École Polytechnique Fédérale de Lausanne

Laboratory of Astrophysics (LASTRO)

Tracing the Cosmic Web

A Dual Study on Galaxy Spectra and Large-Scale Percolation Analysis

---

Tommaso Marzan
Bachelor student in Physics
`tommaso.marzan@epfl.ch`

Supervisor
Aurélien Verdier
Laboratory of Astrophysics - EPFL

Laidlaw Scholars Leadership and Research Program

Summer Research Internship

July - September 2026

Academic Year 2025/2026

Abstract

Modern observational cosmology relies heavily on automated pipelines to process massive spectroscopic surveys, making data validation and analysis crucial for further research. This study first evaluates the performance of the 4MOST automated pipeline through the Visual Inspection (VI) of 1,015 spectra, assessing classification and redshift (z) accuracy across galaxies, stars, quasars and unknown objects. The pipeline achieved an overall accuracy of 86.70% for redshift estimation and 83.45% for object classification. While stars showed 100% agreement and high-confidence quality flags reliably implied accurate redshifts, the pipeline exhibited a systematic bias by over-classifying low-quality galaxy spectra as quasars and assigning low-confidence flags rather than marking objects as unknown. These results will be later on exploited by the Laboratory of Astrophysics of EPFL for further research in observational cosmology.

We then analyse the statistical properties of large-scale galaxy clustering using 53,592 Luminous Red Galaxies (LRGs) from the Dark Energy Spectroscopic Instrument (DESI) Data Release 1. Their 3D spatial distribution was reconstructed under the standard Planck 2018 model and compared against a reference catalog with a perfectly uniform spatial distribution over the same volume.By applying 3D site percolation theory, we evaluated how galaxies link together as the connection distance increases. The analysis reveals clear phase transitions, marked by the sudden emergence of a single "super-cluster" spanning the entire sample. Crucially, this percolation phase transition occurs at a smaller distance threshold in the real DESI data than in the control sample. This earlier transition is directly consistent with theoretical expectations, as gravitational attraction pulls galaxies into dense structures, making it significantly easier for large-scale connected networks to form compared to a uniform distribution. This interdisciplinary approach demonstrates how applying statistical physics to the study of galaxies provides a comprehensive framework to map the cosmic web and quantify galaxy clustering across current and upcoming spectroscopic observations.

Tracing the Cosmic Web
A Percolation Analysis of Large-Scale Structure

Compiled with LaTeX. Source code available upon request.

Contents

1 Introduction

The current paper is the result of an 8-weeks summer internship carried out at the Laboratory of Astrophysics of EPFL in the scope of the Laidlaw Scholars Leadership and Research Program. My work during these 8 weeks was mainly divided into two different sub-projects. The first one was directly proposed to me by my supervisor and deals with the visual inspection (VI) of around a thousand of spectra coming from the 4MOST Survey Validation (SV) database (see section 1.2). This work will later on be essential to the laboratory in order to produce some more advanced results. The second part of my work is a personal project that investigates the large-scale structure of the universe using the theory of percolation and phase transitions (from which the name of this report), starting from the DESI Data Release 1 database.

1.1 Classification of astronomical bodies

Astronomy and astrophysics deal with a vast variety of celestial bodies in order to study their properties. While stars are luminous spheroids of hot gas held together by (self-)gravity (like the sun), galaxies are formed by group of stars and other objects gravitationally linked to each other. In the scope of this research, the most relevant celestial bodies can be resumed as follows.

- Luminous Red Galaxies (**LRG**): particularly luminous old or dead galaxies that therefore present a big fraction of red stars.
- Bright Galaxies (**BG**): all kind of galaxies that are intrinsically very bright.
- Emission Light Galaxies (**ELG**): galaxies that present an intense star formation activity.
- Quasi-Stellar Objects (**QSO**): also called quasars, they're galaxies that contain a supermassif black hole at their centre.

The main way to cathegorise them is to look at their spectrum, as the composition and the age of a galaxy impacts the way the light is emitted and absorbed. Some basics of spectroscopy are covered in section 1.3, whereas the instruments used for this scope are presented in the next sub-section.

