



**HAL**  
open science

# Graph Based Approach for Galaxy Filament Extraction

Louis Hauseux, Konstantin Avrachenkov, Josiane Zerubia

► **To cite this version:**

Louis Hauseux, Konstantin Avrachenkov, Josiane Zerubia. Graph Based Approach for Galaxy Filament Extraction. Complex Networks 2023 - The 12th International Conference on Complex Networks and their Applications, Nov 2023, Menton, France. 10.1007/978-3-031-53472-0\_32 . hal-04231772

**HAL Id: hal-04231772**

**<https://inria.hal.science/hal-04231772v1>**

Submitted on 6 Oct 2023




**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 4.0 International License

# Graph Based Approach for Galaxy Filament Extraction

Louis Hauseux <sup>\*\*</sup>, Konstantin Avrachenkov , and Josiane Zerubia 

Inria, Université Côte d’Azur, Sophia-Antipolis, France.  
FirstName.LastName@inria.fr,

**Abstract.** We propose an original density estimator built from a cloud of points  $\mathcal{X} \subset \mathbb{R}^d$ . To do this, we consider geometric graphs  $\mathcal{G}(\mathcal{X}, r)$  on the cloud. These graphs depend on a radius  $r$ . By varying the radius, we see the emergence of large components around certain critical radii, which is the phenomenon of *continuum percolation*. Percolation allows us to have both a local view of the data (through local constraints on the radius  $r$ ) and a global one (the emergence of macro-structures). With this tool, we address the problem of galaxy filament extraction. The density estimator gives us a relevant graph on galaxies. With an algorithm sharing the ideas of the Fréchet mean, we extract a subgraph from this graph, the galaxy filaments.

**Keywords:** geometric graphs, continuum percolation, Fréchet mean, galaxy filaments

## 1 Introduction

At scales of billions light-years, the observable universe—matter and light—does not follow a uniform distribution but forms what are known as ‘large-scale structures’ [3, 13]. These structures seem arranged hierarchically: 1° super-clusters of galaxies (hyper-dense small volumes, sometimes called ‘knots’ or ‘nodes’); 2° ‘sheets’ or ‘walls’ of galaxies ; 3° ‘filaments’ of galaxies. These different clusters delimit large “voids” regions that are virtually empty of galaxies: they shape the “cosmic web”, like a giant sponge or a spider’s web.

Astronomical surveys [1, 2] now contain millions of galaxies, making it impossible to extract these structures with the naked eye. Various types of algorithms have been proposed to extract automatically these clusters, and particularly the galaxy filaments. (*Cf.* the survey “Tracing the cosmic web” [22]). Most are based on density estimators<sup>1</sup> (two comparative studies: [12, 14]). The density-based

---

<sup>\*\*</sup> The first author would like to thank the Université Côte d’Azur (UCA) DS4H Investments in the Future project managed by the National Research Agency (ANR, reference number ANR-17-EURE-0004) and 3IA Côte d’Azur for partial funding of his PhD thesis. All the authors acknowledge a partial support by Nokia Bell Labs “Distributed Learning and Control for Network Analysis” and Bpifrance in collaboration with Airbus D&S (LiChIE contract, 2020-2024).

<sup>1</sup> Those ‘filaments’ are relatively thick and have a non-negligible width; a real 1D-manifold would not have a density w.r.t. the Lebesgue measure.

methods are often based on the Delaunay density estimator DTFE [28], estimator derived from Delaunay triangulation; the estimated density being inversely proportional to the area of the neighbouring triangles (the analogous exists with Voronoï tiling). We will look at another classical density estimator used: The  $K$ -Nearest Neighbours ( $K$ -NN) algorithm [6]. This very simple algorithm can produce—with some refinements—impressive results. For example, the HDB-SCAN hierarchical clustering algorithm [10] is based on the High-Density Levels of the  $K$ -Nearest Neighbours density estimator.

Obtaining a filamentary structure naturally led to introduce graphs on the galaxies (considered as points in space). As early as 1985, the pruned minimal spanning tree [5] was proposed as a filamentary model.

More interesting is the idea proposed by Colberg [11] who also pruned minimal spanning tree and studied what happens at the percolation stages. The percolation thresholds are directly linked to the types of structure that appear. This is the key point to observe.

The Delaunay and  $K$ -Nearest Neighbours estimators have only a local view of the data.  $K$ -NN estimator has good properties (consistency, calculation speed, see monograph by Biau & Devroye [6]). However, obtaining consistency requires that  $k$  tends to infinity. In practice,  $k$  is taken smaller than 10 and the number of galaxies is insufficient too much hope in this estimator.

Percolation is a phenomenon which, under local constraints, can be observed macroscopically: It is the precise moment when macro-structures appear. Percolation allows us to have both a local view of the data (through local constraints on the graph) and a global one (through the emergence of macro-structures).

If we assume galaxies are IID points plotted in space by an unknown measure of density  $f$ , thanks to the hierarchical structures, we could identify galaxy clusters with the highest density clusters [18, 25], *i.e.*  $f^{-1}([h; +\infty)) = \{x \in \mathbb{R}^d \mid f(x) \geq h\}$ .

