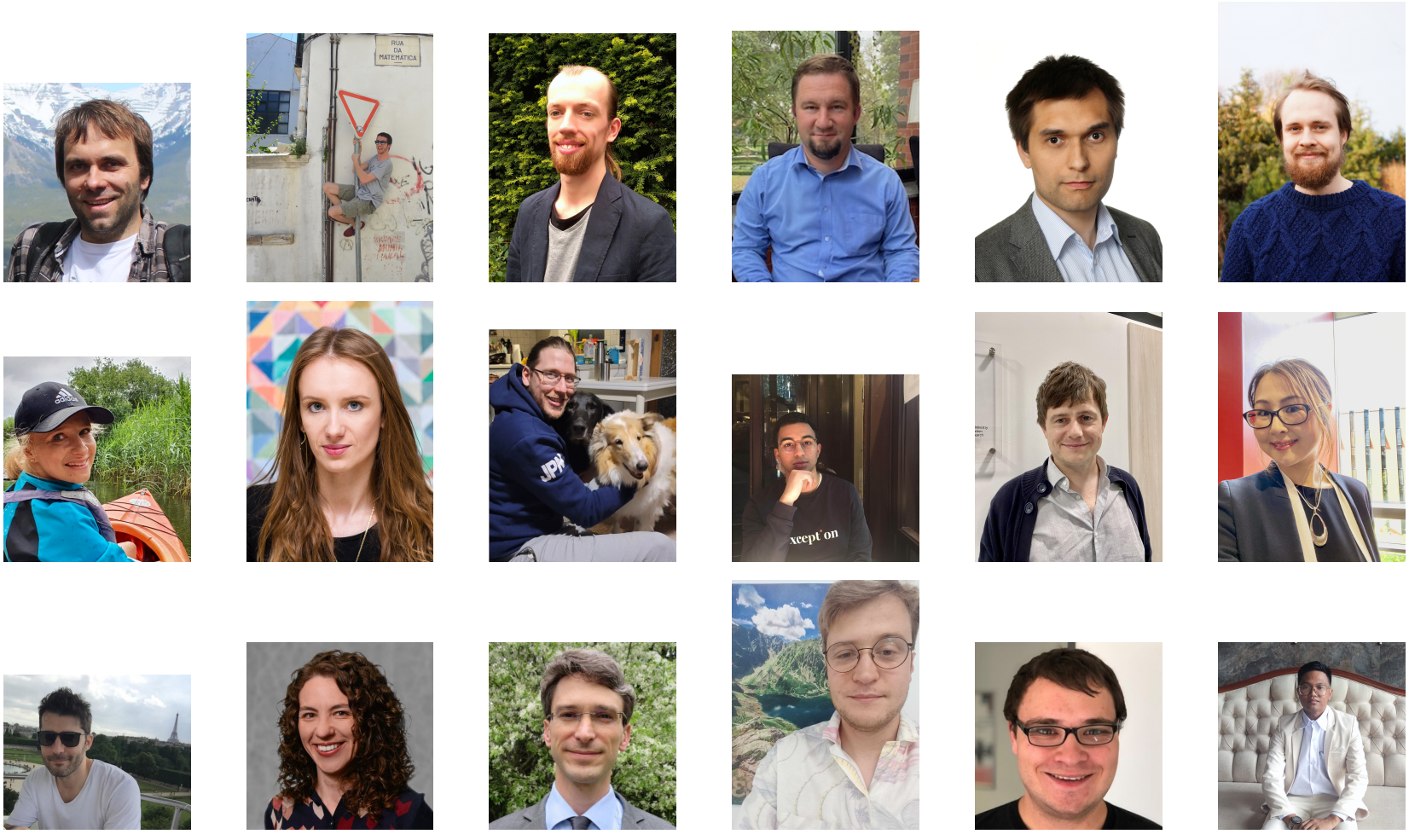
- TDA provide more powerful statistics,
- Some asymptotic properties of those statistics can be shown,
- But most time theoretical results are hard to get
- We need to be estimated using Monte Carlo simulations.

Outreach articles on the topic

(A) Topological features, EMS Magazine, https://euromathsoc.org/magazine/articles/190

(B) Topological visualization, journals.pan.pl/Content/125751/PDF/66-69_Dlotko_pol.pdf

The TDA-Team



We, the people of the Dioscuri Centre in Topological Data Analysis

Thank you!

