



HAL
open science

How can we theoretically measure the performance of density-based clustering algorithms?

Louis Hauseux

► **To cite this version:**

Louis Hauseux. How can we theoretically measure the performance of density-based clustering algorithms?. ACM SIGMETRICS 2024 Student Research Competition, Jun 2024, Venice, Italy. hal-04674019

HAL Id: hal-04674019

<https://hal.science/hal-04674019>

Submitted on 21 Aug 2024


HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

How can we theoretically measure the performance of density-based clustering algorithms?

Louis HAUSEUX
 Inria, Université Côte d’Azur
 NEO & AYANA teams
 Sophia Antipolis, France
 louis.hauseux@inria.fr 

Categories and Subject Descriptors

Measurement [Metrics]: Performance—we define the “percolation rate” to measure the performance of density-based clustering algorithms; Theory [Methodologies]: Analytical modeling techniques and model validation—*Analysis of the mathematical phenomenon behind these algorithms: the percolation*

Keywords

density-based clustering, percolation, geometric graphs, Nearest Neighbors density estimator

1. INTRODUCTION

Many of clustering algorithms for a point cloud $\mathcal{X}_n \subset \mathbb{R}^d$ in the Euclidean space are based on density estimates [1]. In fact, the density function f of point generation contains the relevant information. It is quite natural to try to extract what HARTIGAN called ‘high-density clusters’ [2].

One elegant solution to do this task consists in constructing a graph whose nodes are the points of the cloud and whose edges connect nearby points. We want the connected components of this graph to reflect the *high-density clusters*.

Some very classical algorithms such as (Robust) Single-Linkage [3] or (H)DBSCAN [4, 5] work in this way. It is particularly helpful because its connected components correspond exactly to the *high-density clusters* of the estimator $\hat{f}_{1\text{-Nearest Neighbor}}$.

An example of the Single-Linkage will show us the mathematical phenomenon at the heart of these algorithms: the *percolation* [6, 7].

We define and measure the *percolation rate* to evaluate the performance of such algorithms.

By way of example, we look at the Robust Single-Linkage algorithm and calculate its percolation rate. This will show theoretically why it is actually preferable to use K -Nearest Neighbours rather than 1-NN. However, convergence in K towards a perfect estimator is very slow, so this analysis explains why in practice $K = 10$ is often a good trade-off.

2. MATHEMATICAL MODEL

Let $\mathcal{X}_n := \{x_1, \dots, x_n\}$ be a cloud of n points all plotted IID according to a probability measure with density $f : \mathbb{R}^d \rightarrow \mathbb{R}_+$. With the very intuitive idea that the different

classes of the point cloud are represented by the “peaks” of the density function f , HARTIGAN [2] defines the high-density clusters of level r as the different connected components of the level set $L_r := \{x \in \mathbb{R}^d : f(x) \geq r\}$.

By varying the level r , we can obtain an *hierarchical clustering* representable by a tree, the *dendrogram*.

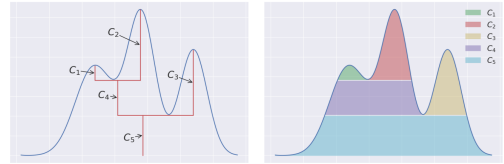


Figure 1: Hierarchical clustering obtained via density level sets. © Images taken from [5].

3. SINGLE-LINKAGE AND PERCOLATION

The *Single-Linkage* algorithm constructs a hierarchical clustering as follows: It starts with the trivial initial clustering (n points for n clusters) $\mathcal{C}_0 = \{C_1^0, \dots, C_n^0\}$ with $C_i^0 = \{x_i\}$. At each step, we merge the two clusters that are closest for the distance: $d_{Clust}(C, C') = \min_{x \in C, y \in C'} \|x - y\|$.

At step t , the resulting clustering $\mathcal{C}_t = \{C_1^t, C_2^t, \dots, C_{n-t}^t\}$ corresponds to the connected components of a geometric graph $\mathcal{G}(\mathcal{X}_n, r_t)$ built on \mathcal{X}_n [7]. Therefore, Single-Linkage performs *persistent analysis* on geometric graphs $\mathcal{G}(\mathcal{X}_n, r)$. Furthermore, the connected components of $\mathcal{G}(\mathcal{X}_n, r)$ match exactly with the high-density clusters of the 1-Nearest Neighbor density estimator $\hat{f}_{1\text{-NN}}$.