In this article we propose a new estimator for *density levels* and filament extraction using geometric graphs [25]. If  $\mathcal{X}$  denotes the cloud points and  $\mathcal{G}(\mathcal{X}, r)$  the geometric graph of radius  $r$  built on these points, we vary  $r$  from 0 until the percolation phases. At each radius  $r$ , we associate a cluster  $\Sigma_r \subset \mathbb{R}^d$ , the *density level* for radius  $r$ .  $\Sigma_r$  increases with  $r$  (like  $\mathcal{G}(\mathcal{X}, r)$ ).

An intuitive idea for extracting filaments from the cluster  $\Sigma_r$  is to take its medial axis [4, 8]. However, we would not take advantage of the persistent information ( $r$  can vary), nor the fact that we have a graph (with an induced distance). This is why we prefer to proceed as follows: increasing the radius  $r$  until big components in  $\mathcal{G}(\mathcal{X}, r)$  appear. At this moment, we initialise a new filament within the big component by its Fréchet mean [15]. As the component grows with  $r$ , we add points to the filament so that the augmented filament satisfies a minimum condition (similar to Fréchet’s mean minimization).

We show an example of such filament extraction on a synthetic 2D-image of galaxies. At a glance, we compare our results with a stochastic method [29].

For a quantified comparison, we compare our density level estimator with conventional density estimators for this type of problem (Delaunay estimator,

$K$ -Nearest Neighbours) on point cloud generated by a known density function  $f$ . Our estimator is already showing very good results, especially for high-density clusters.

## 2 Preliminaries

In this section, we introduce the mathematical background.

**Geometric graphs** Given a set  $\mathcal{X}$  of points in  $R \subset \mathbb{R}^d$  and a radius  $r$ , the geometric graph  $\mathcal{G}(\mathcal{X}, r)$  is the undirected graph whose nodes are the points in  $\mathcal{X}$ , and whose edges join all the nodes that are at a distance less than  $r$ .

**Percolation phenomenon [9, 23, 25]** Let  $\mathcal{H}_\lambda$  be a Poisson point process on  $\mathbb{R}^d$  of intensity  $\lambda$  and  $\mathcal{H}_{\lambda,0} := \mathcal{H}_\lambda \cup \{0\}$ . Denote  $p_1(\lambda)$ ,  $p_2(\lambda)$ ,  $\dots$  the probabilities that the component containing origin in  $\mathcal{G}(\mathcal{H}_{\lambda,0}, 1)$  has exactly 1, 2,  $\dots$  nodes. And  $p_\infty(\lambda)$  the *percolation probability* (this component is of infinite size):

$$p_\infty(\lambda) := 1 - \sum_{k=1}^{\infty} p_k(\lambda).$$

$p_\infty(\cdot)$  is an increasing function of  $\lambda$ ; there exists a *critical value*  $\lambda_c$  (which depends on the dimension  $d$  of the space) below which  $p_\infty(\lambda) = 0$  (for  $\lambda < \lambda_c$ ) and above which  $p_\infty(\lambda) > 0$  (for  $\lambda > \lambda_c$ ). In the latter case,  $p_\infty(\lambda)$  can be seen as the proportion of points that fall into the giant component (the second component being of negligible size compared to the first).

This phenomenon of percolation, *i.e.* the appearance of a giant connected component, is very interesting for modelling and studying numerous problems. For example, the spread of a forest fire (the nodes being the trees, the neighbourhood radius  $r$  the threshold below which a tree devoured by flames sets fire to its neighbours). We can then deduce from the density of the forest whether an outbreak of fire is likely to be naturally confined to a limited area or not.

### 2.1 Density estimator and density levels

Various indicators exist for comparing probability measures, such as the Kullback-Leibler divergence [14, 21] or Wasserstein distance [24, 32]. But what do these tools mean when the provided density estimator is not integrable, like the  $K$ -Nearest Neighbours estimator? There is a much stronger objection: The Kullback-Leibler divergence and the Wasserstein distance do not take into account the specific features of our problem: galaxy clusters to be identified are highly hierarchical. We are asking for good *relative* accuracy (preserving density hierarchy), not necessary *absolute*.

To have a tool that conforms to the hierarchical structure of clusters, we are going to define a notion of density level inspired by the ‘‘High-Density Clusters’’ introduced by Hartigan [18], *cf.* also the introduction of Penrose’s book [25].

The High-Density Clusters of level  $h$  are the different connected components of  $f^{-1}([h; +\infty))$ . Hartigan [19] showed that the connected components of geometric graphs is a consistent estimator of these clusters in dimension 1.

Let  $P \in [0; 1]$  be a parameter representing the proportion of classified points (those of highest density). To this proportion  $P$ , we can associate the density-height  $h_P$  defined as follows:

$$h_P = \inf \left\{ h \mid \int_{f \geq h} f(x) dx \leq P \right\}.$$

Now, given a point  $x \in \mathbb{R}^d$ , we attribute to  $x$  the first  $P$  such that  $x$  lies into one of the clusters of level  $h_P$ :

$$\mathcal{P} : x \in \mathbb{R}^d \mapsto \mathcal{P}(x) := \inf \{ P \in [0; 1] \mid x \in f^{-1}([h_P; +\infty)) \}.$$

Intuitively,  $\mathcal{P}(x)$  represents the proportion of points that must be taken in the cloud points for  $x$  to appear in one of the High-Density Clusters.