(A) Topological features

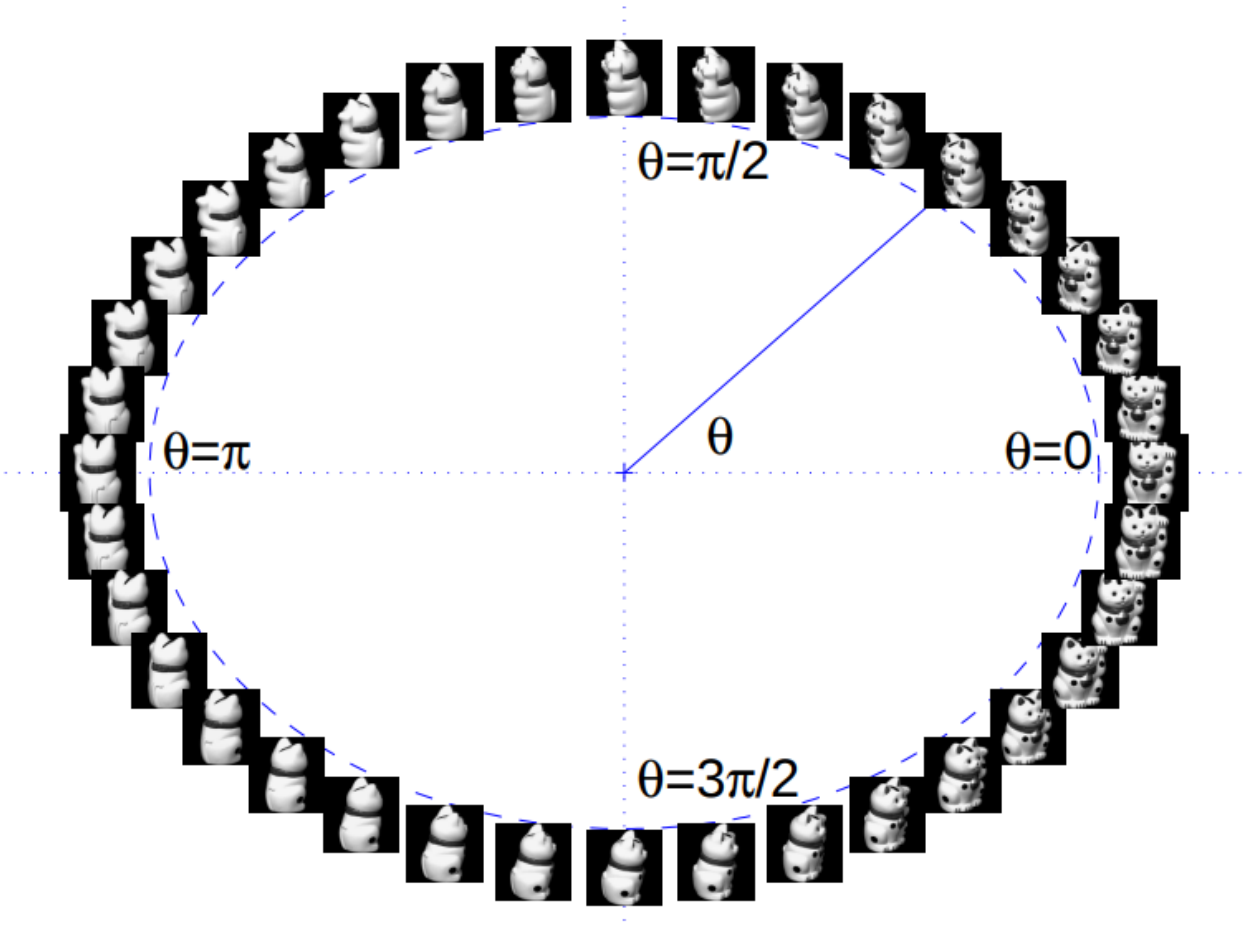
(B) Topological visualization

Paweł Dłotko, Dioscuri Centre in Topological Data Analysis
pdlotko @ gmail, http://dioscuri-tda.org/members/pawel.html