1.2 Spectroscopic instruments: 4MOST and DESI

The datasets analysed in this research were acquired using two optical spectrographs: instruments designed to measure light properties across specific bands of the electromagnetic spectrum. A detailed description of both instruments is provided below.

- The 4-Meter Multi-Object Spectroscopic Telescope (**4MOST**) is a wide-field optical telescope based at ESO[1]'s Observatory in Cerro Paranal, Chile. Its design allows to obtain millions of spectra during five-year surveys, even for objects spread over a significant area of the sky. Its main purpose is to support research about galactic stellar archaeology and extragalactic surveys [1]. The telescope achieved first light in October 2025.
- The Dark Energy Spectroscopic Instrument (**DESI**) is also an optical telescope, based at the Kitt Peak National Observatory in Arizona, USA. Thanks to DESI's unprecedented precision and capacity, the first public data release contained spectra of about 47 millions of galaxies [2]. This will enable to do research in many branches of cosmology and fundamental physics. Its main aim can be described as being to measure the history of the expansion of the universe, which is related to dark energy (from which the name DESI) [3].

Figure 1 shows a subset of the observable universe mapped by DESI Data Release 1. The plot displays the spatial distribution of targets out to redshift $z \simeq 4$, with the earth positioned at the vertex of the slice. Different tracer populations - ranging from BGs and LRGs at lower redshifts to ELGs and QSOs farther out - are color-coded to highlight their respective radial depth. The inset magnification emphasises the large-scale structure of the universe (further detailed in section 3).

[1] European Southern Observatory, international astronomical organisation founded in 1962.

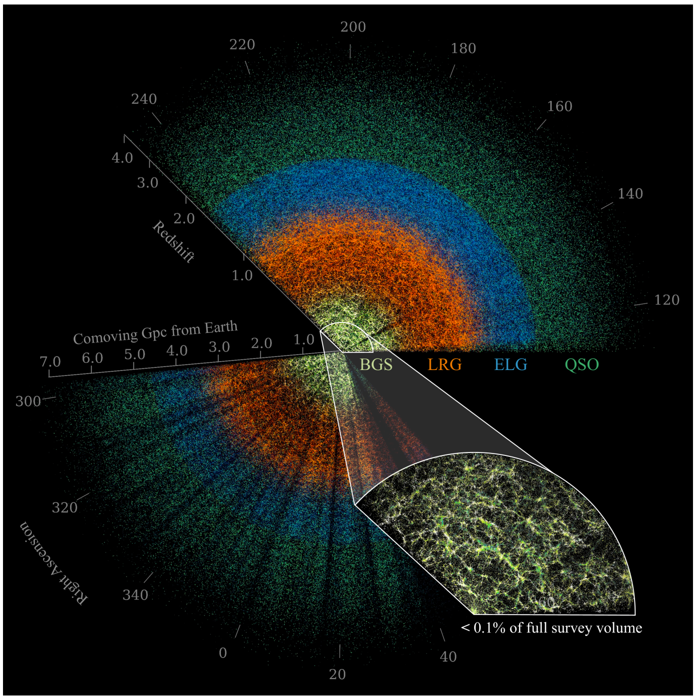


Figure 1: A slice of the universe mapped by DESI Data Release 1. Redshifts from 0 to 4 and different galaxy types (BGs, LRGs, ELGs and QSOs) are rendered, with the earth at centre. Image credits: [2].

1.3 Redshift and distance measurements

Spectroscopy allows us to observe how the irradiance of light behaves as a function of the wavelength emitted or absorbed by an object. In astrophysics, the targets of our observations are light-emitting bodies like galaxies, stars, or black holes. Typically, spectroscopic analysis leads to the identification of chemical elements by detecting emission or absorption peaks at specific wavelengths. As a matter of fact, molecular transitions produce unique spectral lines at well-defined rest-frame wavelengths.