For convenience, we will consider the function  $1 - \mathcal{P}$  instead, which is thus an increasing function of the density. It is this  $1 - \mathcal{P}$  function that we call the *map of density levels*.

This hierarchical classification is perfectly suitable if we have a good estimate of the proportion of galaxies which lie in each kind of clusters (superclusters, walls, filaments [20]). If this knowledge is lacking, the proportion  $P(r)$  of classified points as a function of  $r$  can still be used to highlight percolation phases.

**Comparison for the identification of a specific cluster** We may wish to compare two estimators for the correct identification of a particular cluster. We can then use the following protocol, inspired by the Precision/Recall method: Let  $\mathcal{C}_P$  be a cluster of level  $P$  (a connected component of level  $h_P$  for the true density function  $f$ ) with volume  $|\mathcal{C}_P|$ . Let  $\hat{\mathcal{C}}_{P'}$  be the corresponding empirical cluster and  $|\hat{\mathcal{C}}_{P'}|$  its volume. We can then define *Precision* and *Recall* :

$$\text{Precision}(P, P') = \frac{|\mathcal{C}_P \cap \hat{\mathcal{C}}_{P'}|}{|\hat{\mathcal{C}}_{P'}|}, \quad \text{Recall}(P, P') = \frac{|\mathcal{C}_P \cap \hat{\mathcal{C}}_{P'}|}{|\mathcal{C}_P|}.$$

**Comparison of density level maps** Now suppose that we have a complete density level map of a region  $\mathcal{R} \subset \mathbb{R}^d$  observed:

$$1 - \hat{\mathcal{P}} : x \in \mathbb{R}^d \mapsto 1 - \hat{\mathcal{P}}(x) \in [0; 1]$$

which is an estimator of the ground truth function  $1 - \mathcal{P}$ . We can then take the  $p$  norms (from the  $L^p$  space) to compare our estimators with the original function.

### 3 Our method

Let us describe more accurately our method in this section. In three main steps:

– Starting with a radius  $r$  equal to 0 and increasing it. For each radius  $r$ , we construct  $\mathcal{G}(\mathcal{X}, r)$ . From this graph, we retain only the connected components with more nodes than a certain *percolation threshold*. We then look at the proportion  $P$  of points lying into one of these major connected components.

– The associated estimator of the High-Density Clusters of level  $h_P$  is a set  $\Sigma_P \subset \mathbb{R}^d$  containing the points of the major connected components.

– Each time a large component appears, a new filament is created. (Filaments are modelled by sub-graphs of connected components). Filaments are initialized with the Fréchet mean of the component (for the distance induced on the graph). Then, they grow progressively with the High-Density cluster.

**Persistent ingredients** By analogy with persistent homology (see the survey by Bobrowski & Kahle [7]), we call ‘persistent’ the variational method of observing what happens when the radius  $r$  varies.

### 3.1 Theoretical advantage: The percolation rate

Percolation is a ‘fast’ phenomenon. Let  $\mathcal{H}_1$  be a Poisson point process on  $\mathbb{R}^2$  with fixed intensity  $\lambda = 1$ : We only vary the radius  $r$  of the geometric graph  $\mathcal{G}(\mathcal{H}_1, r)$ .

Starting with  $r = 1$ , percolation has not yet taken place. The largest component therefore contains a proportion  $p_\infty(1) = 0$  of the points. For  $r := r_c = \sqrt{\lambda_c} \approx 1.2$  [27, 34], percolation occurs: A giant component appears<sup>2</sup>.

As soon as the giant component appears (for  $r \geq r_c$ ), if we increase radius  $r$  slightly, the probability of percolation  $p_\infty(r)$  approaches 1 very quickly. At this point, the giant component includes almost all the points (a proportion  $p_\infty(r)$ ).

The plot below<sup>3</sup> shows a simulation of  $p_\infty(r)$  on  $\mathbb{R}^2$ .

How can we measure the speed of percolation? From a certain radius,  $r_{\min}$ , the giant component becomes non-negligible in size compared with the cloud of points. Suppose it contains  $\varepsilon \leftarrow 5\%$  of the points. That is:

$$r_{\min} := p_\infty^{-1}(\varepsilon).$$

For another larger radius  $r_{\max} \geq r_{\min}$ , the giant component encompass almost all the points (a proportion  $1 - \varepsilon$ ). We can consider the percolation to be complete:

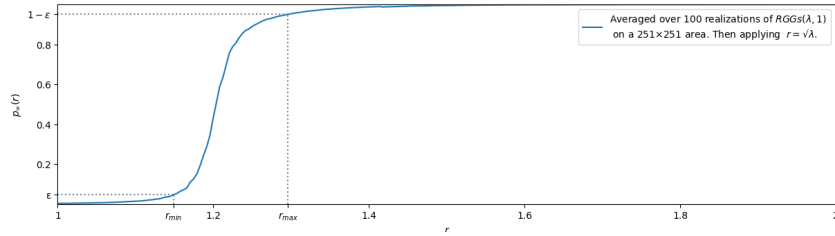
$$r_{\max} := p_\infty^{-1}(1 - \varepsilon).$$

The quantity of interest is

$$\frac{r_{\max}}{r_{\min}}.$$

<sup>2</sup> But still  $p_\infty(r_c) = 0$ . Although it is widely accepted that for any  $d \geq 2$ ,  $p_\infty(r_c) = 0$ , this has only been proved for  $d = 2$  (*cf.* theorem 4.5 by Meester & Roy [23] and by Tanemura for  $d$  sufficiently large [30]).

