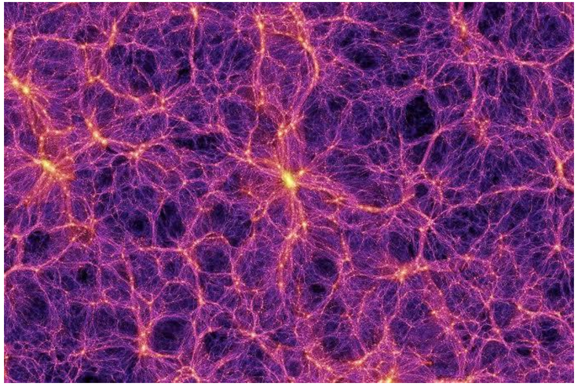
At the heart of density-based clustering: The percolation phenomenon.

Applications: Galaxy filament network extraction and Italian Olive-Oil

Louis Hauseux PhD student
(K. Avrachenkov, éq. NEO – J. Zerubia, éq. AYANA),
INRIA d'Université Côte d'Azur

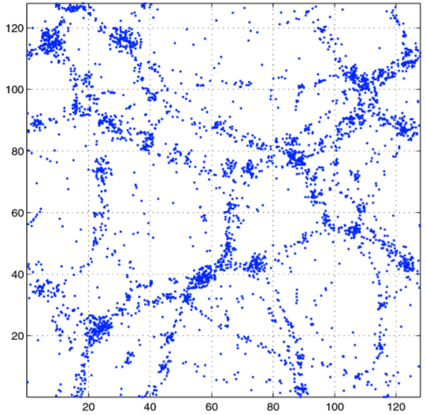
I – Introduction – Galaxy Filaments

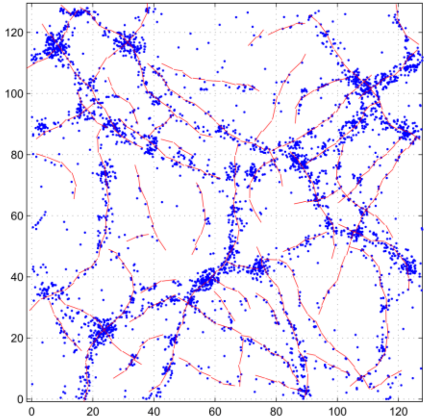
The problem: Galaxy Filament Extraction



Simulated 3D image of the CosmicWeb.

Galaxies are not distributed uniformly
- ➔ **'Large Scale Structures'**
- ➔ **Filamentary extraction**



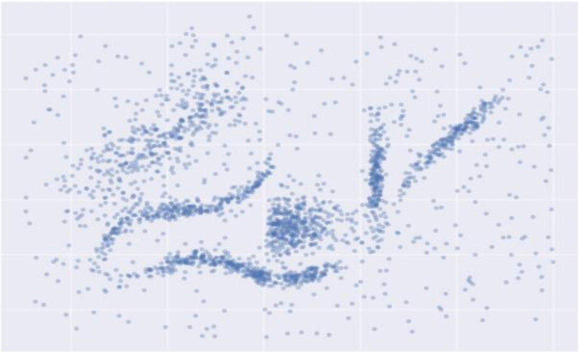


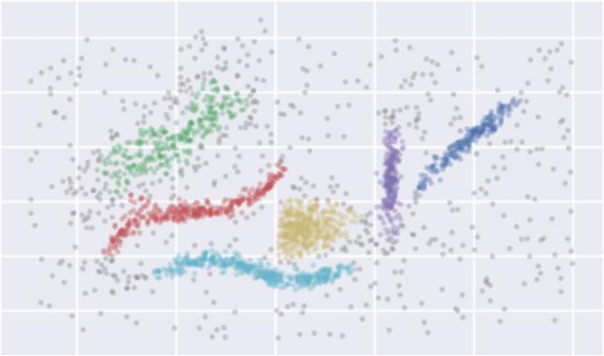
« Candy Model »
(Stochastic Geometry)*

* R. Stoica *et al. Detection of cosmic filaments using the Candy model* (2005)

I – Introduction

Our approach: Clustering view





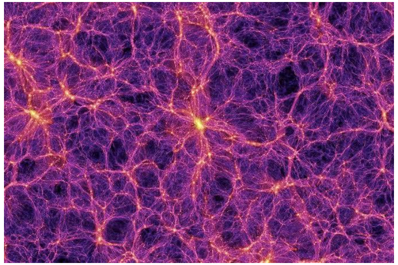
Hierarchical Density-Based SCAN Algorithm *

* McInnes and Healy “Accelerated hierarchical density based clustering” (2017)

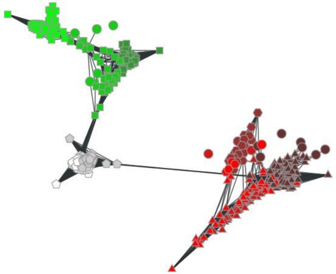
Overview

I – Introduction and illustration

- The illustrative Datasets: Galaxies and Italian Olive-Oil
- The mathematical model for density-based clustering:
  High-Density Clusters
- A classical algorithm: (Robust) Single-Linkage
  (and its mathematical limits)
- The benefits of hypergraphs. New model proposed.



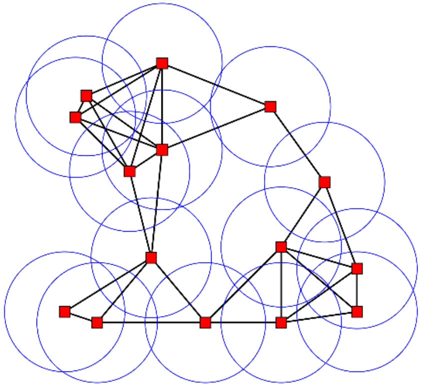
The Cosmic Web



Olive-Oil Dataset

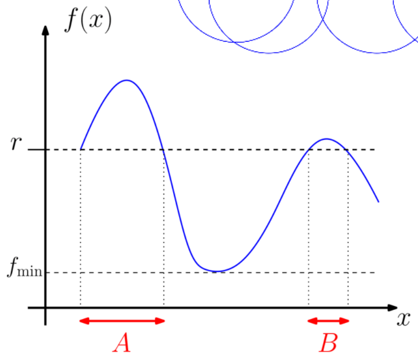
Overview

« Cluster » ou « Community » detection…
What is this ?



I – Introduction and illustration

- The illustrative Datasets: Galaxies and Olive-Oil
- **The mathematical model for density-based clustering: High-Density Clusters**
- A classical algorithm: (Robust) Single-Linkage (and its mathematical limits)
- The benefits of hypergraphs. New model proposed.

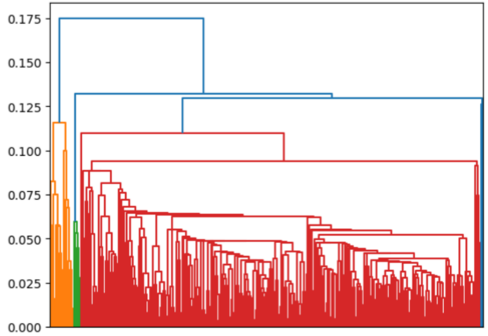


High-Density Clusters ***A*** and ***B***

Overview

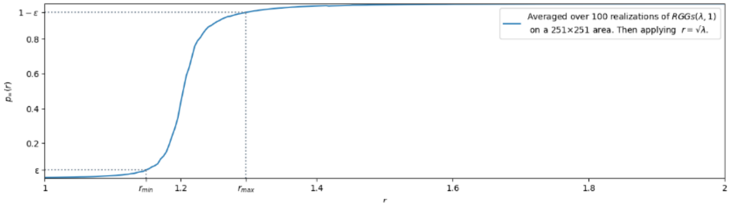
I – Introduction and illustration

- The illustrative Datasets: Galaxies and Olive-Oil
- The mathematical model for density-based clustering:
  High-Density Clusters
- A classical algorithm: (Robust) Single-Linkage
  (and its mathematical limits)
- The benefits of hypergraphs. New model proposed.



The ***Dendrogram*** produced by Single-Linkage

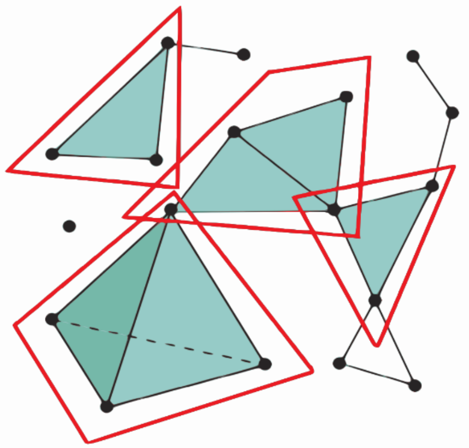
The ***rate of percolation***



Overview

I – Introduction and illustration

- The illustrative Datasets: Galaxies and Italian Olive-Oil Clustering
- The mathematical model for density-based clustering:
  High-Density Clusters
- A classical algorithm: (Robust) Single-Linkage
  (and its mathematical limits)
- The benefits of hypergraphs. New model proposed.