However, as light travels through space to reach a telescope, these characteristic lines appear systematically shifted towards longer wavelengths. This effect is called relativistic Doppler effect and can be seen as a generalisation of the classical Doppler effect: it's the same phenomenon that explains why - for example - the siren of an ambulance moving towards us is heard at a higher frequency than when at rest. Denoting λ_S the wavelength of a light pulse at rest (e.g. in the reference frame of the object we're observing) and λ the wavelength perceived by an observer (e.g. the telescope), the **redshift** is defined as follows:

$$z = \frac{\lambda - \lambda_S}{\lambda_S}. \tag{1}$$

In observational cosmology, redshift plays a fundamental role, as it is the direct physical quantity measured from galaxy spectra, revealing the shift of known spectral features. It is important to remark that redshift only depends on the relative velocity *along the line of sight* between the source and the observer. In astrophysics, the velocity of an object is always given by the sum of two different contributions. The first contribution is the "intrinsic" velocity of the object, that is mainly due to gravitational forces. The second contribution is due to the expansion of the universe: by calling d the distance between the observer and the source, this contribution follows **Hubble's law**:

$$v = H_0 \cdot d \tag{2}$$

where H_0 is the Hubble constant. This contribution is a direct consequence of the theory of general relativity and shall be interpreted as a change in the geometry of space itself. Mathematically, we can say that a time-dependent metric applies. A nice analogy to visualise this phenomenon is an inflating balloon, where the points drawn on its surface recede from one another not because they are moving across the material, but because the fabric itself is stretching. The actual value of H_0 is currently unknown as observations lead to two different possible values, and stating which is the correct one remains one of the major open questions in modern physics.

2 Spectra Analysis

This section focuses on spectral data retrieved from the 4MOST Survey Validation database.

2.1 Motivation

The primary task in this work consists of visual inspection (VI), which involves comparing the automatic classification generated by the 4MOST pipeline with human expert evaluation. In this context, a pipeline refers to the automated software workflow that processes the raw data collected by the telescope. It performs reduction, calibration, and preliminary analysis - such as extracting spectra, estimating redshifts, and classifying objects - without human intervention. Specifically, the pipeline automatically estimates the redshift of a given spectrum and classifies the type of astronomical object observed.

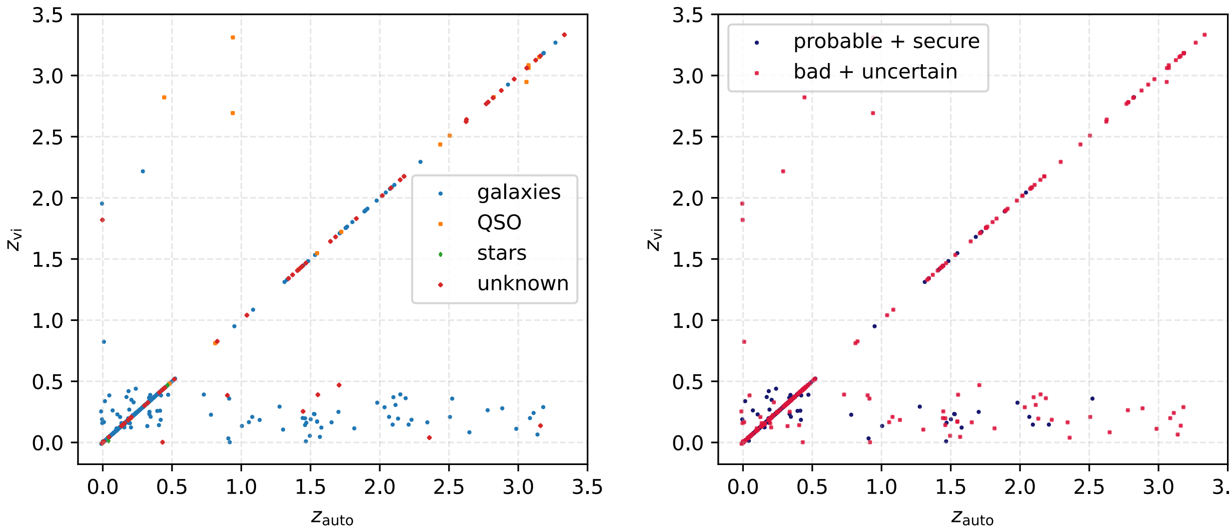
The purpose of this visual inspection is threefold and can be summarised as follows:

- to assess whether the automatic classification and redshift estimates from the pipeline are consistent or biased, and to identify potential systematic errors;
- to verify the overall spectral quality and evaluate the performance of the 4MOST instrument;
- to construct a reliable dataset suitable for training and evaluating machine learning algorithms.