<sup>3</sup> Thanks to Vinay Kumar [33] for sharing the data.



**Fig. 1.** Estimation of percolation probability  $p_\infty(r)$  in  $\mathbb{R}^2$  by simulation on giant Random Geometric Graphs. © [33]. With  $\varepsilon = 0.05$ , it gives the following results:  $r_{\min} = 1.15$  and  $r_{\max} = 1.30$ . Note that on this curve,  $p_\infty(r)$  is positive even if  $r \lesssim r_c \approx 1.2$ ; this is due to the approximation of  $\mathbb{R}^2$  by a finite square  $251 \times 251$ .

Suppose there are two large contiguous regions, of intensity  $\lambda_1$  and  $\lambda_2$  with  $\lambda_2 < \lambda_1$ . From a certain radius  $r_{\min}^{(1)}$ , percolation begins in the first region with the highest  $\lambda_1$  intensity. To identify this region correctly without confusing it with the neighbouring region of lower intensity  $\lambda_2$ , the first percolation phase must be ‘completed’ before percolation begins in the second region. In other words, we want to have :

$$r_{\max}^{(1)} < r_{\min}^{(2)}.$$

Now, in  $\mathbb{R}^d$ ,  $r_{\min}^{(2)} = \left(\frac{\lambda_2}{\lambda_1}\right)^{\frac{1}{d}} \times r_{\min}^{(1)}$  and  $r_{\max}^{(1)} = \left(\frac{\lambda_2}{\lambda_1}\right)^{\frac{1}{d}} \times r_{\max}^{(1)}$ , the two regions can therefore be correctly and distinctly identified if and only if :

$$\frac{r_{\max}}{r_{\min}} < \left(\frac{\lambda_2}{\lambda_1}\right)^{\frac{1}{d}}.$$

(The quantity  $\frac{r_{\max}}{r_{\min}}$  being independent of the intensity  $\lambda$  of the region).

Percolation will be all the faster as the ratio  $\frac{r_{\max}}{r_{\min}}$  becomes close to 1. On Fig. 1, for  $\varepsilon = 0.05$ , we can see that this ratio is indeed close to one:  $r_{\min} = 1.15$  and  $r_{\max} = 1.30$ . Thus:  $\frac{r_{\max}}{r_{\min}} \approx \frac{1.30}{1.15} \approx 1.13$  in  $\mathbb{R}^2$ .

**The ‘percolation’-graph** Introducing percolation ingredients into the geometric graph is made in a very simple way: We set a robust percolation threshold (e.g.  $PercolThreshold \leftarrow 50$ ), and consider only connected components with more than  $PercolThreshold$  nodes. We denote  $\mathcal{G}_1(\mathcal{X}, r)$  the graph pruned of the small connected components.

### 3.2 Filament extraction from a graph

In this sub-section, we fix the radius  $r$  of the pruned geometric graph  $\mathcal{G}_1(\mathcal{X}, r)$  and look at one of the big connected components, which is a sub-graph  $G(V, E)$  with vertices  $V$  and edges  $E$ . A filament *Filament* may already have been drawn on

this component (for previous radii). We have a distance  $d$  on this graph induced by Euclidean distance between two neighbouring points. If  $x, y \in V$  are two vertices,  $ShortPath(x, y)$  denotes the set of vertices of the shortest path (for the distance  $d$ ) from  $x$  to  $y$  in  $G$ .

The variable  $Centres$  denotes the set of vertices which are chosen to represent *Filament*. The first centre is the Fréchet mean of  $G$ .  $FilNodes$  are the vertices of *Filament* which join the  $Centres$  such that *Filament* is the Minimal Tree spanning  $Centres$  nodes.