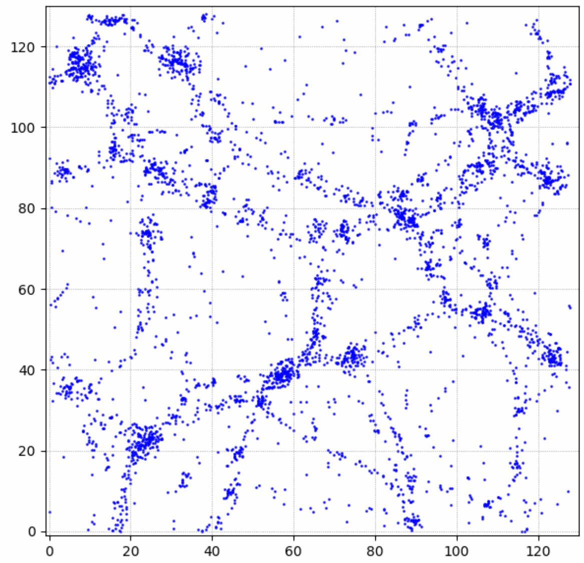


The ***Hypergraphs*** and their ***Components***

Overview

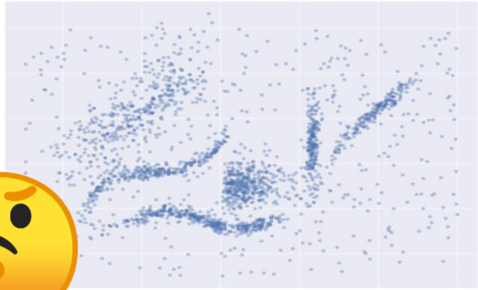
II – Results (illustrative & math)

- Asymptotic percolation rate on the discrete grid
- Correspondance between High-Density Clusters for K-NN and Čech complexes
- The persistant extraction of Galaxy Filaments
- The clustering of olive-oils matches with italian geography !



Variational extraction of Galaxy Filaments

I – The problem



Our approach: Clustering view

What is the the **underlying structure** (= the **clusters**) ?

➔ The **density** f of point-generation

Goal: identify the « **High-Density Clusters** »*

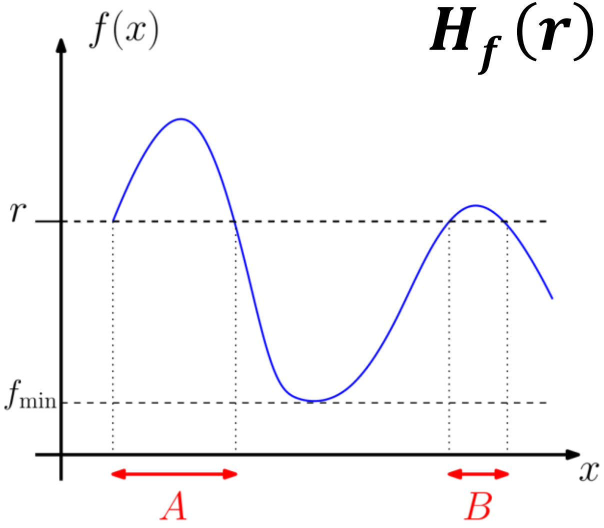
* J. Hartigan. *Clustering Algorithms* (1975)

I – High-Density Clusters

High-Density Clusters $\boldsymbol{H_f(r)}$ of level $\boldsymbol{r}$: the connected components of:

$$\boldsymbol{H_f(r)} = \{\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d}} \mid \boldsymbol{f(x)} \geq \boldsymbol{r}\} \subseteq \mathbb{R}^{\boldsymbol{d}}$$

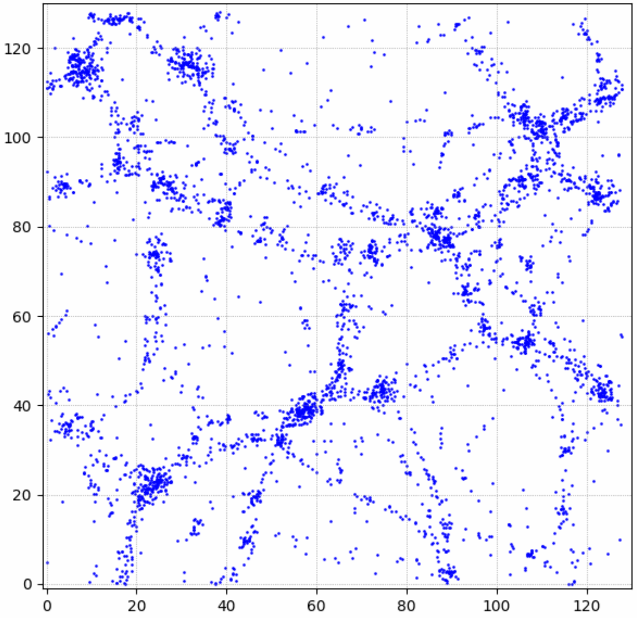
Important !

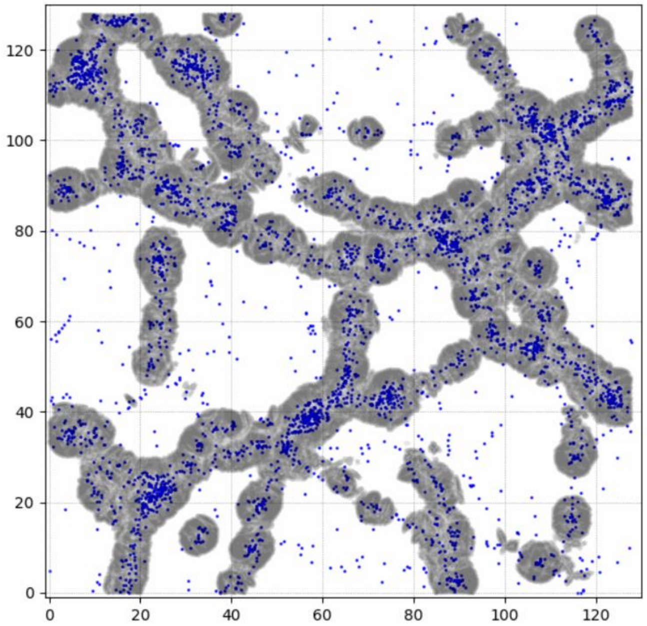


High-Density Clusters ***A*** and ***B***

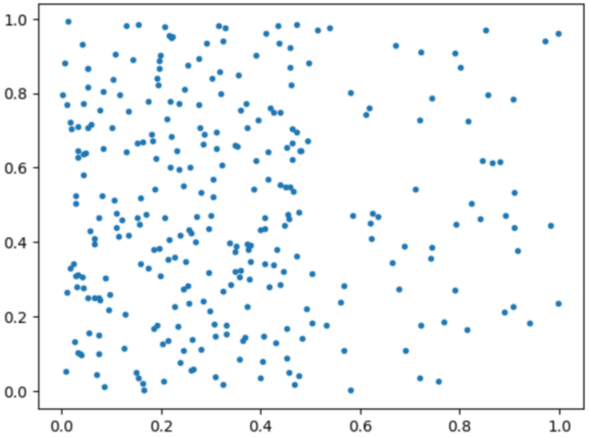
I – High-Density Clusters

High-Density Clusters $\boldsymbol{H_{\hat{f}}(r)}$ of level $\boldsymbol{r}$ for $\hat{f}$ = density estimator of ***10-Nearest Neighbors***





I – Single-Linkage



Practical situation : the unit square $[\mathbf{0}\,;\mathbf{1}]^2$ is split in 2:

- The left-rectangle with high density.
- The right-one with low density.

Single-Linkage algorithm: Start with the trivial classification: $\boldsymbol{n}$ points for $\boldsymbol{n}$ clusters.