The spectra can belong to four distinct categories of objects: galaxies (`GALAXY`), stars (`STAR`), quasars (`QSO`), or unknown objects (`UNKNOWN`). In order to determine the reliability of the redshift deduced from a spectrum, four confidence labels are used: `1(bad)`, `2(uncertain)`, `3(probable)`, `4(secure)`.

2.2 Results

Overall, 1015 spectra have been visually inspected, and the primary results from this data analysis are presented below. Figure 2a illustrates the distribution of the automatically assigned redshifts ($z_{\rm auto}$) against those assigned via visual inspection ($z_{\rm vi}$) for each object category. Overall, the data points lie predominantly along the diagonal, demonstrating excellent agreement between $z_{\rm vi}$ and $z_{\rm auto}$ across all object types. However, a non-negligible subset of points - mainly representing galaxies -



(a) Redshifts categorised by object type.

(b) Redshifts categorised by quality flags.

Figure 2: Comparison of automatic and visually-inspected redshifts.

scatters significantly away from the diagonal for $z_{\mathrm{vi}} \leq 0.5$, spreading uniformly across $z_{\mathrm{auto}} \in [0, 3.5]$. This indicates that the pipeline misestimates the redshift mainly for a specific fraction of galaxies.

To determine whether these mismatched redshifts correspond to spectra of low or high quality, it is instructive to examine Figure 2b, which presents the same distribution categorised by quality labels. It can be observed that the vast majority of mismatched redshifts possess low-confidence labels (`bad` or `uncertain`), particularly at higher redshifts. Consequently, high-confidence quality labels (`probable` or `secure`) almost strictly imply an accurate pipeline redshift. Conversely, a low-quality label does not necessarily imply a mismatch, as many low-quality spectra still show strong agreement between automated and visual estimates. The overall accuracy is 86.70%.

Figure 3 shows the comparison of the distribution of object types between the visual inspection (VI) and the pipeline (auto). Different behaviours emerge depending on the type of object considered. Accordance is perfect for stars: 100% of object automatically labelled as `STAR` have been confirmed

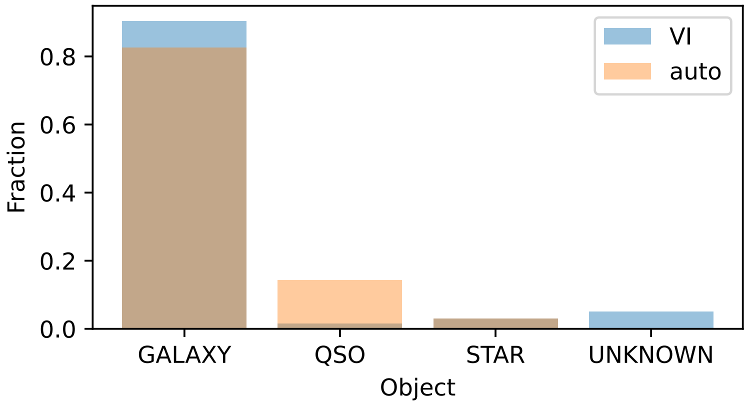


Figure 3: Distribution of object types: automatic and visually-inspected.

by the visual inspection. Regarding other objects, a clear tendency emerges: only 11% of objects that are classified as `QSO` by the pipeline are confirmed by the visual inspection. The remaining 89% is labelled either as `GALAXY` either as `UNKNOWN` after visual inspection. A further analysis also reveals that around 50% of visually-inspected quasars show a mismatch between z_{vi} and z_{auto}. Furthermore, 0% of objects are classified as `UNKNOWN` by the pipeline.