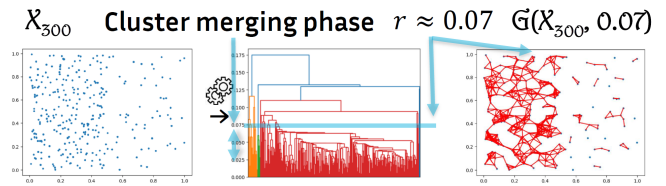


Figure 2: From left to right: 1) \mathcal{X}_{300} 2) Dendrogram of Single-Linkage 3) Geometric graph $\mathcal{G}(\mathcal{X}_{300}, 0.07)$.

See Fig. 2 for an illustration. The density is constant on each half-rectangle (left and right), and is larger on left-hand side. We observe on the dendrogram the first percolation phase (on left). Suddenly, for $r \lesssim 0.07$, plenty of clusters merge. The associated geometric graph $\mathcal{G}(\mathcal{X}_{300}, 0.07)$ has a giant component almost corresponding to the left high-

I would like to thank my supervisors Konstantin AVRACHENKOV (Inria, NEO team) and Josiane ZERUBIA (Inria, AYANA team) for advice.

density cluster. This sudden appearance of a giant connected component is called *percolation* [6, 7].

4. PERCOLATION RATE

HARTIGAN [2] showed that the Single-Linkage algorithm is a consistent estimator of high-density clusters in dimension $d = 1$, but only *fractionally* consistent in dimension $d \geq 2$.

How can we measure this recoverable ‘fraction’? This led us to define a *percolation rate* in [8]. Percolation is a ‘fast’ phenomenon. Once it appears for a critical radius r_c , the proportion $p_\infty(r)$ of points within the giant component grows quickly to 1. From a certain radius r_{\min} , the giant component contains $\varepsilon \leftarrow 5\%$ of the point and becomes detectable. For another larger radius $r_{\max} > r_{\min}$, the giant component encompass almost all the points (a proportion $1 - \varepsilon$): the cluster is recovered. Namely, let $r_{\min} := p_\infty^{-1}(\varepsilon)$ and $r_{\max} := p_\infty^{-1}(1 - \varepsilon)$. The quantity of interest – we call the *percolation rate* v – is:

$$v := \frac{r_{\min}}{r_{\max}}.$$

Note that $v \leq 1$. This percolation rate depends on the

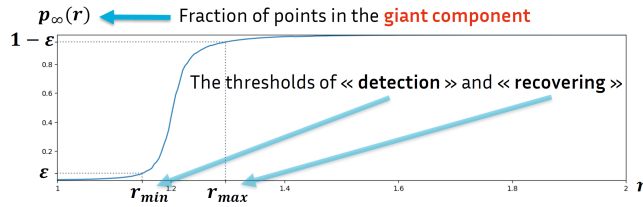


Figure 3: Estim. of $p_\infty(r)$ in \mathbb{R}^2 on random geometric graphs [7]. $\varepsilon \leftarrow 0.05 \implies r_{\min} = 1.15, r_{\max} = 1.30$.

kind of objects one ‘percolates’ and the associated notion of ‘connected components’. For example, in order to increase v , it would be a good idea to consider *hypergraphs* rather than classical graphs [9]. A percolation rate $v = 1$ is synonymous of an almost perfect clustering.

5. PERCOLATION RATE OF DISCRETIZED ROBUST SINGLE-LINKAGE

To gain in robustness, a *robust* version of the Single-Linkage was proposed [3], inspired by the consistency of the K -NN density estimator [10].

The main difference is that in the K -Robust version, a point $x \in \mathcal{X}_n$ must have at least $K - 1$ other points in his r -neighbourhood to appear in the geometric graph.

Let us consider the grid \mathbb{Z}^d rather than \mathbb{R}^d . A cell $s \in \mathbb{Z}^d$ – a *site* in the jargon of percolation – is activated (= *open*) once it contains K points of an homogeneous Poisson point process $\mathcal{X} \subset \mathbb{R}^d$. Then, we look at the *clusters* formed by

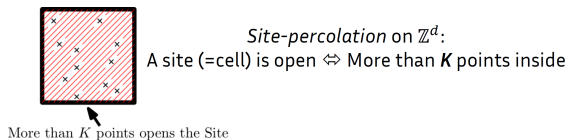


Figure 4: Discretization of the Robust Single-Linkage (RSL) on \mathbb{Z}^d : we look at *site-percolation*.

the open sites: this is the *discretized* Robust Single-Linkage.