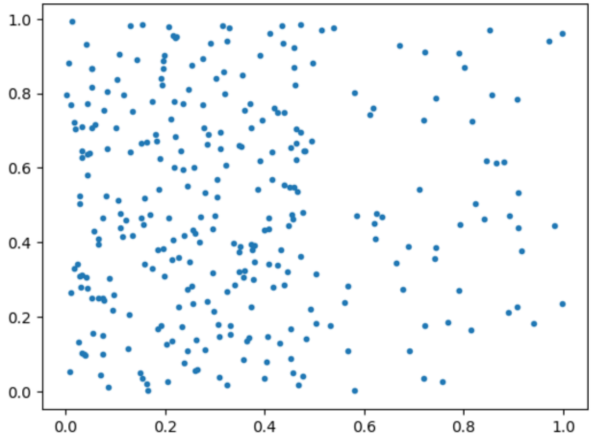
Partition on the points

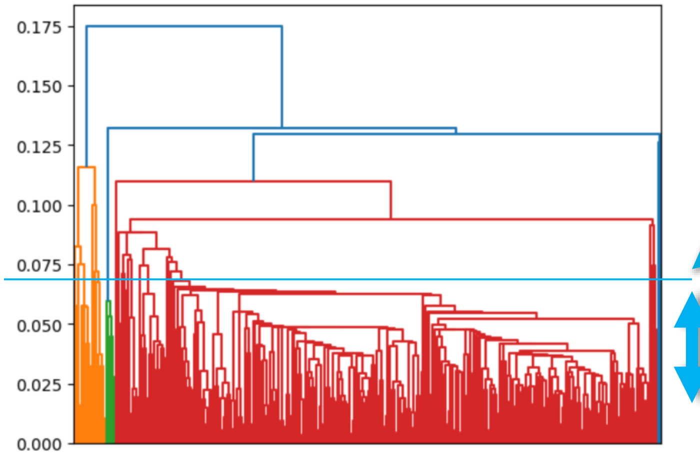
Initial **Clustering**:

$$\boldsymbol{C} = \{\boldsymbol{C}_1, \dots, \boldsymbol{C}_n\} \text{ with } \boldsymbol{C}_1 = \{\boldsymbol{x}_1\}, \dots, \boldsymbol{C}_n = \{\boldsymbol{x}_n\}$$

At each step, we merge <u>the two closest clusters</u>

I – Single-Linkage





$r \approx 0{,}7$

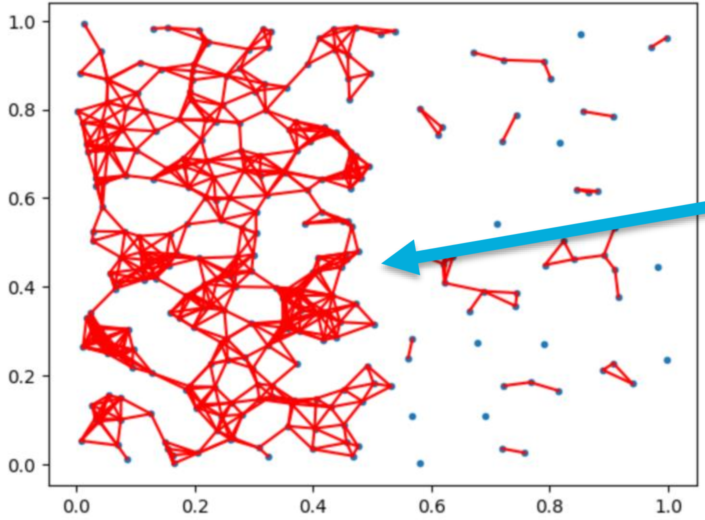
« Cluster merging phase » = Transition phase *

Hierarchichal Clustering produced by Single-Linkage

A tree ➔ the ***Dendrogram***

* G. Parisi. *Statistical Field Theory* (1988)

I – Single-Linkage ~Geometric Graphs

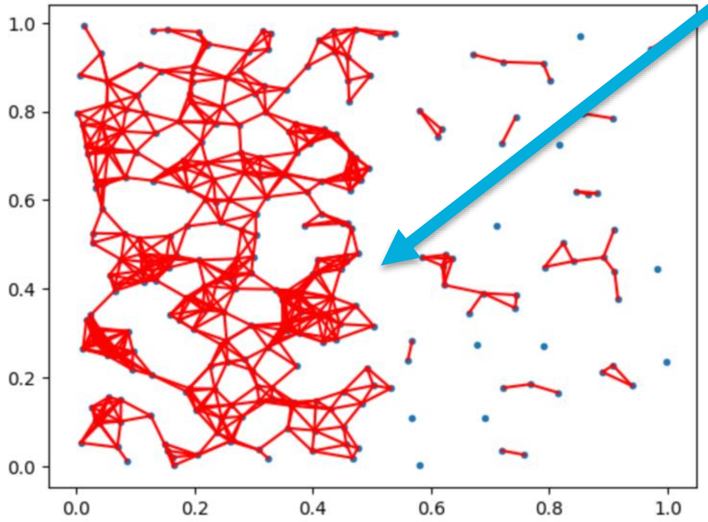


Geometric graph with same radius
$r \approx 0{,}7$

We almost recover the
High-Density Cluster

I – Single-Linkage ~Geometric Graphs

We almost recover the
High-Density Cluster… Why ? 🤔



Single-Linkage
~ Geometric Graphs
~ Empirical High-Density Clusters 1-Nearest Neighbor:

$$\hat{f}_{\text{1-NN}}(y) = \frac{1}{R_y^d} \qquad \text{where } R_y = \min_{x \in \mathcal{X}_n} ||x - y||$$

I – Geometric Graphs ~ High-Density Cluster

Geometric Graphs ~ Empirical H.-D. Clusters 1-NN:

$$\hat{f}_{\text{1-NN}}(y) = \frac{1}{R_y^d} \qquad \text{where } R_y = \min_{x \in \mathcal{X}_n} ||x - y||$$

Create a graph $\boldsymbol{G}$ on the cloud point χ such that:

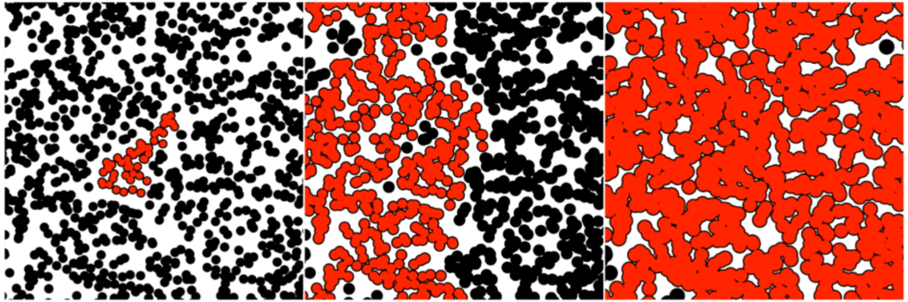
High-Density Clusters ⟷ Connected Components of $\boldsymbol{G}$

Subsets of $\mathbb{R}^{\boldsymbol{d}}$ .... Ok for $\boldsymbol{d} = \boldsymbol{2}$ or $\boldsymbol{3}$
... but computationally expensive for $\boldsymbol{d}$ large

I – Percolation *

Model : the $x_1, \dots, x_n$ IID plotted uniformally on $[0\,;1]^2$.

→ Density λ of **Poisson Point Process**



3 **Geometric Graphs** / greatest component

$\lambda < \lambda_c$ $\lambda = \lambda_c$ $\lambda > \lambda_c$

Percolation = giant component

→ **Very fast phenomenon**

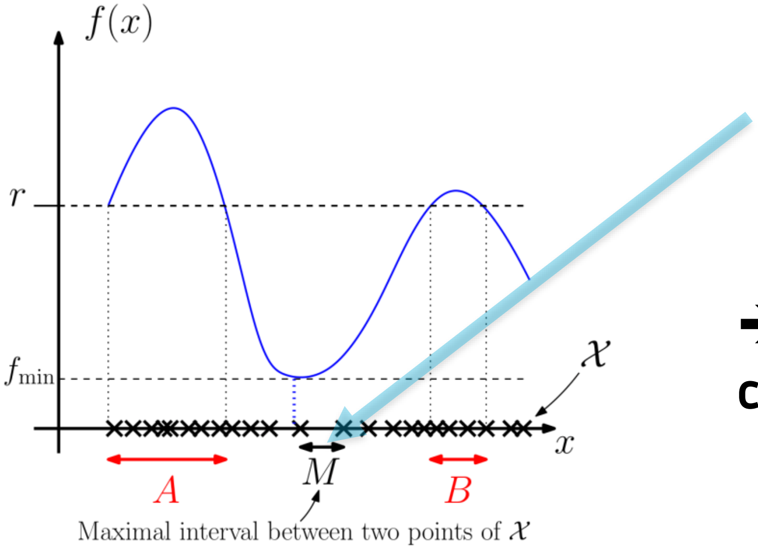
Two sub-cases : there exists a critical value $\boldsymbol{\lambda_c} \in (\mathbf{0}\,; \infty)$:

- If $\lambda < \lambda_c$, no great component (greatest of size $\Theta(\log(n))$)
- If $\lambda > \lambda_c$, **one giant component** (greatest of size $\Theta(n)$)

* M. Penrose. *Random Geometric Graphs* (2003)

I – Consistency and Single-Linkage

➔ Single-Linkage is **consistent** in $\mathbb{R}$!!!



