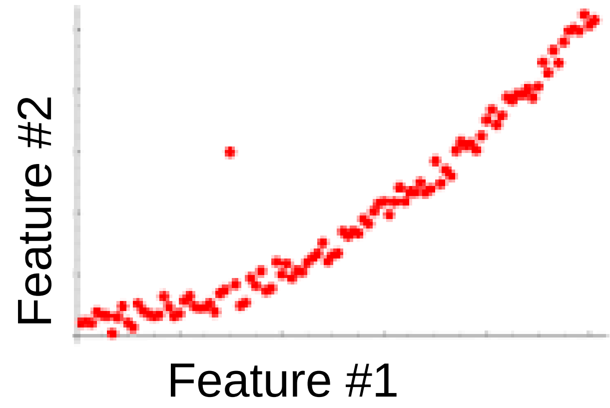


Outlier mining in high-dimensional datasets based on Jensen-Shannon distance and graph structure analysis

A. S. O. Toledo[1,2], R. Silini[1], L. C. Carpi[3], **Cristina Masoller[1]**

1. Departament de Fisica, Universitat Politecnica de Catalunya, Spain.
2. Centro Federal de Educacao Tecnologica de Minas Gerais, Brazil.
3. Universidade Federal de Minas Gerais, Brazil.

Complex Networks
Palermo, November 9, 2022

cristina.masoller@upc.edu @cristinamasoll1

An outlier or an anomaly?

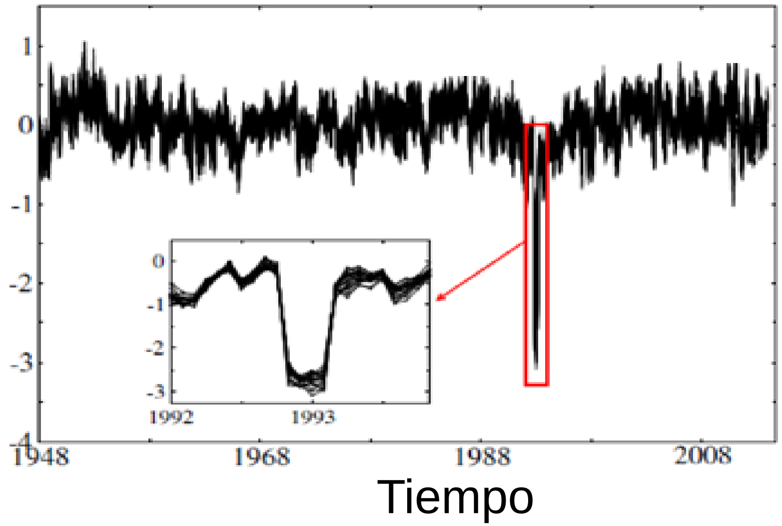
- *Outlier*: a legitimate data point that is far from the center of the distribution that characterizes the process
- *Anomaly*: a data point that cannot be explained given current knowledge of the process generating the data.
- Types:
  - *Point anomalies*: a data point that is anomalous with respect to the rest of the data.
  - *Contextual anomalies*: a data point that is anomalous in a specific context.
  - *Collective anomalies*: a set of data points that are not anomalies by themselves, but their collective occurrence is anomalous.