---

**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  $|Centres| < \text{int}(|V|/PercolThreshold)$  do    ▷ We search for a new centre
   $D \leftarrow \{\}$                                 ▷ The sums of the distances to minimise
  for  $x \in V$  do                                ▷  $x$  is the hypothetical new centre
     $NodesFilament \leftarrow \text{copy}(FilNodes)$ 
     $Branch_x \leftarrow ShortPath(x, NodesFilament)$     ▷ Hypothetical new branch
     $NodesFilament \leftarrow NodesFilament \cup Branch_x$ 
     $D[x] \leftarrow 0$ 
    for  $y \in V$  do
       $D[x] \leftarrow D[x] + d(y, NodesFilament)^2$ 
    end for
  end for
   $x \leftarrow \text{argmin}(D)$                                 ▷ The new centre chosen
   $Centres \leftarrow Centres \cup \{x\}$ 
   $FilNodes \leftarrow FilNodes \cup Branch_x$ 
end while
Filament ← MinimalSpanningTree(FilNodes)
Returns Filament

```

---

Note that the first centre chosen with the Algo. 1 is the Fréchet mean [15] of the graph  $G(V, E)$ . Moreover, the *Filament* result is a tree.

In some cases, we might want to ‘close’ the tree by inserting a loop and introduce a closing post-processing algorithm.

A final post-processing consists of pruning the filamentary network obtained of branches that are too small (*e.g.* those shorter than the radius  $r$  of the geometric graph).

### 3.3 The density level estimator

Thanks to all the concepts and tools defined above, we are now able to provide an estimator of density levels. Let  $P \in [0; 1]$  be the proportion of ‘classified’



points (lying in one of the great components). The associated radius is

$$r_P = \inf \left\{ r \mid \frac{|\text{Vertices}(\mathcal{G}_1(\mathcal{X}, r))|}{|\text{Vertices}(\mathcal{G}(\mathcal{X}, r))|} \geq P \right\}.$$

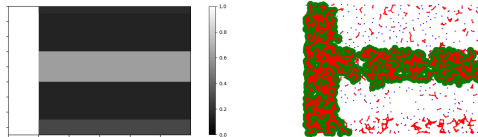
At radius  $r_P$ , the connected components of  $\mathcal{G}(\mathcal{X}, r_P)$  of size  $\geq \text{PercolThreshold}$  represent a proportion  $P$  of the cloud point  $\mathcal{X}$ . We have now to find a volume of the space  $\Sigma_P \subset \mathbb{R}^d$  that fits best the classified points. An intuitive solution, inspired by percolation theory (*Boolean model*), would be to take the union of balls centred on these classified points:

$$\Sigma_P := \bigcup_{x \in \text{Vertices}(\mathcal{G}_1(\mathcal{X}, r_P))} B(x, R).$$

The radius  $R$  needs to be chosen. Usually, in continuum percolation theory, we consider  $R = r_P/2$ . But this radius is too small for our purpose: If we consider the volume on the entire Boolean model at percolation stage:

$$\Sigma := \bigcup_{x \in \text{Vertices}(\mathcal{G}(\mathcal{X}, r_c))} B(x, r_c/2)$$

(also  $\Sigma \supset \Sigma_P$ ), in  $\mathbb{R}^2$ ,  $\Sigma$  occupies only a proportion  $\phi_c \approx 0,676$  [27, 34] of the space. ( $\phi_c$  is called the *space coverage* [17]). That is,  $\Sigma$  will not recover a proportion  $1 - \phi_c$  of the space, equals by ergodicity to  $e^{-\lambda_c \theta_d / 2^d}$  ( $\theta_d$  being the volume of the unit ball in  $\mathbb{R}^d$ ). With a radius twice as large (our choice:  $R = r_P$ ; see on Fig. 2 an example), this un-recovered proportion of space is reduced to:  $e^{-\lambda_c \theta_2} = (1 - \phi_c)^4 \approx 0.01$ . Our experiments show that taking a larger radius  $R$  (up to  $1.5 \times r_P$ ) produces better results. Increasing  $R$  increases the *Recall*. Taking  $R$  too large, however, can end up lowering *Precision*.



**Fig. 2.** Left: Density levels of  $f$ . Function support is a rectangle  $24 \times 17$  subdivided into sub-rectangles. A left high density ‘blob’ ( $f \propto 4$ ); At the center, a thick ‘Filament’ ( $f \propto 3$ ); Below a thinner one ( $f \propto 2$ ) and upper a very thin one ( $f \propto 1$ ). Between ‘Filaments’, some ‘voids’ ( $f \propto \frac{1}{2}$ ). Right: 2000 IID **points** generated by  $f$ . The  $\mathcal{G}(\mathcal{X}, r \leftarrow 0.4)$  **graph edges** and the density level volume associated  $\Sigma_r$ , with  $\text{PercolThreshold} \leftarrow 50$ .

**The density level map** The definition of the empirical density level map  $1 - \hat{\mathcal{P}}$  follows naturally: let  $x \in \mathbb{R}^d$ ,  $\hat{\mathcal{P}}(x)$  is the first  $P$  for which  $x$  lies in  $\Sigma_P$ , *i.e.*

$$\hat{\mathcal{P}}(x) := \inf \{ P \in [0; 1] \mid x \in \Sigma_P \} \quad (\text{with the convention } \inf(\emptyset) = 0).$$