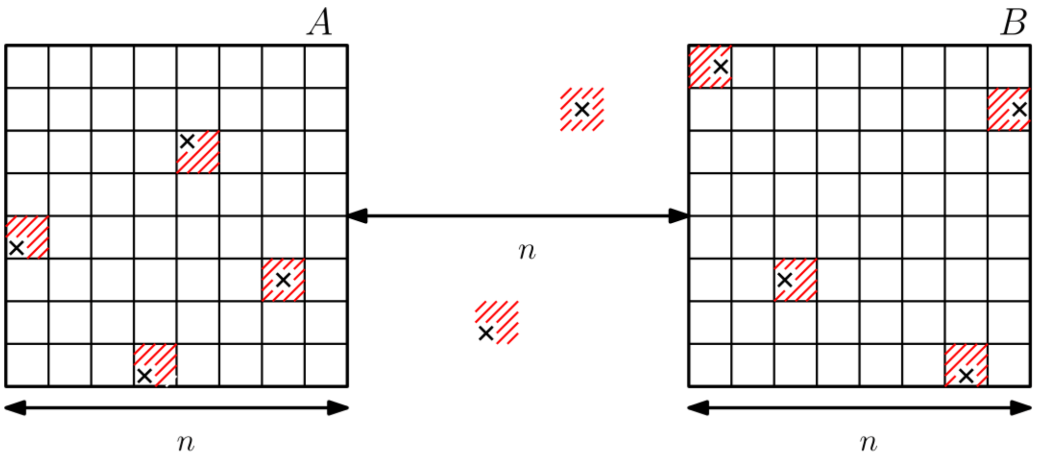
The « cut » between the two clusters, containing ***A*** or ***B***

➔ But Single-Linkage is <u>not</u> **consistent** in $\mathbb{R}^d$ for $d \geq 2$ …

I – Consistency and Single-Linkage

➔ Single-Linkage is <u>not</u> **consistent** in $\mathbb{R}^d$ for $d \geq 2$ … !!!

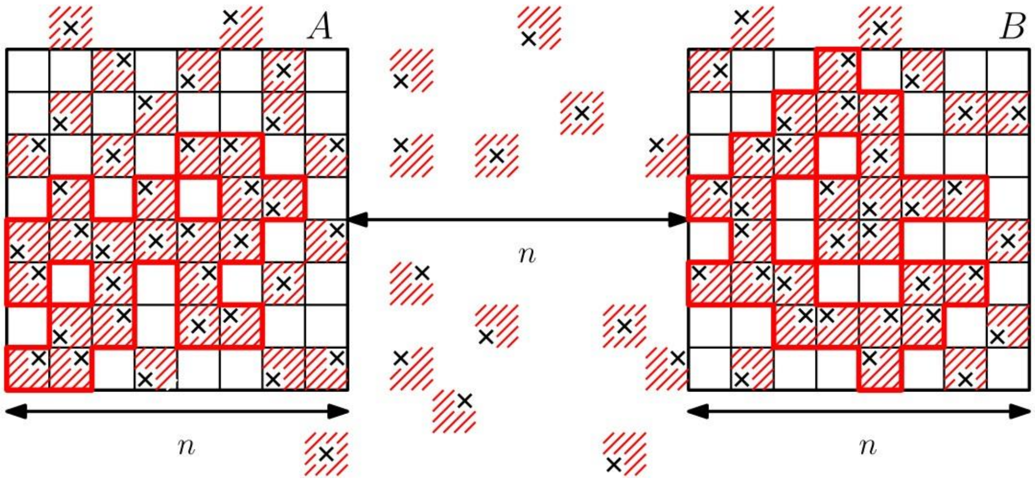
Discrete **Site-Percolation** on $\mathbb{Z}^2$



I – Consistency and Single-Linkage

➔ Single-Linkage is <u>not</u> **consistent** in $\mathbb{R}^d$ for $d \geq 2$ … !!!

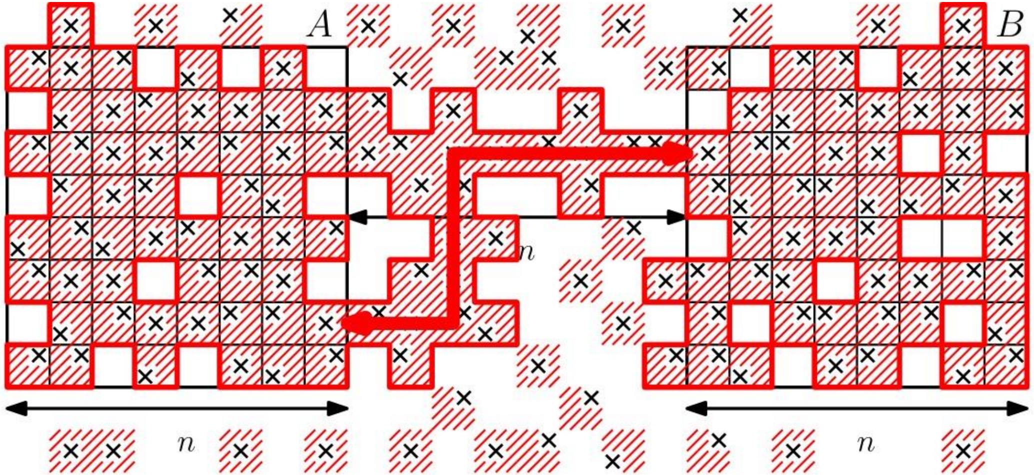
Discrete **Site-Percolation** on $\mathbb{Z}^2$



I – Consistency and Single-Linkage

➔ Single-Linkage is <u>not</u> **consistent** in $\mathbb{R}^d$ for $d \geq 2$ … !!!

Discrete **Site-Percolation** on $\mathbb{Z}^2$



The two clusters merge before having recovered ***A*** and ***B*** …

I – Consistency and Single-Linkage

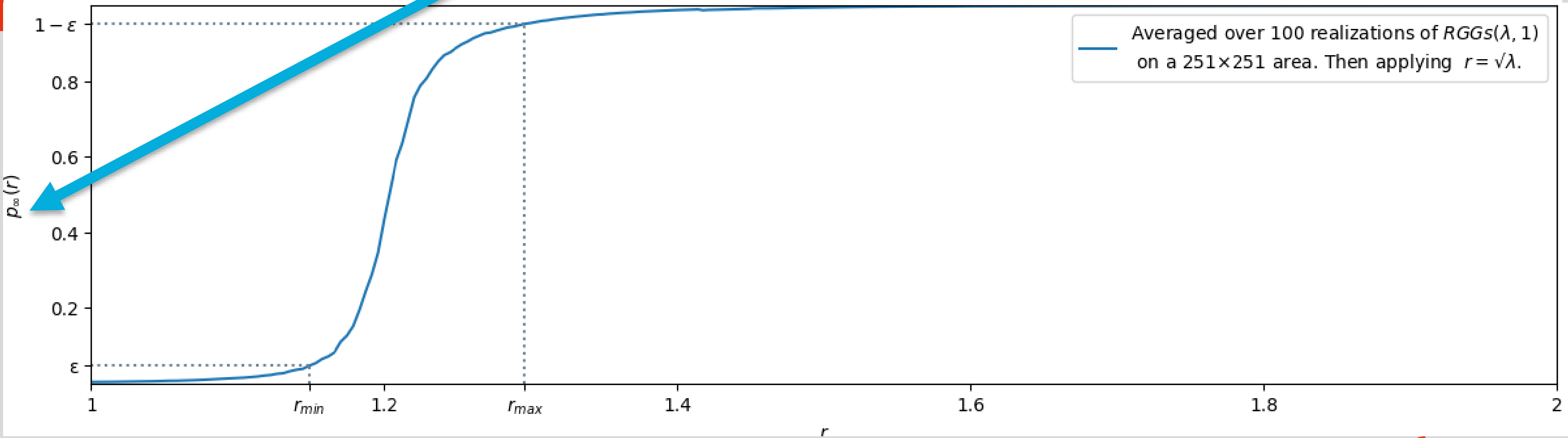
➔ Single-Linkage is <u>not</u> **consistent** in $\mathbb{R}^d$ for $d \geq 2$ ... !!!