V. Chandola et al., ACM Comput. Surveys 41, 15 (2009)

cristina.masoller@upc.edu @cristinamasoll1

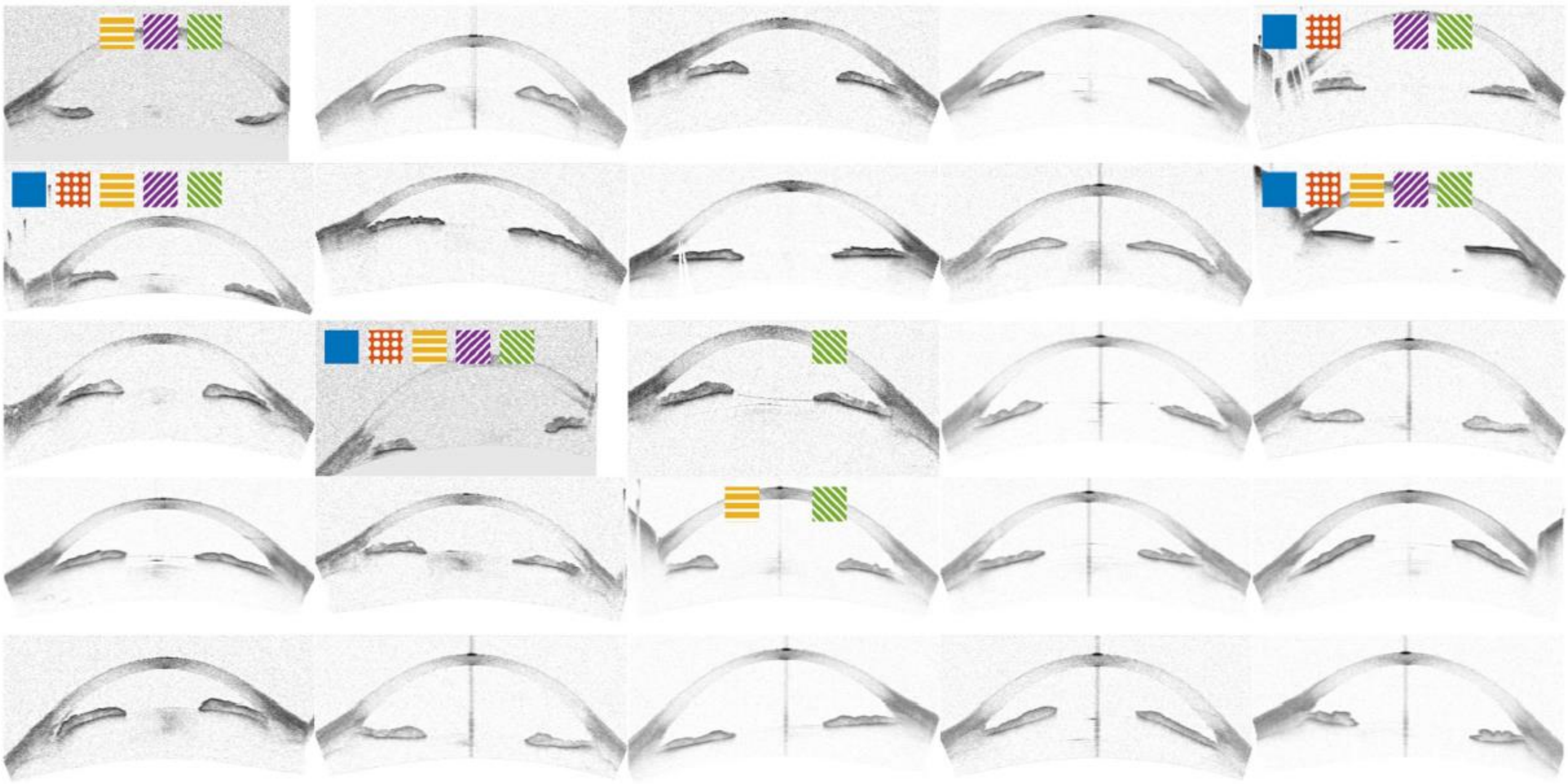
Outlier/anomaly detection algorithms: many applications

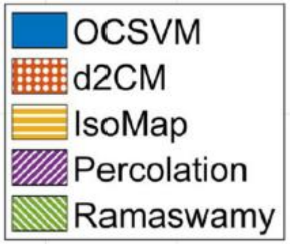
- Intrusion detection
- Failure detection
- Machine learning: filtering outliers from the training data improves algorithm performance.



cristina.masoller@upc.edu @cristinamasoll1

Our motivation: analysis of OCT anterior chamber images





P. Amil, N. Almeida and C. Masoller: Outlier Mining Methods Based on Graph Structure Analysis, Front. Phys. 7, 194 (2019).