Bonus quiz

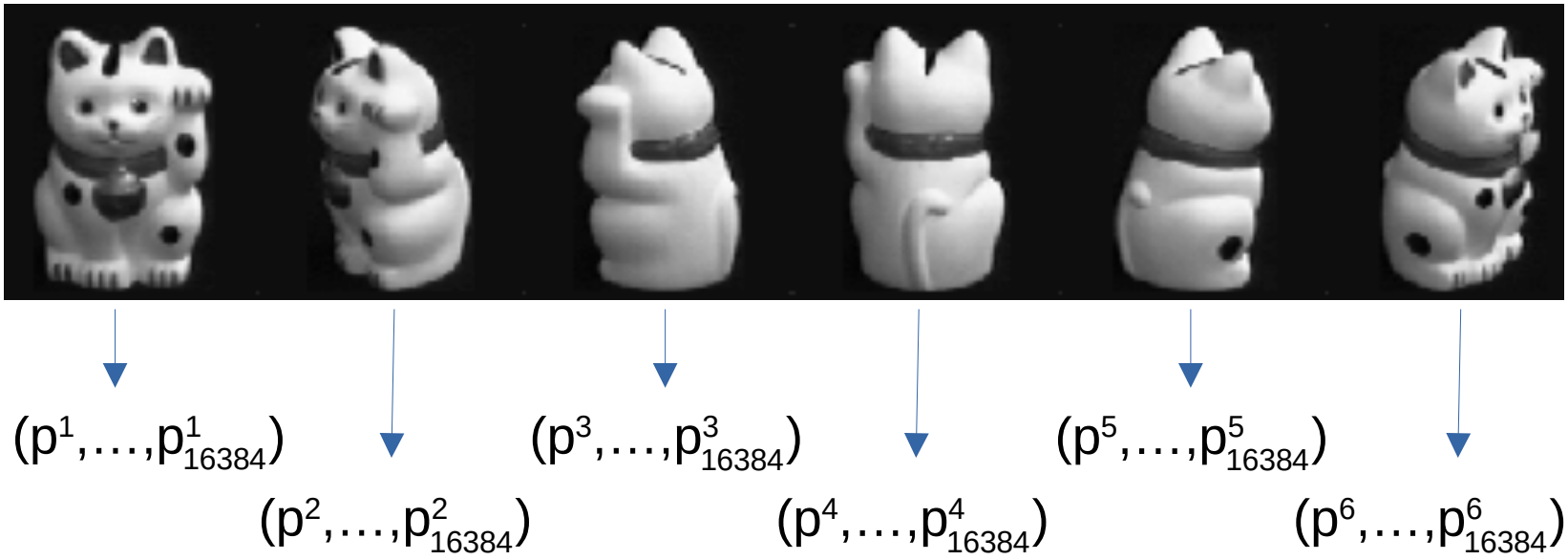
Exercise: Can you see in high dimensions?

Meet the Lucky Cat. Brings