Two conclusions can be thus made following these results. On one hand, the pipeline seems to have a bias towards classifying galaxies or unknown objects as quasars. Second, the pipeline prefers placing a low-confidence label rather than classifying an object as unknown. This being said, the results remain encouraging, as the pipeline achieved an overall 83.45% accuracy for object classification, which is comparable to the redshift accuracy. Table 1 resumes the accuracies of the pipeline.

Classification	**Total mismatching**	**Pipeline accuracy [%]**
Redshift	168	83.45
Object type	135	86.70

Table 1: Redshift and object classification accuracies of the pipeline.

3 Cosmic Web

On large scales, the distribution of matter in the Universe exhibits a non-uniform organization, resulting in the formation of galaxy aggregations across various scales. These galaxies tend to align into a cosmic network of filaments and dense nodes. This pattern primarily originates from initial density fluctuations following the Big Bang, which led overdense regions to gravitationally attract surrounding matter. These structures formed hierarchically, meaning that smaller systems merged over time to give rise to progressively larger cosmic structures. The inset magnification of Figure 1 shows the filamentary structure of galaxies at low redshifts. See also this page.

The aim of this study is to statistically characterise galaxy clustering. The theory of phase transitions will be applied through 3D site percolation to a subset of DESI Data Release 1 galaxies and will be compared to a uniform random distribution.

3.1 Dataset and target selection

The DESI Data Release 1 dataset contains an extremely large number of galaxies, of different types and at different redshifts. In order to make the current study computationally feasible while remaining representative, a restricted subset of these galaxies was chosen. Figure 4 shows the portion of the sky occupied by Luminous Red Galaxies (LRGs) from the database as a function of their redshift $z \in [0.6, 0.8]$. Within this subset, a limited sky footprint containing 53 592 elements is selected, corresponding to the area bounded by the solid black line in Figure 4.

This portion of the sky is defined by the Right Ascension (R.A.) and Declination (Dec.) as follows:

$$225^\circ \leq \text{R.A.} \leq 270^\circ \quad \text{and} \quad 30^\circ \leq \text{Dec.} \leq 45^\circ. \tag{3}$$

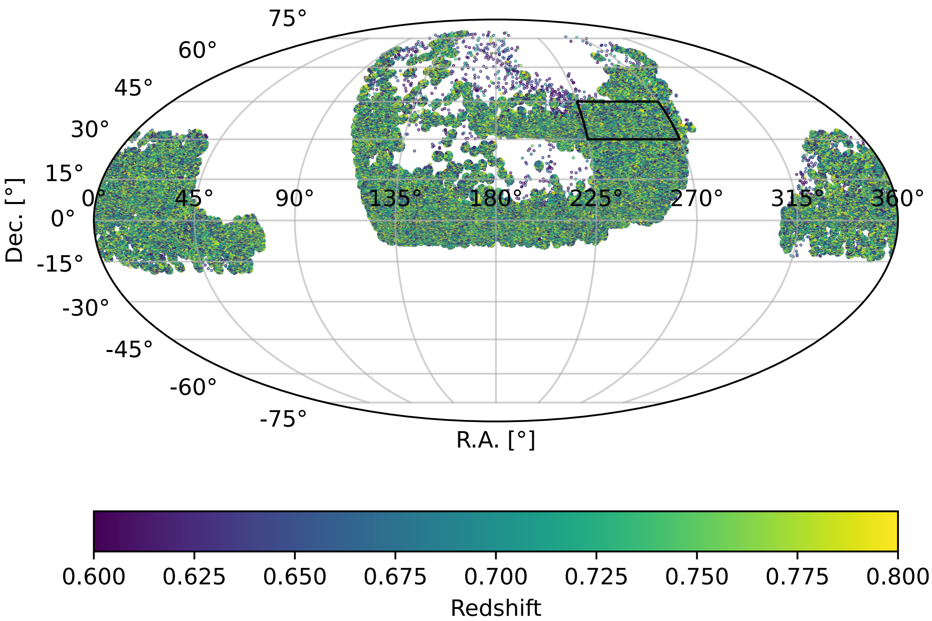


Figure 4: Full-sky distribution of LRG targets, with the survey footprint outlined by the solid black line.