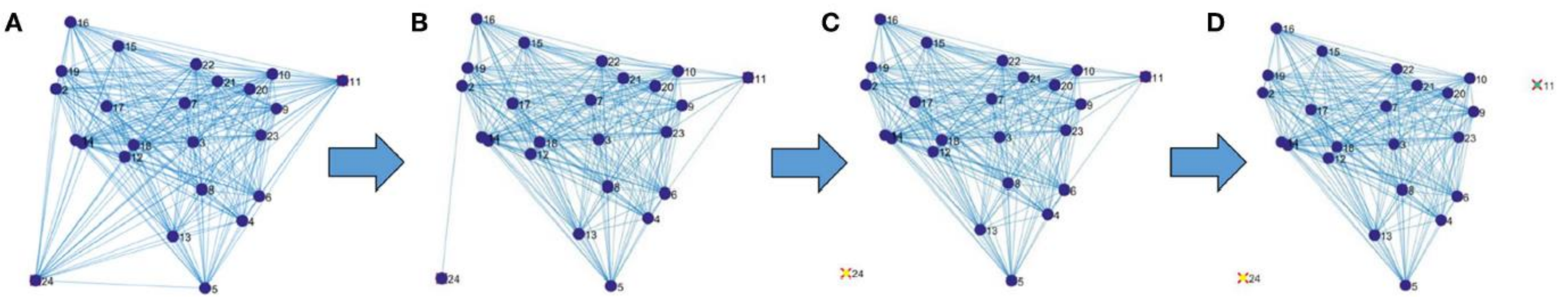
cristina.masoller@upc.edu @cristinamasoll1

We consider high-dimensional datasets where a distance can be defined between items of the dataset

Feature vectors of items *i* and *j*: $\{f_{i1} \dots f_{iM}\}$ $\{f_{j1} \dots f_{jM}\}$

Distance between them: $d_{ij} = \sqrt{\sum_{k=1}^{M} (f_{ik} - f_{jk})^2}$

First method: Outlier detection using percolation (OP)