➔ Only « fractional consistency » is possible

➔ Goal: obtain the largest possible « fraction »

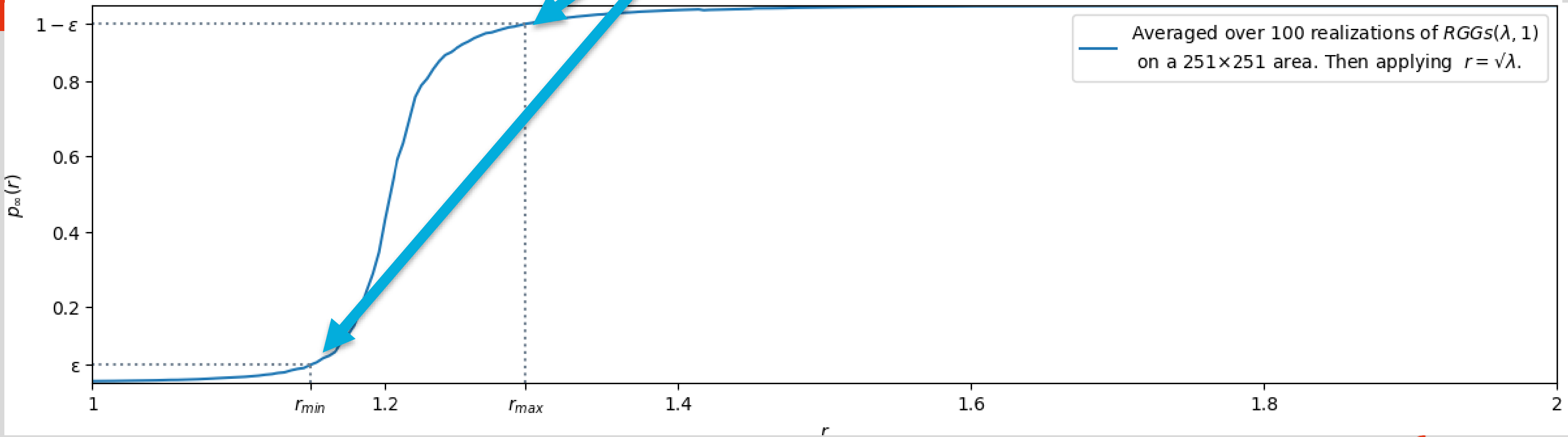
I – Rate of Percolation

Fraction of points in the **giant component**



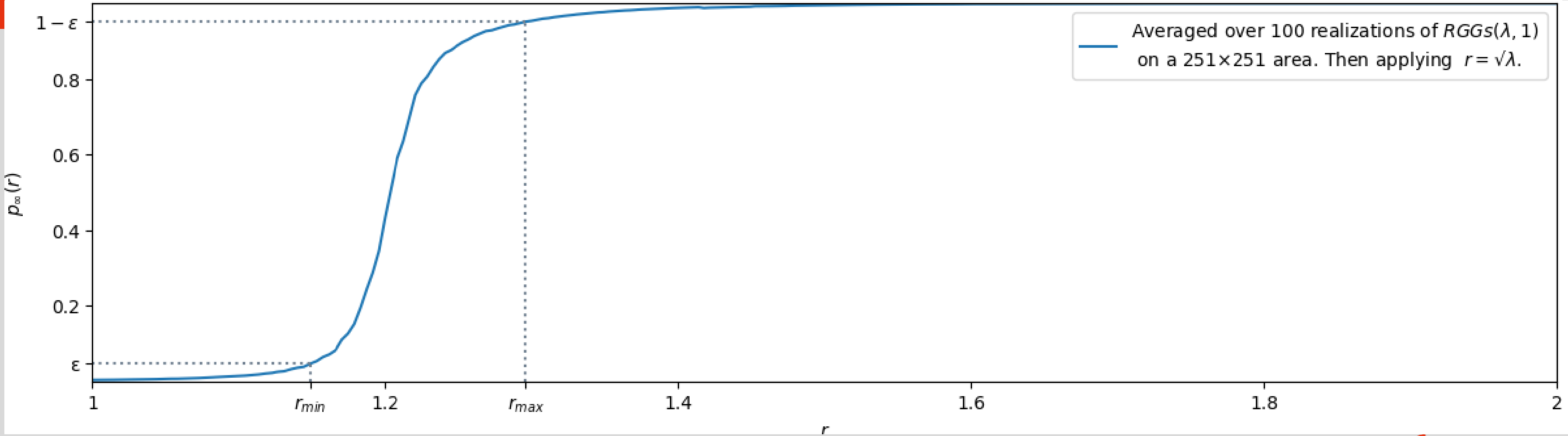
I – Rate of Percolation

The thresholds of « **detection** » and « **recovering** »



I – Percolation rate

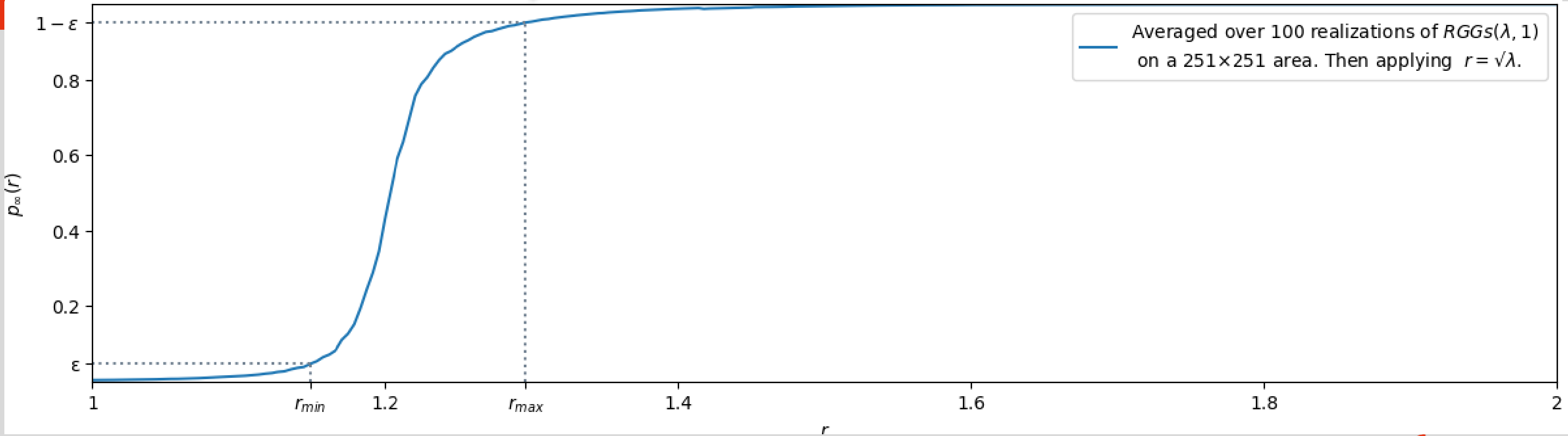
Percolation rate* $= \dfrac{r_{min}}{r_{max}}$



* L. Hauseux *et al.*: "Graph Based Approach for Galaxy Filament Extraction " (2023)

I – Percolation rate

This function depends on the objects (Graphs, **Hypergraphs**, …) and on the **kind of connectivity**

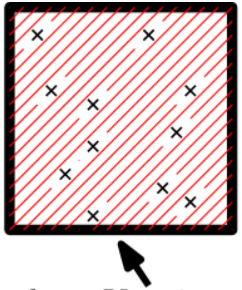


Inria

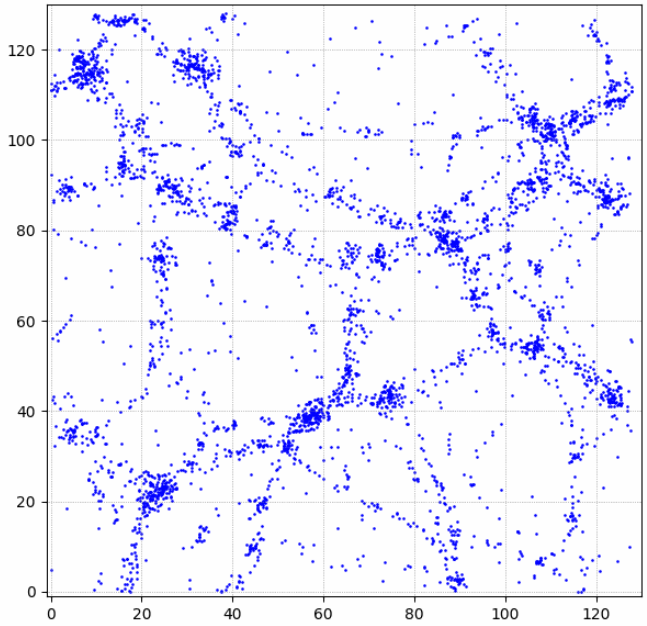
I – Percolation rate

➔ Look at ***K*-NN** instead of **1-NN**

In the percolation on $\mathbb{Z}^d$, a site is open if it contains more than ***K*** points