luck to everyone who solve this puzzle.



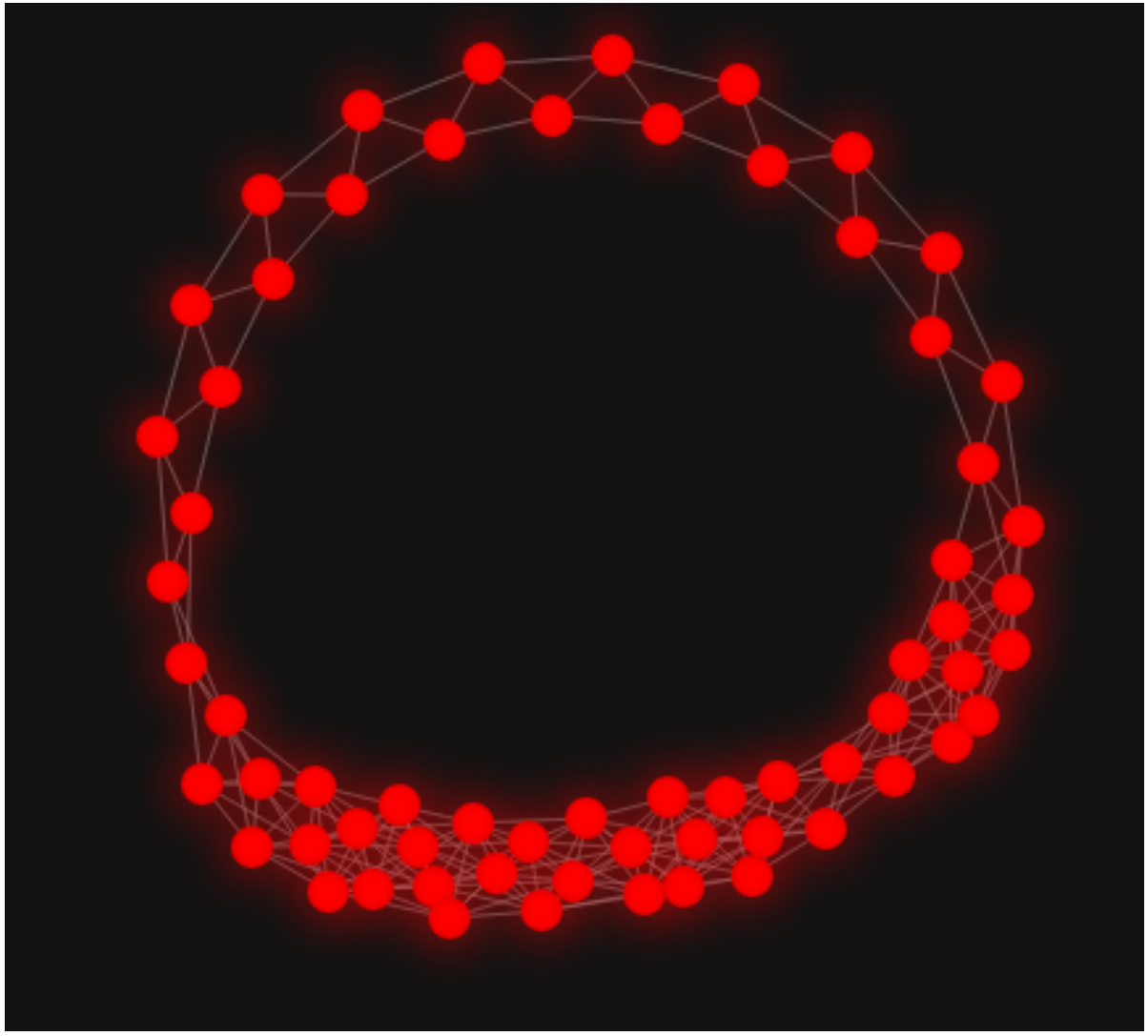
$128 \times 128 = 16384$ dimensional space