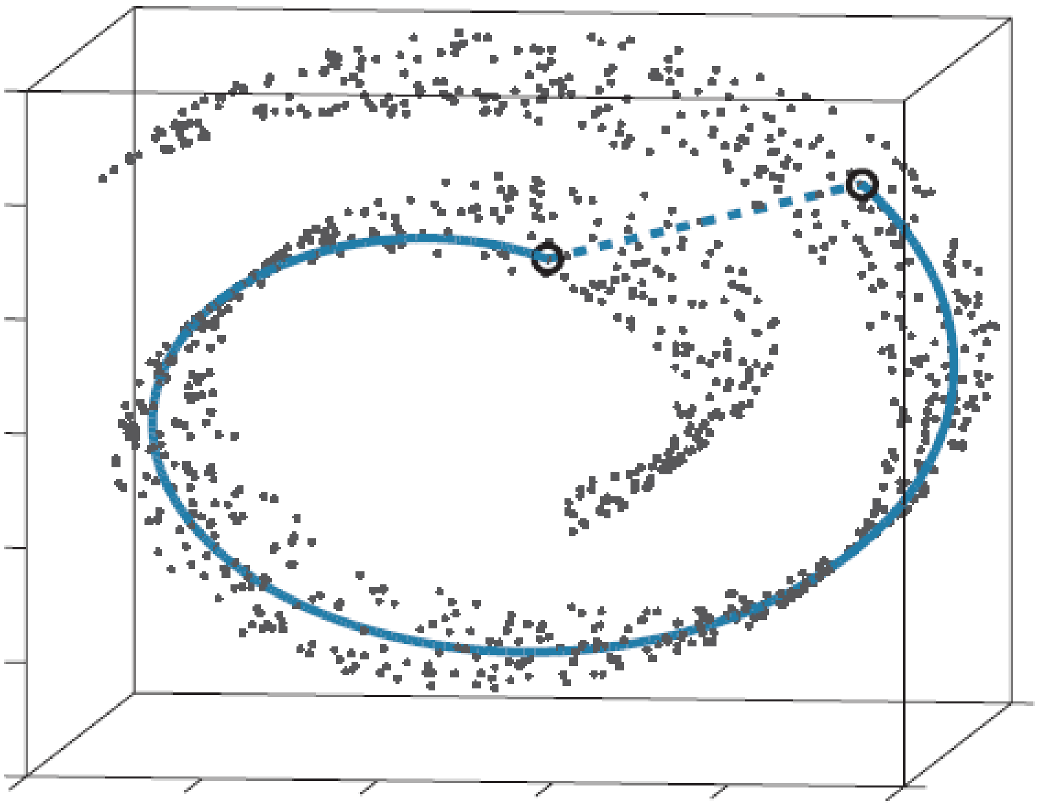


Outlier score OP = order in which elements disconnect from the giant component. **Parameter free.**

cristina.masoller@upc.edu @cristinamasoll1

Second method: nonlinear dimensionality reduction

- *Main idea: how well or how poorly an element fits in the learned manifold*.



Distance in the high dimensional space (dash) and distance in the learned (lower dimensional) manifold (solid).

ISOMAP, Tenenbaum et al., Science 290, 2319 (2000).
P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Steps

- Apply *IsoMap* to the distance matrix $\mathbf{D_{ij}}$ to obtain
  - a new set of features
  - a new distance matrix in the geodesic space, $\mathbf{D^G_{ij}}$
- With the new features, recalculate the distance matrix $\mathbf{D'_{ij}}$
- For each element, calculate correlation between $\mathbf{D^G_{ij}}$ and $\mathbf{D'_{ij}}$
- $AL_i = 1 - {\rho_i}^2$
- Two parameters (integers):
  - Dimension of reduced space
  - # of geodesic neighbors

Note: we don't use the features returned by Isomap to assign outlier scores

cristina.masoller@upc.edu @cristinamasoll1

Comparison with other outlier detection methods

- Distance to center of mass (d2CM): an outlier score is assigned according to the distance of an element to the center of mass.

- Ramaswamy: an outlier score is assigned according to the distance of an element to its kth nearest neighbor.

- One Class Support Vector Machine (OCSVM): uses the scalar product to define a function that returns +1 in the region where normal elements are located and −1 elsewhere.