More than K points opens the Site



I – Percolation rate

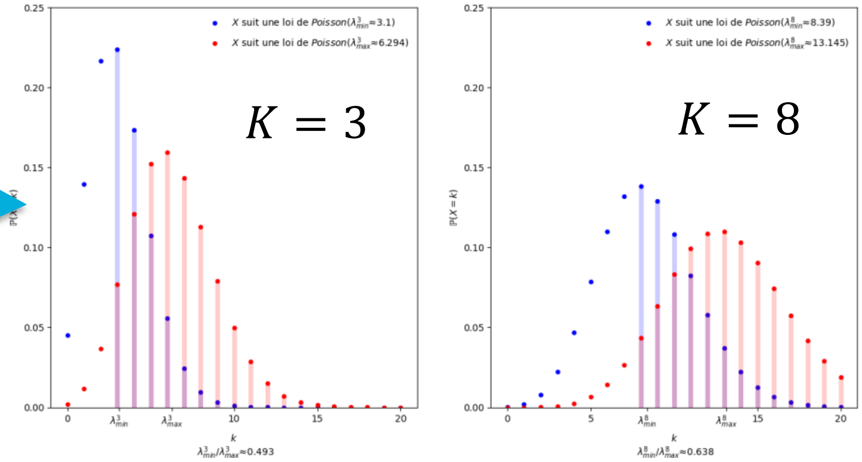
➔ Look at ***K*-NN** instead of **1-NN**

In the percolation on $\mathbb{Z}^d$, a site is open if it contains more than ***K*** points

Activation proba $\boldsymbol{p}$ = Queue of a **Poisson law**.

We* compute the intensities of $\boldsymbol{p_{detection}}$ and $\boldsymbol{p_{recovering}}$

$$\lambda_{min}^{3}/\lambda_{max}^{3} \approx 0.493 \quad < \quad \lambda_{min}^{8}/\lambda_{max}^{8} \approx 0.638$$

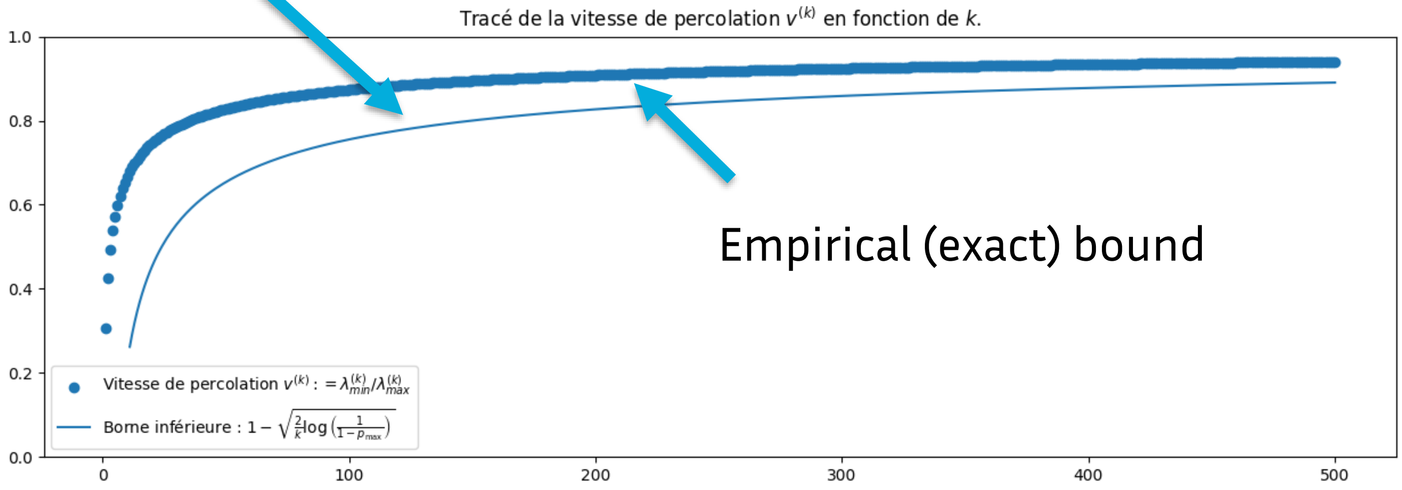


* Unpublished work

I – Percolation rate

➔ Look at ***K*-NN** instead of **1-NN**

Theoretical bound*: the **percolation rate** $v^K = 1 - O\left(\frac{1}{\sqrt{K}}\right)$



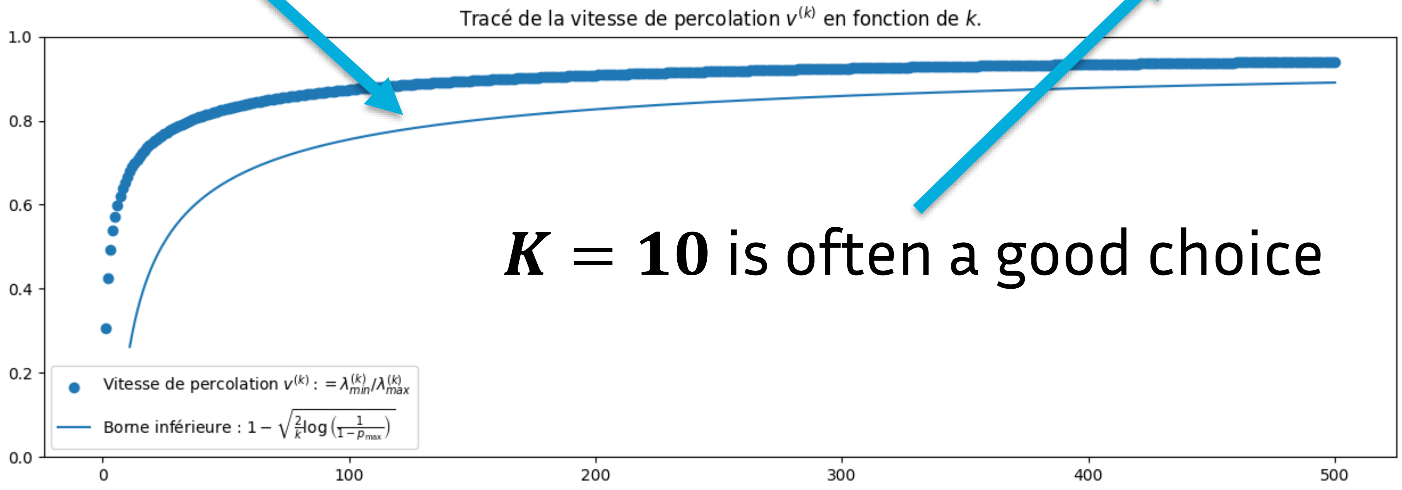
Empirical (exact) bound

* **Unpublished work**

I – Percolation rate

➔ Look at ***K*-NN** instead of **1-NN**

Theoretical bound: the **percolation rate** $\boldsymbol{v^K = 1 - O\left(\frac{1}{\sqrt{K}}\right)}$



$\boldsymbol{K = 10}$ is often a good choice

I – Robust Single Linkage

➔ Weakness of RSL (or DBSCAN)

Algorithm 1 Robust Single-Linkage *

Input: $\mathcal{X}$ the cloud point, $K \in \mathbb{N}$ and $\alpha \in [1;2]$ two parameters
Output: A hierarchical clustering $r \mapsto \hat{H}(r)$

for $x_i \in \mathcal{X}$ **do**
  $R_K(x_i) \leftarrow \inf \{r \mid |\mathcal{X} \cap B(x_i, r)| \geq K\}$
end for
for $r \in \{R_K(x_i) \mid x_i \in \mathcal{X}\}$ **do**
  $G_r(V, E) \leftarrow$ the graph with nodes

- $V = \{x_i \mid R_K(x_i) \leq r\}$
- and edges $E = \{\{x_i, x_j\} \mid ||x_i - x_j|| \leq \alpha r\}$