From a gray scale image to a point



Gray scale images converted to vectors in high dimensional space

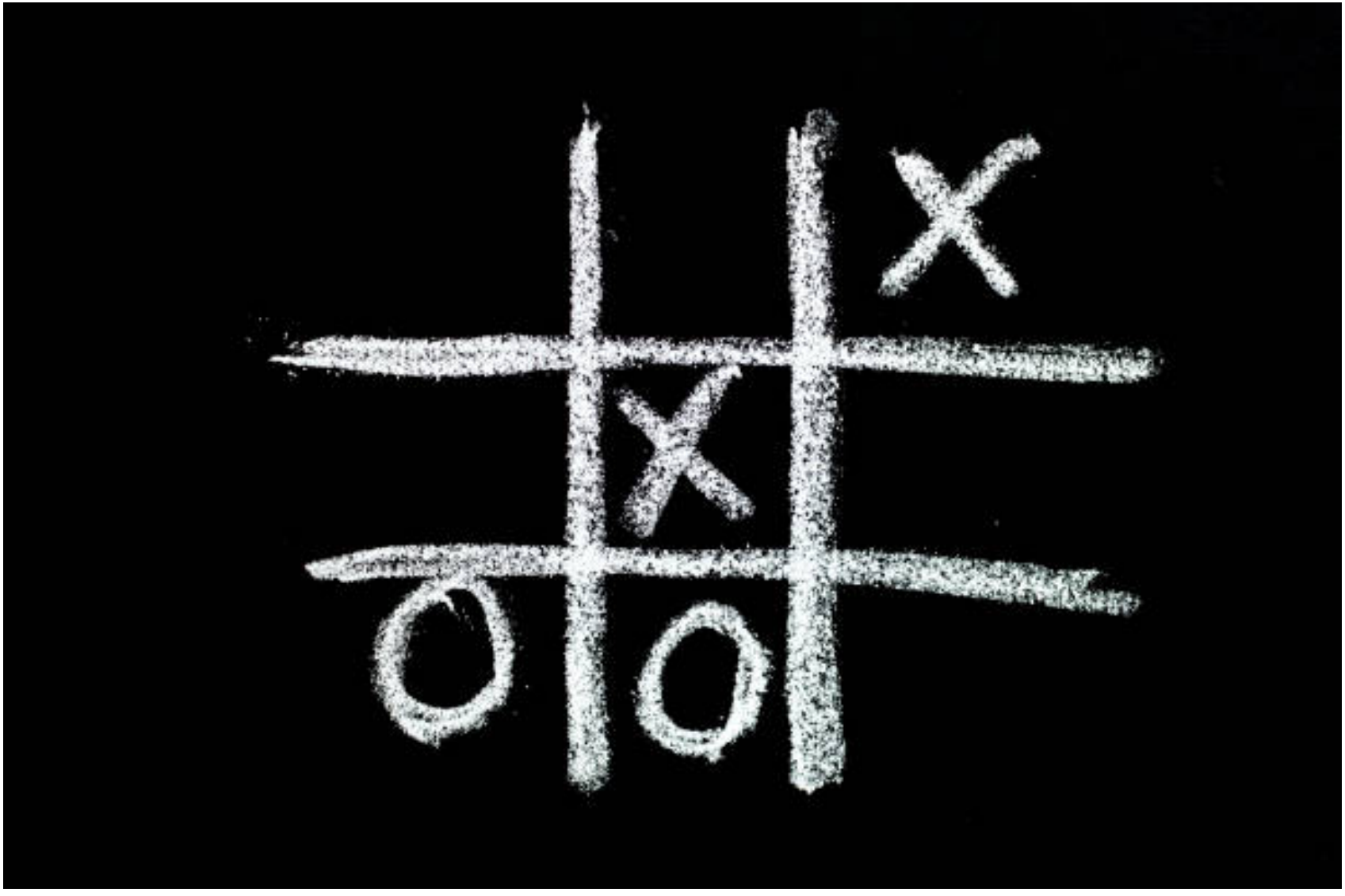
Network based landscapes of data



$128 \times 128 = 16384$ dimensional space

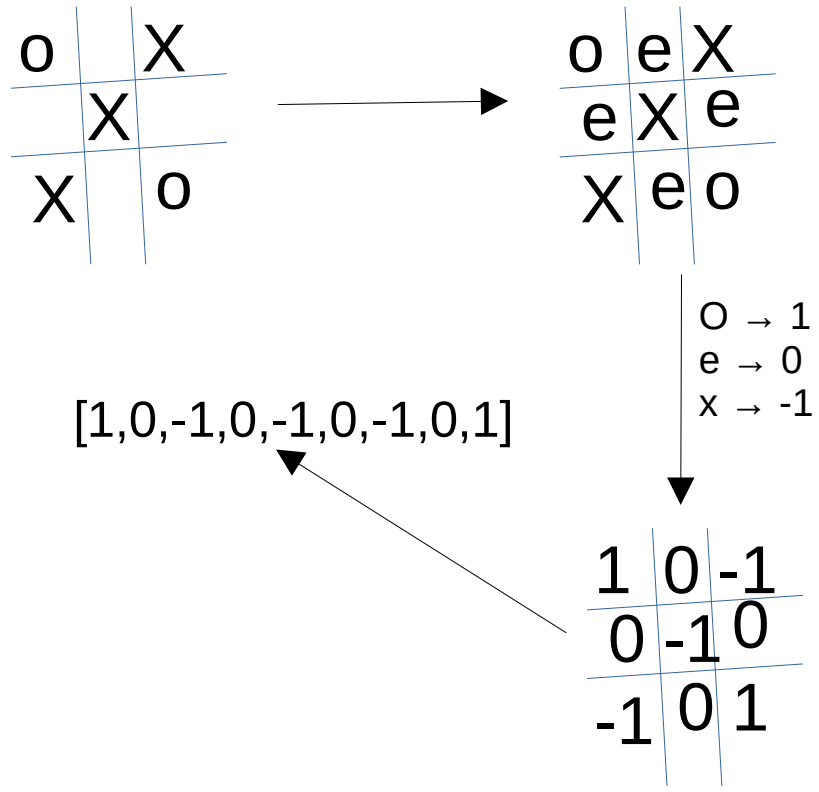
Bonus, tic-tac-toe

Let us get a bit less serious...