On the grid \mathbb{Z}^d , we are now able to compute the exact percolation rate v^K of the K -discretized Robust Single-Linkage. In fact, the probability of a site to be open is $p = \mathbb{P}[\mathcal{P} \geq K]$ where $\mathcal{P} \sim \text{Poisson}(\lambda)$, λ being the intensity of the Poisson point process. We define as previously λ_{\min}^K and λ_{\max}^K and

$$v^K := \frac{\lambda_{\min}^K}{\lambda_{\max}^K}.$$

THEOREM. *Discretized RSL is asymptotically almost instantaneous: $v^K = 1 - O\left(\frac{1}{\sqrt{K}}\right)$. The constant for $\frac{1}{\sqrt{K}}$ depends on ε and the dimension d .*

Since v^K increases slowly towards 1 (in $1/\sqrt{K}$), it is easy to understand why the choice of $K \leftarrow 10$ is often a good trade-off in practice.

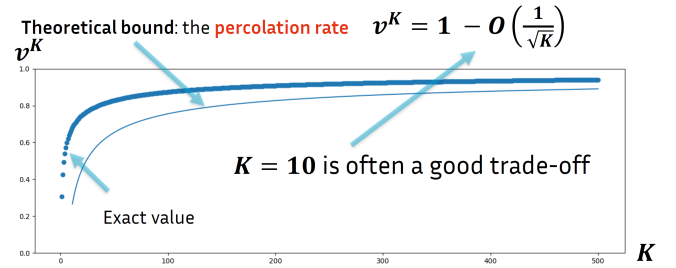


Figure 5: Theoretical bound and exact value of the discretized RSL percolation rate in dimension $d = 2$.

References

- [1] A. Rolle and L. Scoccola, *Stable and consistent density-based clustering*, arXiv.v3, 2023. DOI: 10.48550/arXiv.2005.09048. arXiv: 2005.09048 [math.ST].
- [2] J. A. Hartigan, “Consistency of single linkage for high-density clusters,” *J. of the Am. Stat. Ass.*, vol. 76, no. 374, pp. 388–394, 1981. DOI: 10.1080/01621459.1981.10477658.
- [3] K. Chaudhuri and S. Dasgupta, “Rates of convergence for the cluster tree,” in *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23, Curran Associates, Inc., 2010.
- [4] R. J. G. B. Campello, D. Moulavi, and J. Sander, “Density-based clustering based on hierarchical density estimates,” in *Advances in Knowledge Discovery and Data Mining*, Berlin, Heidelberg: Springer, 2013, pp. 160–172, ISBN: 978-3-642-37456-2. DOI: 10.1007/978-3-642-37456-2_14.
- [5] L. McInnes and J. Healy, “Accelerated hierarchical density based clustering,” in *IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017, pp. 33–42. DOI: 10.1109/ICDMW.2017.12.
- [6] G. Grimmett, *Percolation*. Springer, 1999, ISBN: 978-3-540-64902-1. DOI: 10.1007/978-3-662-03981-6.
- [7] M. Penrose, *Random Geometric Graphs*. Oxford University Press, 2003, vol. 5. DOI: 10.1093/acprof:oso/9780198506263.001.0001.
- [8] L. Hauseux, K. Avrachenkov, and J. Zerubia, “Graph based approach for galaxy filament extraction,” in *Complex Networks & Their Applications XII*, Springer, 2024, pp. 384–396. DOI: 10.1007/978-3-031-53472-0_32.
- [9] L. Hauseux, K. Avrachenkov, and J. Zerubia, “Benefits of hypergraphs for density-based clustering,” in *32nd European Signal Processing Conference (EUSIPCO)*, Submitted, Lyon, France, 2024.
- [10] G. Biau and L. Devroye, *Lectures on the Nearest Neighbor Method*. Springer, 2015, vol. 246, ISBN: 978-3-319-25386-2. DOI: 10.1007/978-3-319-25388-6.