  $\hat{H}(r) \leftarrow ConnectedComponents\,(G_r)$
end for
Return $\hat{H}$

Strong assomption on nodes

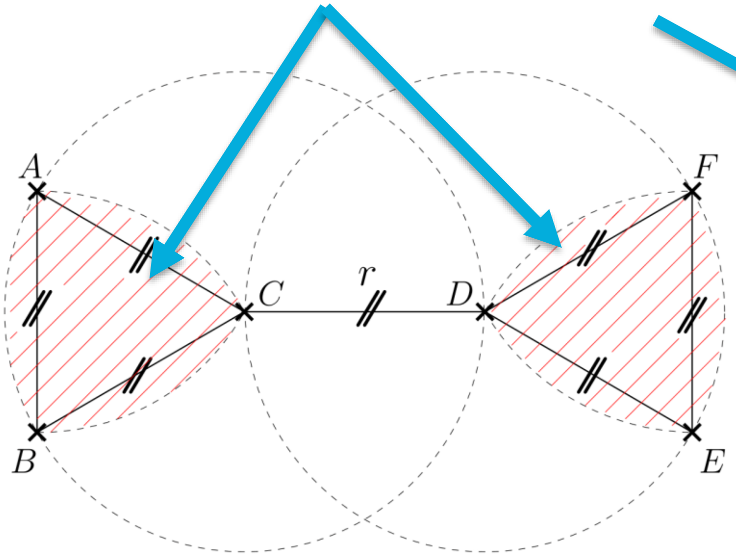
Weak assomption on connectivity

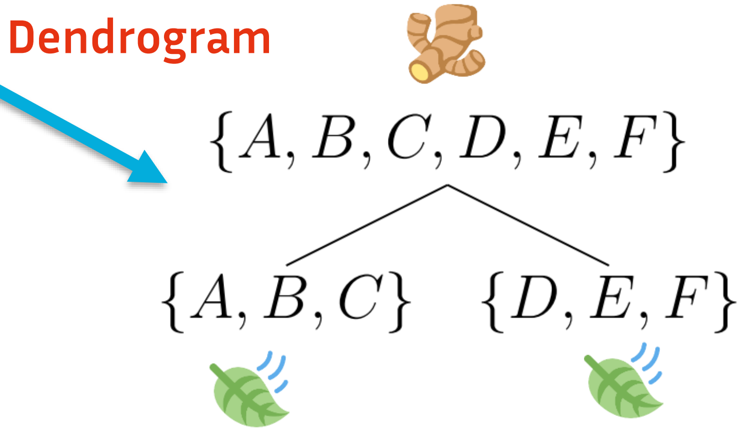
* K. Chaudhuri and S. Dasgupta: “Rates of convergence for the cluster tree” (2010)

I – Robust Single Linkage

➔ Weakness of RSL (or DBSCAN)

The High-Density Clusters of level r for ***3-NN***





I – Robust Single Linkage

➔ Weakness of RSL (or DBSCAN)

The **Dendrogram** of the Robust Single-Linkage …

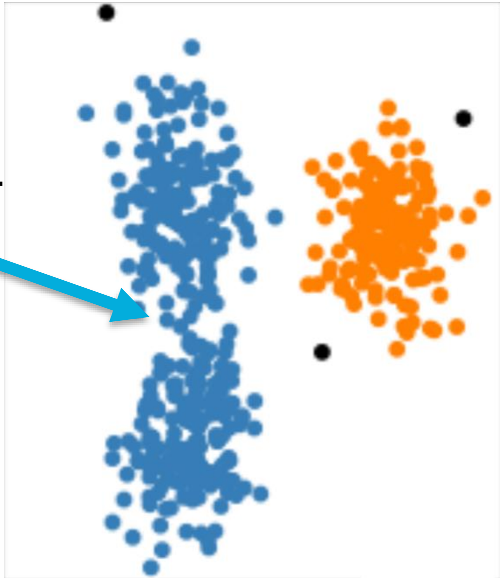
$$\{A, B, C, D, E, F\}$$

I – Robust Single Linkage

➔ Weakness of RSL (or DBSCAN)

Concrete case: Results of HDBSCAN algorithm. © Scikit-Learn's (Python library) Clustering page

The two clusters merge…



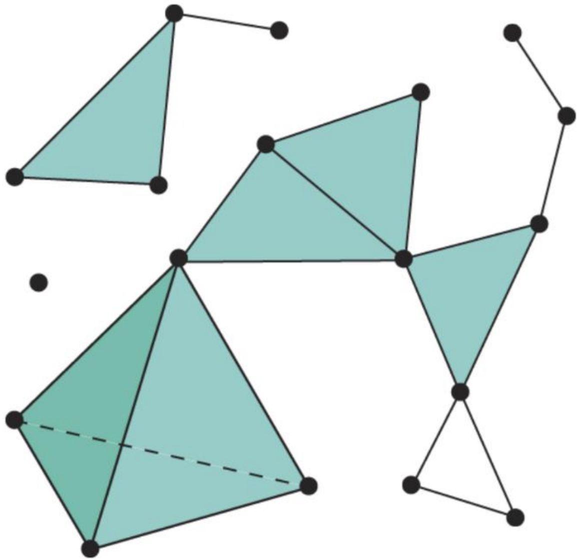
I – High-Density Clusters ~ Hypergraphs

Subsets of $\mathbb{R}^{\boldsymbol{d}}$ .... Ok for $\boldsymbol{d} = \mathbf{2}$ or $\mathbf{3}$
... but computationally expensive for $\boldsymbol{d}$ large

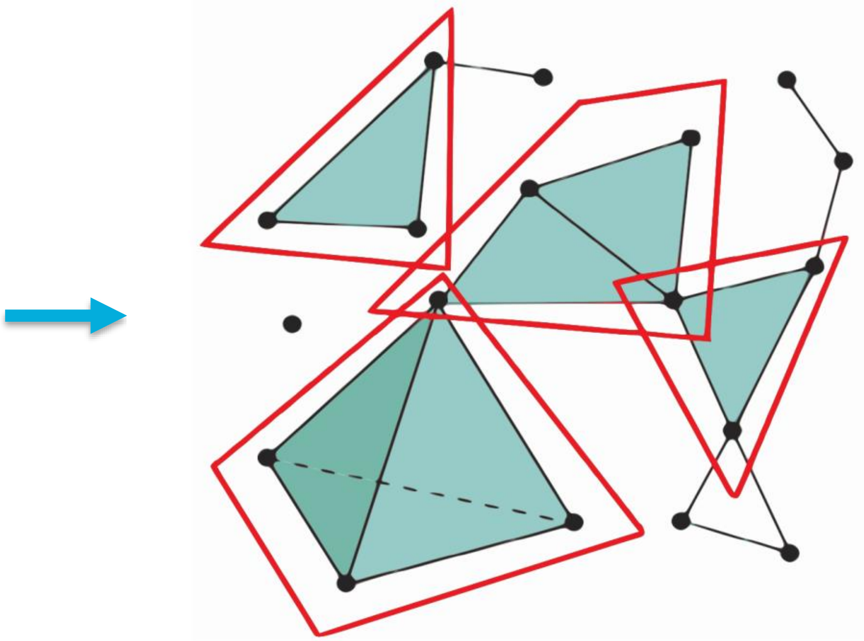
High-Density Clusters of ***1-NN*** ⟷ Connected Components of a graph

High-Density Clusters of ***K-NN*** ⟷ Connect. Comp. of an **hypergraph**

I – Hypergraphs and Q-Connectivity *



An hypergraph



Polyhedra on an hypergraph **

* R. Atkin "From cohomology in physics to q-connectivity in social science" (1972)

** Unpublished work

II – Results – Galaxy Filaments

Choose ***Centres***

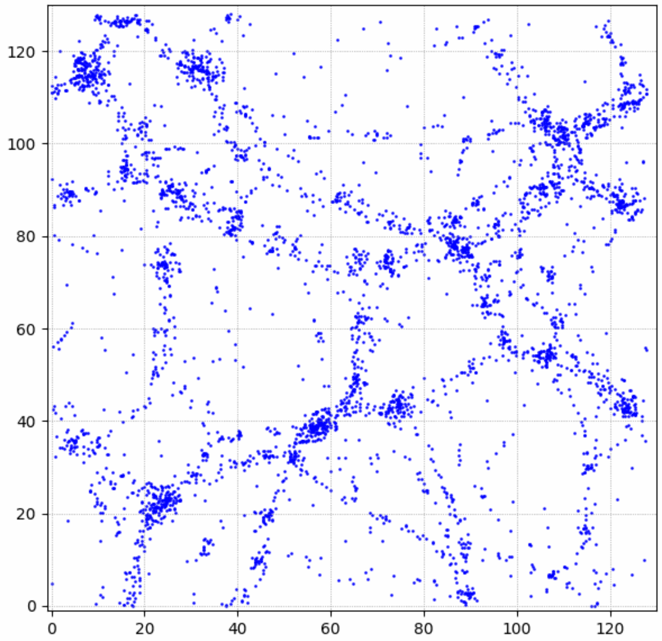
Filaments are MST on the ***Centres*** minimizing ...