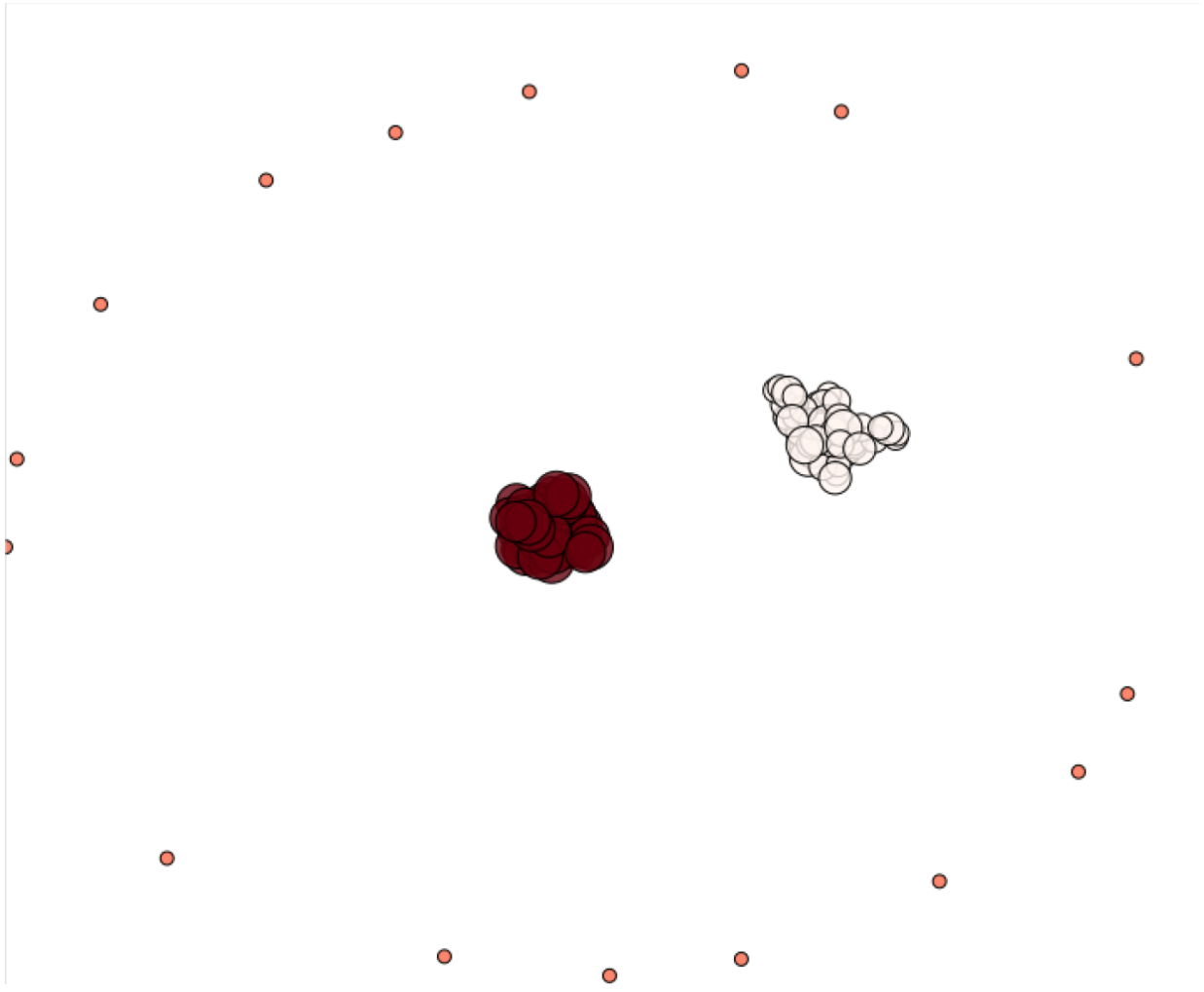
and a bit more combinatorial

Representation of a final configuration

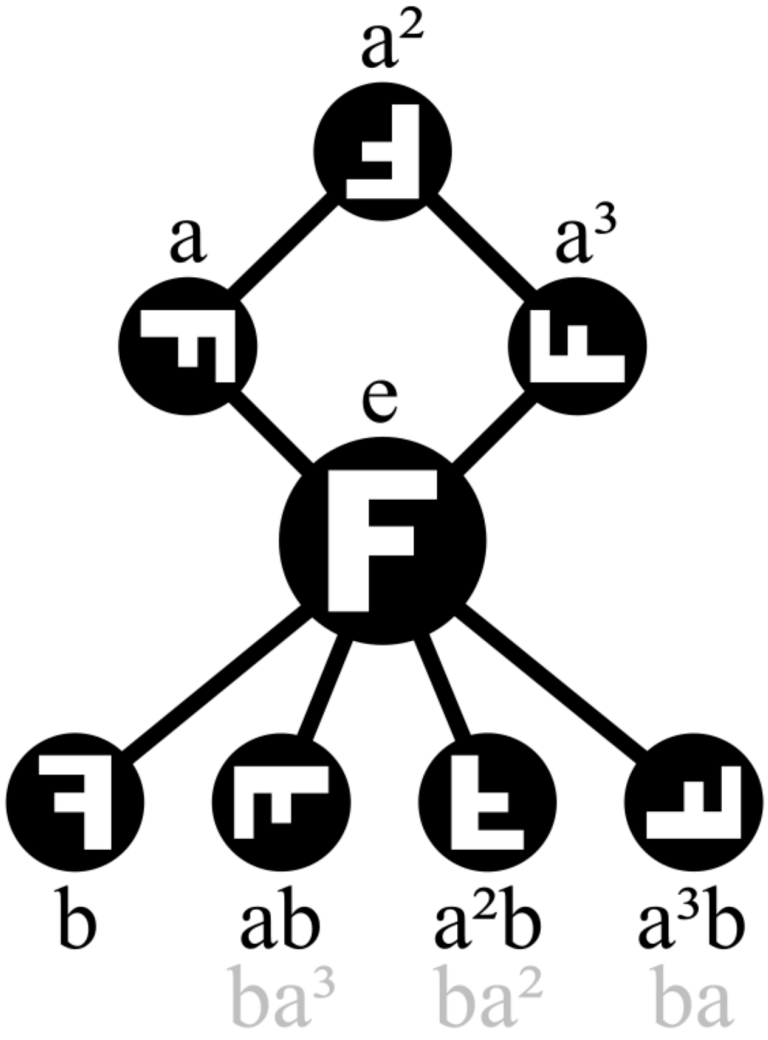


958 configurations labeled as 'first player win', 'first player lose', 'tie'.