On Fig. 4, the reader can see three examples of empirical density level maps estimated on a cloud of points IID generated with density  $f$  (see Fig. 2).

## 4 Results

In this section, we first see an example of filament extractions on a synthetic 2D-image of galaxies<sup>4</sup> and compare visually with a stochastic method [29]. Second, for a more quantified comparison, we compare density level estimators (ours, Delaunay estimator, 10-Nearest Neighbours) on point cloud  $\mathcal{X}$  generated by a known density function  $f$  plotted on Fig. 2: A rectangular-shaped density map (‘blob’ modelled by a large and high-density rectangle, ‘filament’ by a thin one).

**Visual comparison with stochastic method** Stochastic geometry methods have been proposed for extracting galaxy filaments [16, 29, 31]. A sheet of *Filament* is represented by a rectangular box. Geometric priors are then introduced on its shape, its density, its connectivity (or alignment) with the other boxes, ... In the end, using techniques such as simulated annealing, the configuration that best fits the data is obtained. Figure 3 shows the result of such an algorithm.

We apply Algo. 1 (see below) to  $\mathcal{X}$  with  $r$  varying from 0 to 5.2. As  $r$  grows, components increase in size, merge, and *Filaments* grow with the radius. In Fig. 3, the result (= *Filaments* drawn) for a very small stopping radius and a larger one.

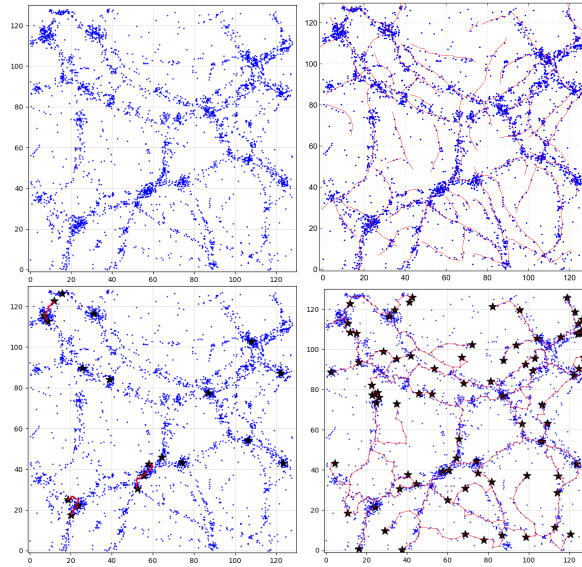
Note that, thanks to persistent ingredients, *Filaments* are robust to the choice of stopping radius: once the majority of points appear in a large component, few new centres are added. So the result is the same except for a few ‘connection-bridges’. As there is no ground truth, comparison is difficult. Our filaments, which are drawn without geometric constraints, are more irregular. However, they form a genuine network and are less prone to over-detection. Our method is also much less computationally intensive. What is more, it can easily be applied to three-dimensional images.

**Comparison with classical density estimators** In order to obtain quantified results, let us now work on clouds of points IID generated according to a density function  $f$  that we know (*cf.* Fig. 2 with a 2000-points cloud  $\mathcal{X}$  scattered).

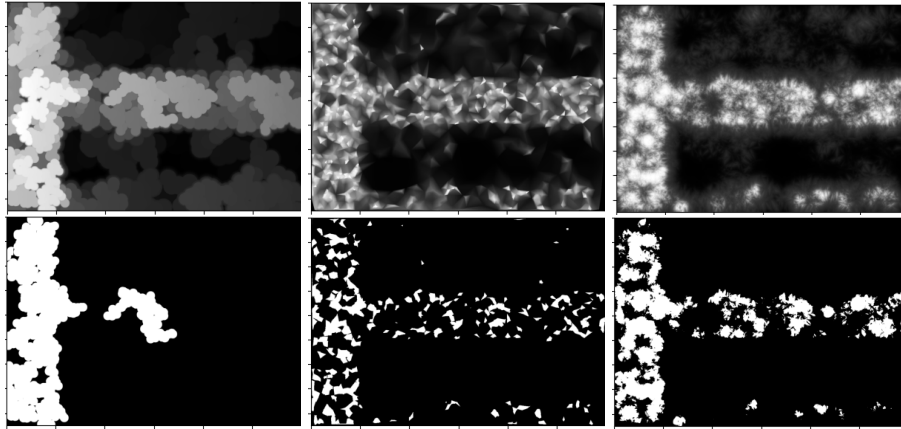
The results of the three density level estimators (ours, Delaunay [28] and  $K$ -Nearest Neighbours [6] with  $K \leftarrow 10$ ) can be seen in Fig. 4. Visually, ours is more homogenous and less prone to local overestimates in low-density zone.

We are now numerically able to compare the estimated density level maps (Fig. 4) with the original one (Fig. 2; density plateaus between two density levels have been replaced by the half-sum). Tab. 1 shows an advantage for our estimator.

<sup>4</sup> Thanks to the authors of the article “Detection of cosmic filaments using the Candy model” [29] for the generation of the data used herein. These data were kindly supplied by Radu Stoica, Enn Saar and Vicent Martínez.



**Fig. 3.** Top Left: Mock cloud *point*  $\mathcal{X} \subset \mathbb{R}^2$  generated by Stoica *et al.* [29]. Top Right: Results of the stochastic ‘Candy Model’ algorithm [29]. Bottom: The persistent extracted Filaments on  $\mathcal{X}$  with *PercolThreshold*  $\leftarrow$  50. Left, for  $r$  varying from 0 to 1. First *Centres* (blacks stars  $\star$ ) and *Filaments* appear. Right,  $0 \leq r \leq 5.2$ .



**Fig. 4.** Top: Three density level maps estimated on the cloud point of Fig. 2. From left to right: Our Percol-Graph estimator, the Delaunay estimator [28], the  $K$ -Nearest Neighbours estimator [6] with  $K \leftarrow 10$ . Bottom: The cluster (white) associated to the density level  $1 - \hat{\mathcal{P}} > 0.617$ . Theoretically, this cluster is the rectangle on the left.

Let us take a closer look. Four levels are interesting, corresponding to the successive appearance of the ‘filaments’:  $1 - \hat{\mathcal{P}} = 0.617$  (highest-density