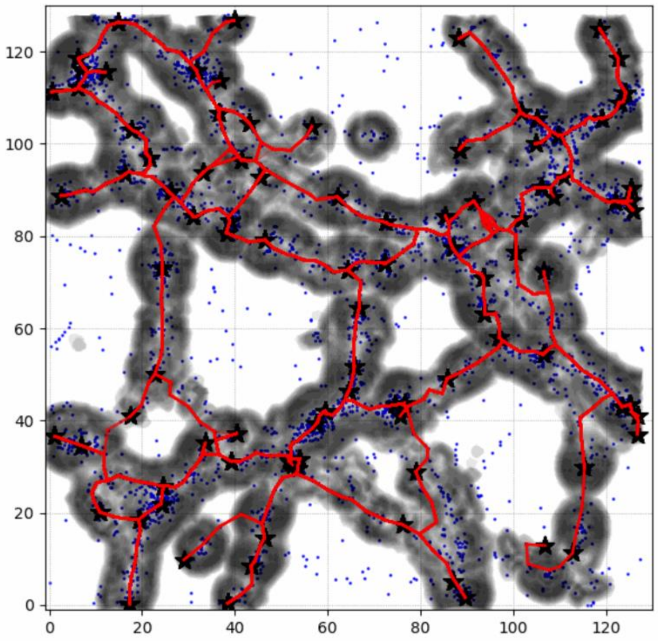
Algorithm 1 Filament extraction of a connected component $G(V, E)$

```
Centres                                         ▷ The centres of the pre-existing filament
FilNodes                                        ▷ Nodes of Filament
PercolThreshold ← 50                            ▷ The percolation threshold
while |Centres| < int(|V|/PercolThreshold) do   ▷ We search for a new centre
    D ← {}                                      ▷ The sums of the distances to minimise
    for x ∈ V do                                ▷ x is the hypothetical new centre
        NodesFilament ← copy(FilNodes)
        Branch_x ← ShortPath(x, NodesFilament)  ▷ Hypothetical new branch
        NodesFilament ← NodesFilament ∪ Branch_x
        D[x] ← 0
        for y ∈ V do
            D[x] ← D[x] + d(y, NodesFilament)^2
        end for
    end for
    x ← argmin(D)                               ▷ The new centre chosen
    Centres ← Centres ∪ {x}
    FilNodes ← FilNodes ∪ Branch_x
end while
Filament ← MinimalSpanningTree(FilNodes)
Returns Filament
```

* L. Hauseux *et al.*: "Graph Based Approach for Galaxy Filament Extraction " (2023)

II – Results – Galaxy Filaments





II – Results – Olive Italian Oil Dataset *

* M. Forina, C. Armanino, S. Lanteri, and E. Tiscornia: "Classification of olive oils from their fatty acid composition" (1983)

** S. Scaldelai, L. Matioli and M. Kleina: (2022) "Multiclusterkde: A new algorithm for clustering based on multivariate kernel density estimation"

*** A. Rolle and L. Scoccola: "Stable and consistent density-based clustering" (2023)

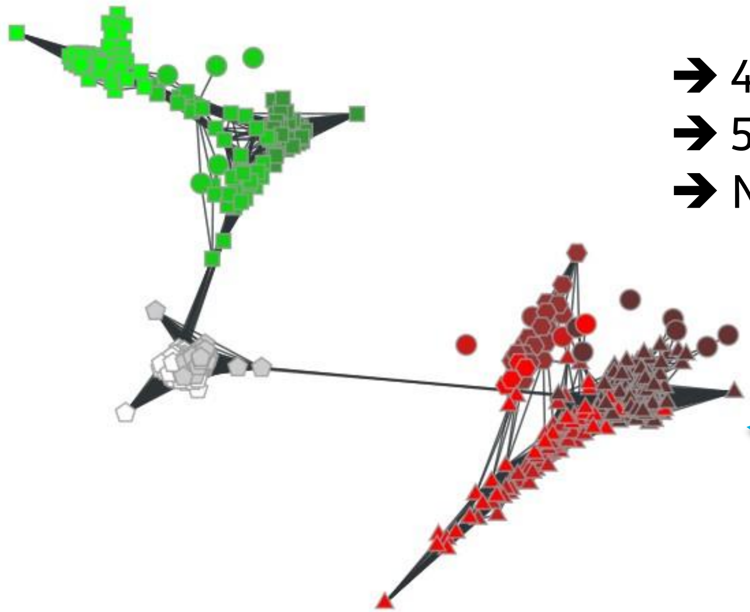
Macro-area	Region
South	1 – North Apulia 2 – Calabria 3 – South Apulia 4 – Sicilia
Sardinia	5 – Inland Sardinia 6 – Coast Sardinia
Centro-settentrionale	7 – East Liguria 8 – West Liguria 9 – Umbria

572 samples of oil composition **(vectors in $\mathbb{R}^8$)**

Fatty acids: Palmitic, Palmitoleic, Stearic, Oleic, Linoleic, Linolenic, Arachidic, eicosenoic

Can we recover the geographic clusters given only the fatty acids ?

II – Results – Olive Oil Dataset *

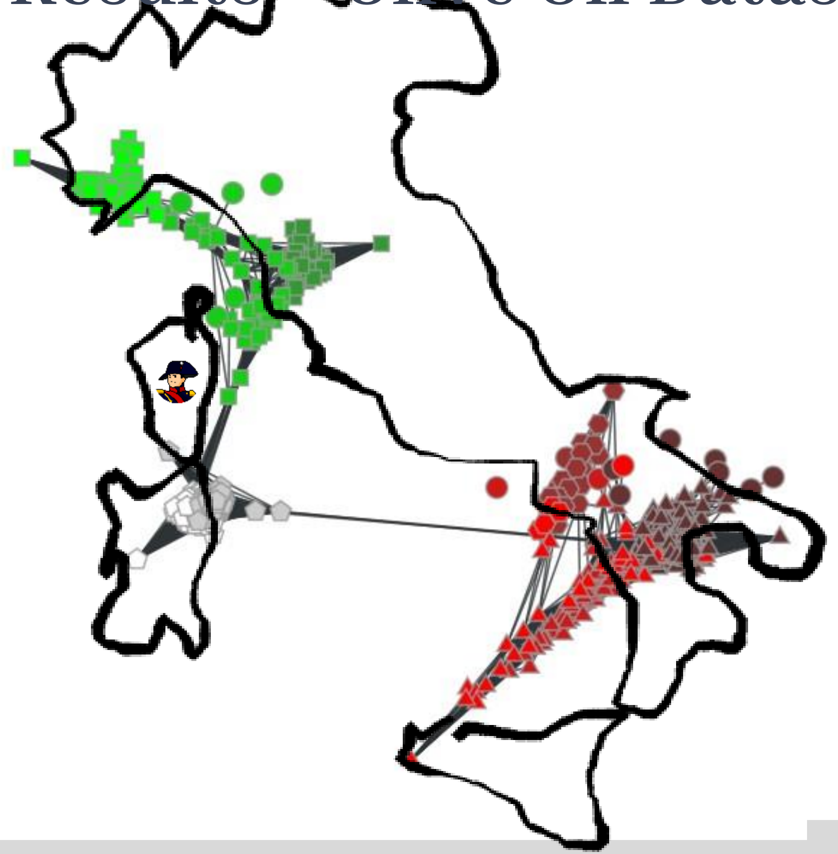


➔ 4 Clusters (Nord Poulia isolated)
➔ 555/572 Classified (**NN** for unclassif.)
➔ No error on the classified points

Ground Truth: Colour

Clustering: Shape

II – Results – Olive Oil Dataset *



Bibliography

- John A. Hartigan. *Clustering Algorithms* (1975). *John Wiley & Sons*.
- M. Forina, C. Armanino, S. Lanteri, and E. Tiscornia: « Classification of olive oils from their fatty acid composition » (1983). *IUFoST Symposium*.
- M. Penrose. *Random Geometric Graphs* (2003). *Oxford Studies in Probability*.
- R. Stoica, Vicent J. Martinez, Jorge Mateu & Enn Saar. « Detection of cosmic filaments using the Candy model » (2005). *Astronomy & Astrophysics*.
- K. Chaudhuri and S. Dasgupta. « Rates of convergence for the cluster tree » (2010). *NIPS*.
- L. McInnes and J. Healy: « Accelerated hierarchical density based clustering » (2017). *ICDMW*.
- S. Scaldelai, L. Matioli and M. Kleina: « Multiclusterkde: A new algorithm for clustering based on multivariate kernel density estimation » (2022). *J. Appl. Stat*.
- L. Hauseux & K. Avrachenkov & J. Zerubia. « Graph Based Approach for Galaxy Filament Extraction » (2023). *Intern. Conf. Of Complex Networks, Menton and HAL*.
- A. Rolle and L. Scoccola: « Stable and consistent density-based clustering » (2023). *Arxiv*.