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**
The problem: Galaxy Filament Extraction

Galaxies are not distributed uniformly
➔ ‘Large Scale Structures’
➔ Filamentary extraction

Simulated 3D image of the CosmicWeb. © MAX-PLACK INSTITUT FÜR ASTROPHYSIK

« 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.
Overview

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.

« Cluster » ou « Community » detection... What is this?
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 persistent 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 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 $H_f(r)$ of level $r$: the connected components of:

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

Important!

High-Density Clusters A and B
I - High-Density Clusters

High-Density Clusters $H_{\hat{f}}(r)$ of level $r$ for $\hat{f}$ = density estimator of 10-Nearest Neighbors
I – Single-Linkage

Practical situation: the unit square $[0; 1]^2$ is split in 2:
- The left-rectangle with high density.
- The right-one with low density.

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

Initial Clustering:

$C = \{C_1, \ldots, C_n\}$ with $C_1 = \{x_1\}, \ldots, C_n = \{x_n\}$

At each step, we merge the two closest clusters
Hierarchical Clustering produced by Single-Linkage

A tree ➔ the Dendrogram

\[ r \approx 0.7 \]

« Cluster merging phase » = Transition phase *

I – Single-Linkage ~ Geometric Graphs

Geometric graph with same radius $r \approx 0.7$

We almost recover the High-Density Cluster
We almost recover the High-Density Cluster... Why?

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

\[ \hat{f}_{1-NN}(y) = \frac{1}{R_y^d} \]

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:

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

Create a graph \( G \) on the cloud point \( \mathcal{X} \) such that:

High-Density Clusters \( \leftrightarrow \) Connected Components of \( G \)

Subsets of \( \mathbb{R}^d \) .... Ok for \( d = 2 \) or 3

... but computationally expensive for \( d \) large
I - Percolation

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

→ Density $\lambda$ of Poisson Point Process

Percolation = giant component

→ Very fast phenomenon

Two sub-cases: there exists a critical value $\lambda_c \in (0; \infty)$:

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

3 Geometric Graphs / greatest component

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

*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 **not consistent** in $\mathbb{R}^d$ for $d \geq 2$ ...

Maximal interval between two points of $\mathcal{X}$
I – Consistency and Single-Linkage

⇒ Single-Linkage is **not consistent** in $\mathbb{R}^d$ for $d \geq 2$ ... !!!

Discrete Site-Percolation on $\mathbb{Z}^2$
I – Consistency and Single-Linkage

⇒ Single-Linkage is not consistent in $\mathbb{R}^d$ for $d \geq 2$ ... !!!

Discrete Site-Percolation on $\mathbb{Z}^2$
Consistency and Single-Linkage

Single-Linkage is not 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 not consistent in $\mathbb{R}^d$ for $d \geq 2$ ... !!!

⇒ Only « fractional consistency » is possible

⇒ Goal: obtain the largest possible « fraction »
1 - Rate of Percolation

Fraction of points in the giant component

Averaged over 100 realizations of RGGs(λ, 1) on a 251x251 area. Then applying $r = \sqrt{\lambda}$. 

$\rho_v(\epsilon)$

$1 - \epsilon$

$\epsilon$

$1$

$r_{\min}$

$r_{\max}$

$2$
I - Rate of Percolation

The thresholds of « detection » and « recovering »
1 – Percolation rate

Percolation rate* \( = \frac{r_{\text{min}}}{r_{\text{max}}} \)

*A. 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.
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 \( K-\text{NN} \) instead of \( 1-\text{NN} \)

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

Activation proba \( p = \text{Queue of a Poisson law} \).

We* compute the intensities of \( p_{\text{detection}} \) and \( p_{\text{recovering}} \).

\[
\frac{\lambda_{\min}^3}{\lambda_{\max}^3} = 0.493 < \frac{\lambda_{\min}^8}{\lambda_{\max}^8} = 0.638
\]

\* 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

$$\nu^K = 1 - O\left(\frac{1}{\sqrt{K}}\right)$$

$K = 10$ is often a good choice
Robust Single Linkage

Weakness of RSL (or DBSCAN)

Algoirthm 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 assumption on nodes
Weak assumption on connectivity

Weakness of RSL (or DBSCAN)

The High-Density Clusters of level \( r \) for 3-NN

Dendrogram

\[ \{ A, B, C, D, E, F \} \]

\[ \{ A, B, C \} \quad \{ D, E, F \} \]
I – Robust Single Linkage

⇒ Weakness of RSL (or DBSCAN)

The Dendrogram of the Robust Single-Linkage ...

\{A, B, C, D, E, F\}
1 - 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 $\mathbb{R}^d$ .... Ok for $d = 2$ or $3$
... but computationally expensive for $d$ large

High-Density Clusters of $1$-NN $\leftrightarrow$ Connected Components of a graph

High-Density Clusters of $K$-NN $\leftrightarrow$ 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
\section*{Results – Galaxy Filaments}

Choose \textbf{Centres}

\begin{algorithm}
\caption{Filament extraction of a connected component $G(V, E)$}
\begin{algorithmic}
\State \textbf{Centres} $\triangleright$ The centres of the pre-existing filament
\State \textbf{FilNodes} $\triangleright$ Nodes of Filament
\State \textbf{PercolThreshold} $\leftarrow$ 50 $\triangleright$ The percolation threshold
\While{$|\textbf{Centres}| < \text{int}(|V|/\text{PercolThreshold})$}
\State $D \leftarrow \{}$ $\triangleright$ The sums of the distances to minimise
\For{$x \in V$}
\State $\textbf{NodesFilament} \leftarrow \text{copy}($\textbf{FilNodes}$)$
\State $\text{Branch}_x \leftarrow \text{ShortPath}(x, \text{NodesFilament})$ $\triangleright$ Hypothetical new branch
\State $\text{NodesFilament} \leftarrow \text{NodesFilament} \cup \text{Branch}_x$
\State $D[x] \leftarrow 0$
\For{$y \in V$}
\State $D[x] \leftarrow D[x] + d(y, \text{NodesFilament})^2$
\EndFor
\EndFor
\State $x \leftarrow \text{argmin}(D)$ $\triangleright$ The new centre chosen
\State $\textbf{Centres} \leftarrow \textbf{Centres} \cup \{x\}$
\State $\textbf{FilNodes} \leftarrow \text{FilNodes} \cup \text{Branch}_x$
\EndWhile
\State $\text{Filament} \leftarrow \text{MinimalSpanningTree}($\text{FilNodes}$)$
\State \textbf{Returns} $\text{Filament}$
\end{algorithmic}
\end{algorithm}

\textbf{Filaments} are MST on the \textbf{Centres} minimizing ...

\textbf{* 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)

<table>
<thead>
<tr>
<th>Macro-area</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>1 – North Apulia</td>
</tr>
<tr>
<td></td>
<td>2 – Calabria</td>
</tr>
<tr>
<td></td>
<td>3 – South Apulia</td>
</tr>
<tr>
<td></td>
<td>4 – Sicilia</td>
</tr>
<tr>
<td>Sardinia</td>
<td>5 – Inland Sardinia</td>
</tr>
<tr>
<td></td>
<td>6 – Coast Sardinia</td>
</tr>
<tr>
<td>Centro-settentrionale</td>
<td>7 – East Liguria</td>
</tr>
<tr>
<td></td>
<td>8 – West Liguria</td>
</tr>
<tr>
<td></td>
<td>9 – Umbria</td>
</tr>
</tbody>
</table>

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