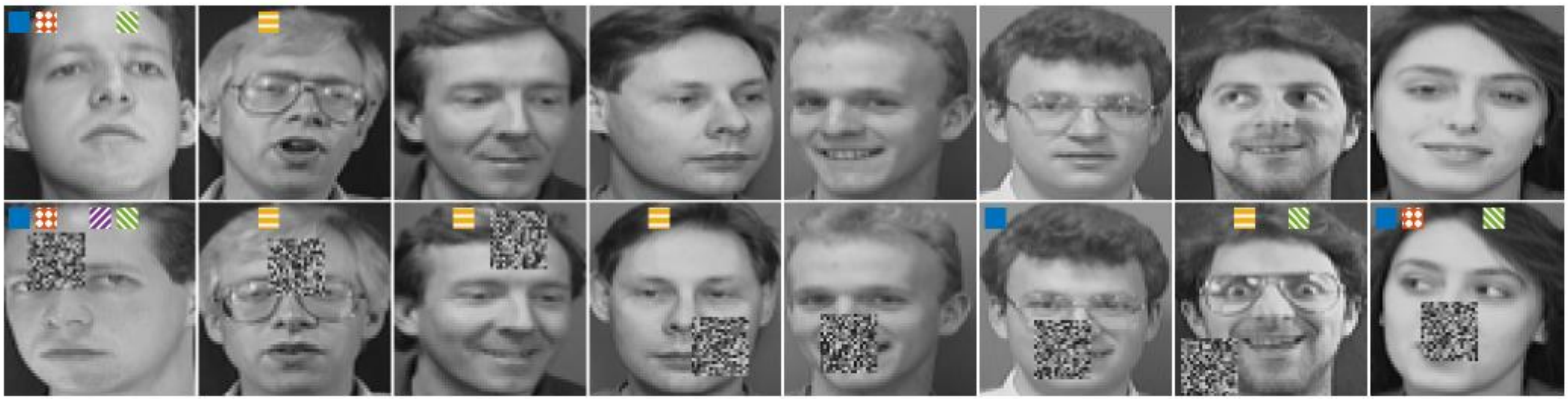
Ramaswamy et al., Efficient algorithms for mining outliers from large data sets. ACM Sigmod Record (Vol. 29, No. 2, pp. 427-438, 2000).
Schölkopf et al., Estimating the support of a high-dimensional distribution. Neural computation13, 1443-1471, 2001.

cristina.masoller@upc.edu **@cristinamasoll1**

Dataset and performance measures

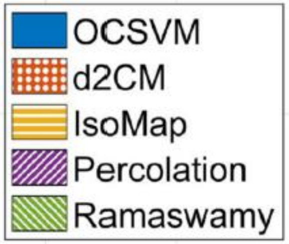
http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

We added to some random images a square with gray-scale pixels whose color distribution is the same as that of the image.



Performance quantification: average precision

(area under the Precision-Recall curve, TP/(TP+FP) vs TP)