Ball Mapper plot of the dataset

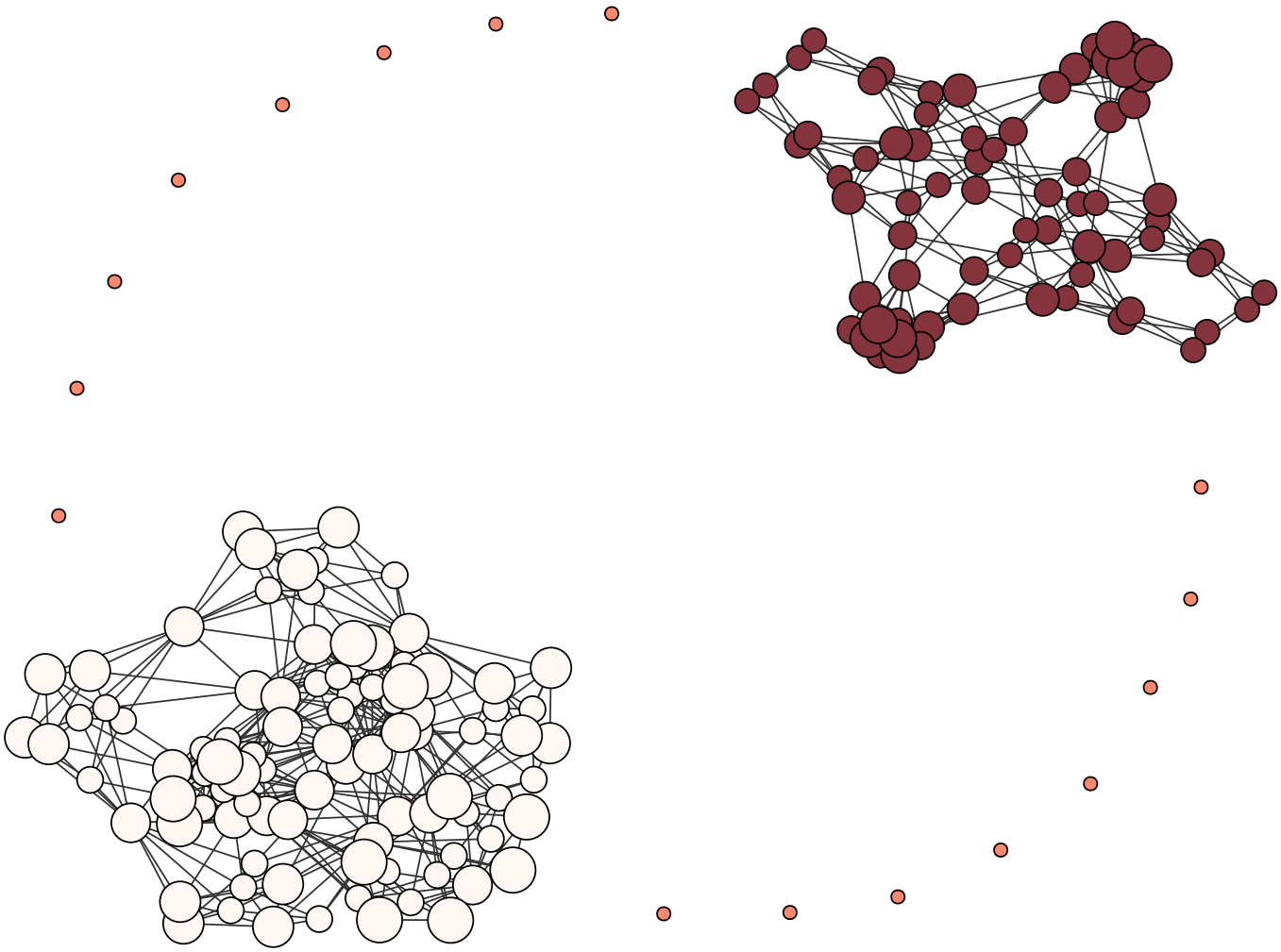


Ball mapper for $\epsilon = 2.5$ colored by the wins of the first player (red), loses (white), disjoint clusters (ties).



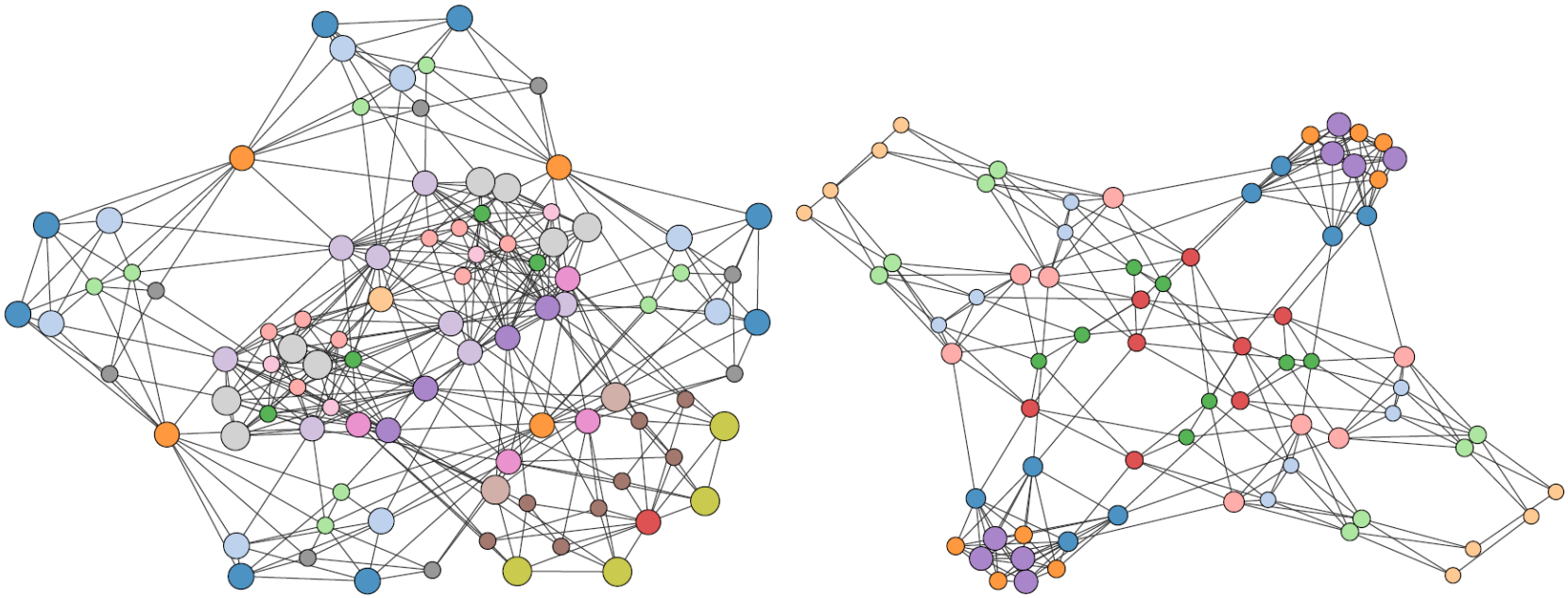
Dihedral group actions

Tik-Tak-Toe, all



P. Dlotko, D. Gurnari, R. Sazdanovic, Mapper-type algorithms for complex data and relations, Journal of Computational and Graphical Statistics, 1-18

Zoom in



- The wins cluster (left) and loses cluster (right) with color denoting the orbits.
- Different orbits might have different lengths. Asymmetric configurations have length 8 orbits.
- The maximally symmetric configuration has an orbit of length 1 -the only red node (left).