**Table 1.** Distance between the estimated density level map and the original one

Algorithm	$L^1 := \sum_x \frac{1}{N}  \hat{\mathcal{P}}(x) - \mathcal{P}(x) $	$L^2 := \sqrt{\sum_x \frac{1}{N} (\hat{\mathcal{P}}(x) - \mathcal{P}(x))^2}$
Graph-Percol	<b>0.095</b>	<b>0.142</b>
Delaunay	0.123	0.187
$K$ -Nearest Neighbours	0.114	0.166

left rectangle).  $2^\circ 1 - \mathcal{P} = 0.280$  (middle thick filament).  $3^\circ 1 - \mathcal{P} = 0.169$  (bottom filament).  $4^\circ 1 - \mathcal{P} = 0.140$ ; (Only “voids” are not in this level). We compute *Precision* and *Recall* for these levels. Results are listed in Tab. 2.

**Table 2.** *Precision* and *Recall* on density levels of filament apparitions.

Algorithme	$1 - \tilde{\mathcal{P}} > 0.617$		$1 - \tilde{\mathcal{P}} > 0.280$		$1 - \tilde{\mathcal{P}} > 0.169$		$1 - \tilde{\mathcal{P}} > 0.140$	
	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>	<i>Precision</i>	<i>Recall</i>
Graph-Percol	<b>0.741</b>	<b>0.827</b>	0.881	<b>0.970</b>	0.814	<b>0.945</b>	0.783	<b>0.896</b>
Delaunay	0.573	0.315	0.857	0.799	<b>0.892</b>	0.899	<b>0.881</b>	0.886
$K$ -NN	0.600	0.528	<b>0.899</b>	0.864	0.861	<b>0.945</b>	0.802	0.893

Our estimator outperforms the other ones for the highest level of density, *i.e.* for the correct detection of the high-density left rectangle. Percolation is in fact a fast enough phenomenon to occur in this zone at density  $f \propto 4$  before taking place in the medium filament with close density  $f \propto 3$ . See Fig. 4: Almost the entire cluster is detected and only one (small) connected component of the thick filament appears. The other estimators have a density level much more uniformly distributed over the main clusters of close densities.

## 5 Conclusion and perspectives

In this paper we propose a new estimator of density levels based on geometric graphs. Looking at what happens persistently allows us to observe percolation phases. Since continuum percolation is a very fast phenomenon, our estimator is able to identify two neighbouring levels of close density.

This estimator of density levels could find a natural application to the problem of identifying galaxy clusters, which are highly hierarchical. In addition, the availability of a graph makes it fairly easy to extract galaxy filaments without having to resort to methods such as calculating the median axis.

Compared with conventional density estimators for this type of problem, it is already showing very good results, especially for high-density clusters.

In the future, we will try to further improve these results focusing on four principal research directions:  $1^\circ$  We mainly looked at one type of clusters, ‘filaments’. Having an estimator of density levels allows us to look at other types,

such as ‘super-clusters’, ‘walls’ and ‘voids’. 2° A galaxy was represented only by a point. Its mass (= its luminosity) could be taken into account using different radii, depending on galaxies. 3° The question of  $\Sigma_P$  for density levels was briefly considered in this paper. There are certainly wiser choices to be made (*e.g.* inspired by Penrose’s works [26]) to approach strong-consistency. 4° We worked with graphs. We could look at other notions of connectivity (*e.g.* connectivity of simplicial complexes).

## References

1. 2df galaxy redshift survey. 2dF Galaxy Redshift Survey URL [2dFGalaxyRedshiftSurvey](http://www.2dF Galaxy Redshift Survey)
2. Sloan digital sky survey. <http://www.sdss.org> URL <http://www.sdss.org>
3. In: M. Longair, J. Einasto (eds.) *The Large Scale Structure of the Universe, International Astronomical Union Symposia*, vol. 79, p. 464. Springer, Tallinn (1978)
4. Attali, D., Boissonnat, J.D., Edelsbrunner, H.: *Stability and Computation of Medial Axes - a State-of-the-Art Report*, pp. 109–125. Springer (2009). DOI 10.1007/b106657\_6
5. Barrow, J.D., Bhavsar, S.P., Sonoda, D.H.: Minimal spanning trees, filaments and galaxy clustering. *MNRAS* **216**(1), 17–35 (1985). DOI 10.1093/mnras/216.1.17
6. Biau, G., Devroye, L.: *Lectures on the Nearest Neighbor Method*, vol. 246. Springer (2015). DOI 10.1007/978-3-319-25388-6
7. Bobrowski, O., Kahle, M.: Topology of rand. geom. complexes: a survey. *Journal of appl. and Comput. Top.* **1**, 331–364 (2018). DOI 10.1007/s41468-017-0010-0