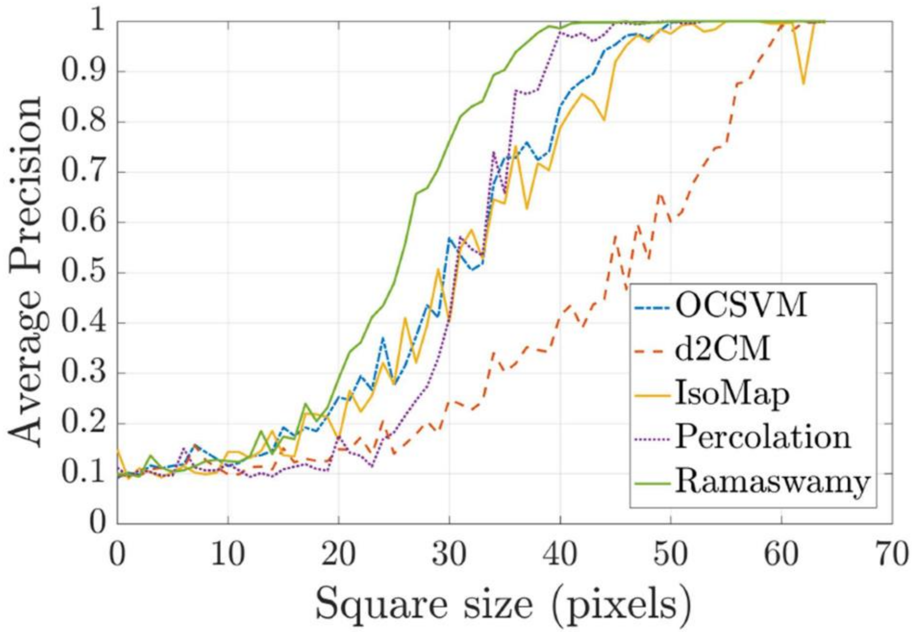


cristina.masoller@upc.edu @cristinamasoll1

Results

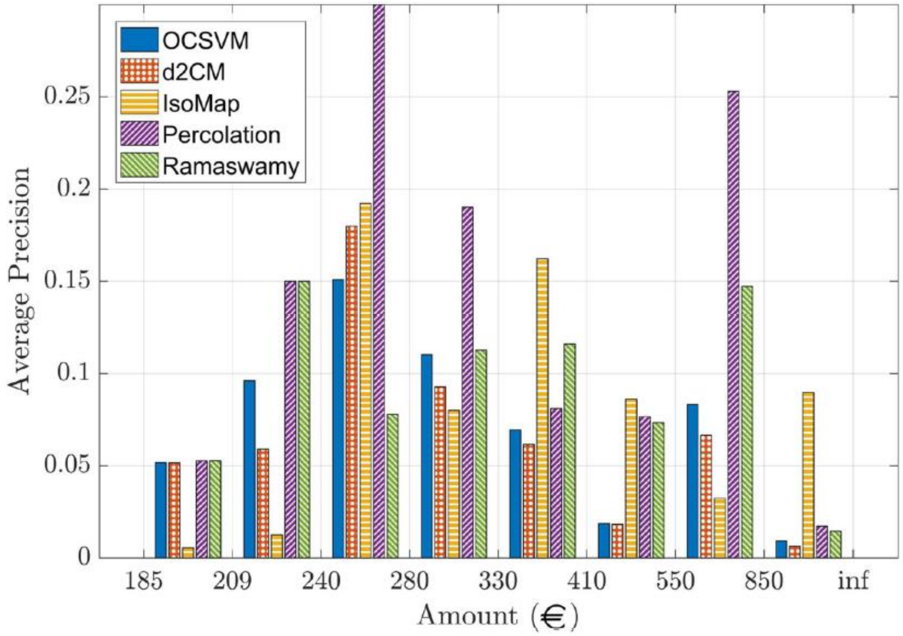
How about other types of elements?

Face database



Credit card transactions (some identified as frauds); each transaction has 28 features from PCA

https://www.kaggle.com/mlg-ulb/creditcardfraud



P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

The methods' performance depends on the data. How do they compare in terms of execution time?

For a database of 1,000 elements with 30 dimensions, run on *Matlab* on an Intel i7-7700HQ laptop:

Distance to center of mass	0.01 s
Ramaswamy	0.04 s
One Class Support Vector Machine	0.2 s
Percolation	6 s
IsoMap	18 s

Can we do better?

cristina.masoller@upc.edu @cristinamasoll1

Two new outlier mining methods

Feature vectors of items *i* and *j*

$$\{f_{i1} \dots f_{iM}\} \qquad \{f_{j1} \dots f_{jM}\}$$

Euclidean distance between them:

$$d_{ij} = \sqrt{\sum_{k=1}^{M} (f_{ik} - f_{jk})^2}$$

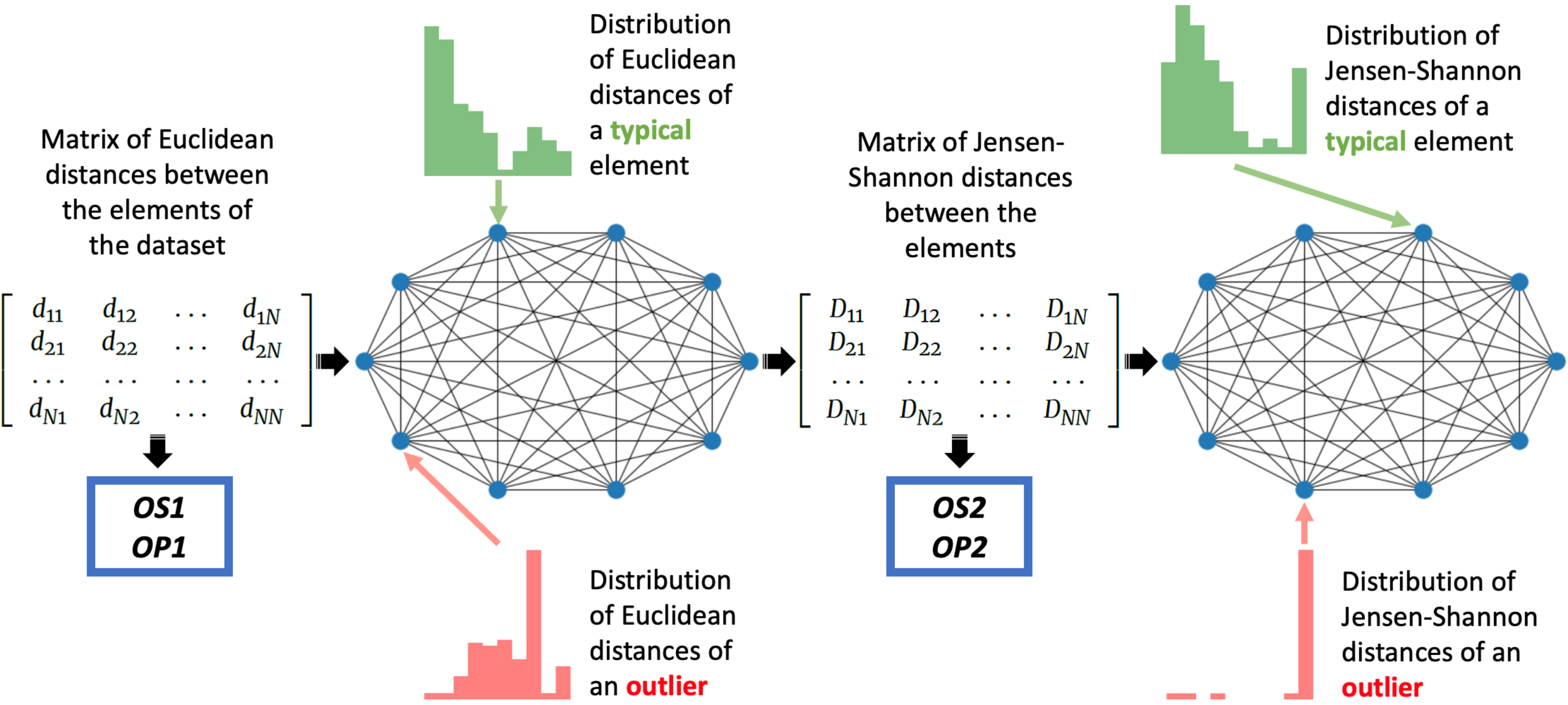
$$\boxed{OS1_i = \frac{1}{N} \sum_{l=1}^{N} d_{il}}$$

Instead of the Euclidean distance, we can use the distance between the distributions of distances $\{d_{il}\}$ and $\{d_{jl}\}$: the Jensen-Shannon distance