8. Boissonnat, J.D., Wintraecken, M.: The reach of subsets of manifolds. *Journal of Applied and Computational Topology* pp. 1–23 (2023). DOI 10.1007/s41468-023-00116-x
9. Bollobás, B., Riordan, O.: *Percolation*. Cambridge University Press (2006). DOI 10.1017/CBO9781139167383
10. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: *Advances in Knowledge Discovery and Data Mining*, pp. 160–172. Springer, Berlin, Heidelberg (2013). DOI 10.1007/978-3-642-37456-2\_14
11. Colberg, J.M.: Quantifying cosmic superstructures. *MNRAS* **375**(1), 337–347 (2007). DOI 10.1111/j.1365-2966.2006.11312.x
12. Darvish, B., Mobasher, B., Sobral, D., Scoville, N., Aragon-Calvo, M.: A comparative study of density field estimation for galaxies. *The Astrophysical Journal* **805**(2), 121 (2015). DOI 10.1088/0004-637X/805/2/121
13. Einasto, J.: Large scale structure of the Universe. *AIP Conference Proceedings* **1205**(1), 72–81 (2010). DOI 10.1063/1.3382336
14. Ferdosi, B. J., Buddelmeijer, H., Trager, S. C., Wilkinson, M. H. F., Roerdink, J. B. T. M.: Comparison of density estimation methods for astronomical datasets. *Astronomy & Astrophysics* **531**, A114 (2011). DOI 10.1051/0004-6361/201116878
15. Fréchet, M.: L’intégrale abstraite d’une fonction abstraite d’une variable abstraite et son application à la moyenne d’un élément aléatoire de nature quelconque. *La Revue Scientifique* (1944)
16. Gernez, P., Descombes, X., Zerubia, J., Slezak, E., Bijaoui, A.: Galaxy filament detection using the quality candy model. In: *IEEE Intern. Conf. on Ac. Speech and Sign. Proc.*, vol. 2 (2006). DOI 10.1109/ICASSP.2006.1660447

17. Hall, P.: Introduction to the theory of coverage processes. Probability and mathematical statistics. John Wiley & Sons (1988)
18. Hartigan, J.A.: Clustering Algorithms. John Wiley & Sons, Inc. (1975)
19. Hartigan, J.A.: Consistency of single linkage for high-density clusters. *J. of the Am. Stat. Ass.* **76**(374), 388–394 (1981). DOI 10.1080/01621459.1981.10477658
20. Kuchner et al.: An inventory of galaxies in cosmic filaments feeding galaxy clusters. *MNRAS* **510**(1), 581–592 (2021). DOI 10.1093/mnras/stab3419
21. Kullback, S., Leibler, R.A.: On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**(1), 79 – 86 (1951). DOI 10.1214/aoms/1177729694
22. Libeskind et al.: Tracing the cosmic web. *MNRAS* **473**(1), 1195–1217 (2017). DOI 10.1093/mnras/stx1976
23. Meester, R., Roy, R.: Continuum Percolation. Cambridge Tracts in Mathematics. Cambridge University Press (1996). DOI 10.1017/CBO9780511895357
24. Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications* **48**, 257–263 (1982). DOI 10.1016/0024-3795(82)90112-4
25. Penrose, M.: Random Geometric Graphs, vol. 5. Oxford University Press (2003). DOI 10.1093/acprof:oso/9780198506263.001.0001
26. Penrose, M.: Random euclidean coverage from within. *Probability Theory and Related Fields* **185**(3-4), 747–814 (2023). DOI 10.1007/s00440-022-01182-5
27. Quintanilla, J., Torquato, S., Ziff, R.: Efficient measurement of the percolation threshold for fully penetrable discs. *Journal of Physics A* **33**(42), L399–L407 (2000). DOI 10.1088/0305-4470/33/42/104
28. Schaap, W.E.: Dtfe : the delaunay tessellation field estimator. Ph.D. thesis, Proefschrift Rijksuniversiteit Groningen (2007)
29. Stoica, R., Martínez, V., Mateu, J., Saar, E.: Detection of cosmic filaments using the candy model. *Astronomy & Astrophysics* **434**(2), 423–432 (2005). DOI 10.1051/0004-6361:20042409
30. Tanemura, H.: Critical behavior for a continuum percolation model. *Probability Theory and Mathematical Statistics* pp. 485–495 (1996)
31. Tempel, E., Stoica, R., Kipper, R., Saar, E.: Bisous model. detect. filam. patterns in p.p. A & C **16**, 17–25 (2016). DOI 10.1016/j.ascom.2016.03.004
32. Vaserstein, L.N.: Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredači Informacii* **5**(3), 64–72 (1969)
33. Vinay Kumar, B., Kashyap, N., Yogeshwaran, D.: An analysis of probabilistic forwarding of coded packets on random geometric graphs. *Performance Evaluation* **160**, 102343 (2023). DOI 10.1016/j.peva.2023.102343
34. Xu, W., Wang, J., Hu, H., Deng, Y.: Critical polyn. in the nonplanar and cont. percol. models. *Phys. Rev.* **103**, 022127 (2021). DOI 10.1103/PhysRevE.103.022127