$$D_{ij} = JS[P_i, P_j] = H[(P_i + P_j)/2] - H[P_i]/2 - H[P_j]/2,$$

$$\boxed{OS2_i = \frac{1}{N} \sum_{l=1}^{N} D_{il}}$$

cristina.masoller@upc.edu @cristinamasoll1



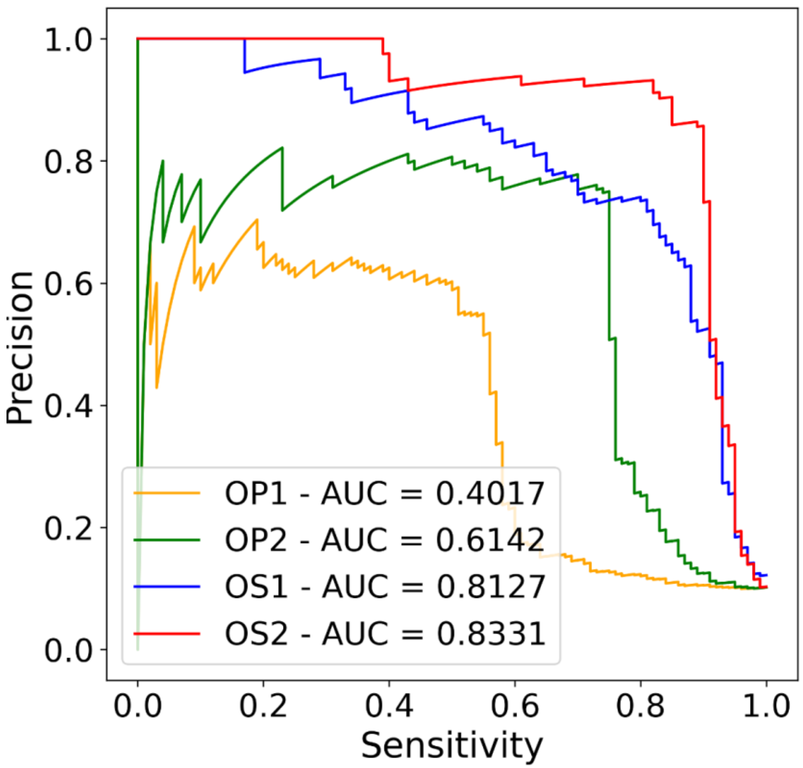
OS1, OS2: Outlier score defined by the sum of distances

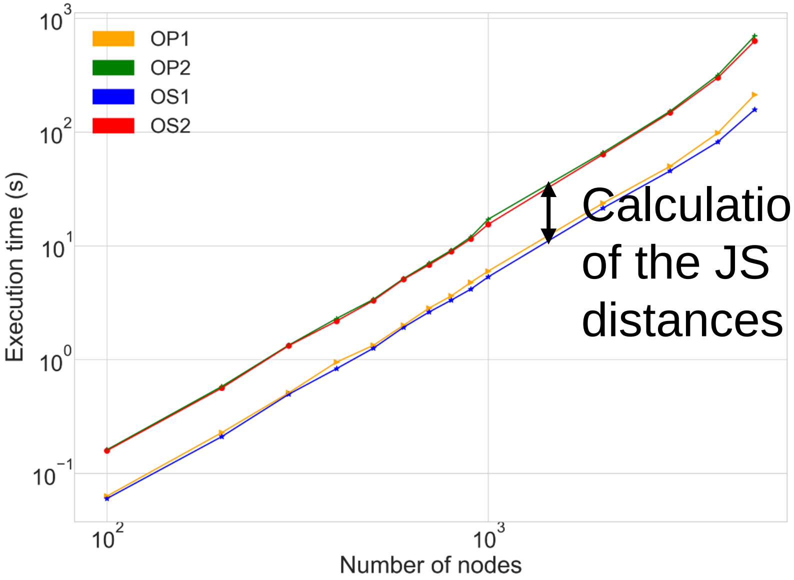
OP1, OP2: Outlier score defined by the percolation order

cristina.masoller@upc.edu @cristinamasoll1

Results for the credit card database, 1000 transactions

10% fraud
(similar with 5% fraud)





Calculation of the JS distances

Python, iMac core i7 with 32 GB RAM:

	t(s)
OP1	5.99
OP2	17.12
OS1	5.33
OS2	16.45

cristina.masoller@upc.edu @cristinamasoll1

Conclusion: The methods are suitable for high dimensional, not too large databases because the execution time grows with the number of features and at least as NxN with the size of the dataset.

P. Amil, N. Almeida and C. Masoller, *Outlier Mining Methods Based on Graph Structure Analysis,* Frontiers in Physics 7, 194 (2019).

A. S. O. Toledo et al., *Outlier mining in high-dimensional data using the Jensen-Shannon divergence and graph structure analysis,* under review (2022).

Thank you for your attention !

cristina.masoller@upc.edu @cristinamasoll1