The Right Ascension (*Ascensio Recta*) and the Declination (*Declinatio*) correspond, respectively, to the longitude and the latitude, and can therefore be interpreted as angles of a spherical coordinates system. In astronomy, it is common to call the *Ascensio Recta* α and the *Declinatio* δ.

To analyse the 3D distribution of galaxies, we convert these coordinates to Cartesian space. Each galaxy in our sample is treated as a point mass (i.e. an object whose spatial extension is negligible with respect to the cosmological scale considered). The Cartesian coordinates (X, Y, Z) are obtained through the transformation:

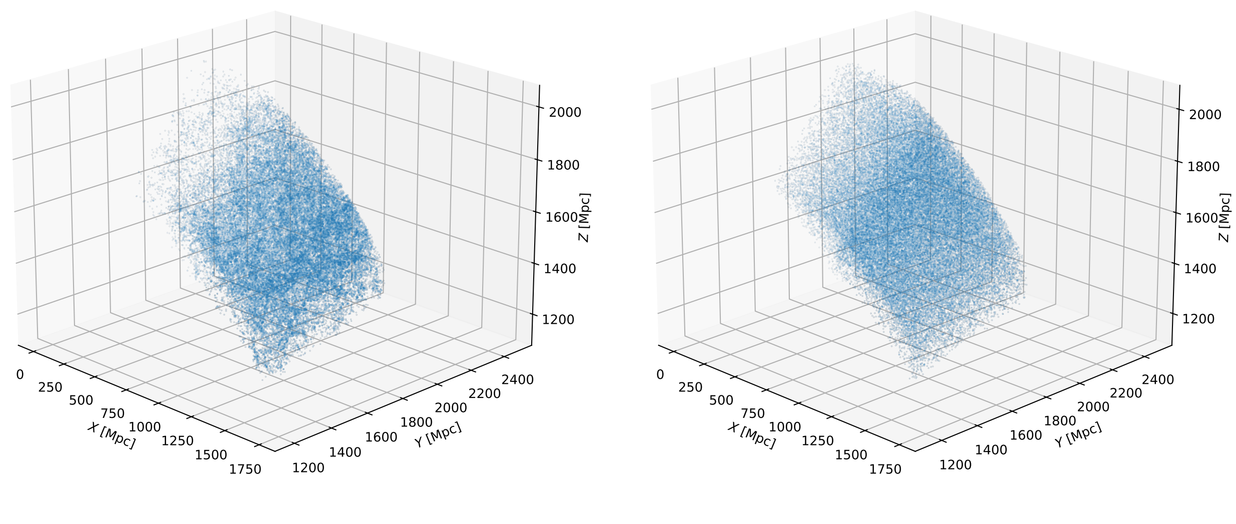
$$\begin{cases} X = d(z) \cdot \cos\delta \cdot \cos\alpha \\ Y = d(z) \cdot \cos\delta \cdot \sin\alpha \\ Z = d(z) \cdot \sin\delta \end{cases} \qquad \text{with} \quad d(z) = \frac{c}{H_0} \int_0^z \frac{dz'}{E(z')} \tag{4}$$

where c is the speed of light, H_0 the Hubble constant and $E(z)$ the dimensionless expansion rate of the universe. As said before, the exact value of the Hubble constant is nowadays still unknown. We adopt the Planck18 cosmological model, which implies $H_0 \simeq 67.66\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ [4]. Applying this transformation yields the 3D cartesian distribution shown in Figure 5a. As a control set, a uniform spatial distribution is generated within the same volume and with the same sample size (53 592 objects). This sample is obtained using `numpy`'s module `random.uniform`. Since a volume element in spherical coordinates is given by

$$dV = dX\,dY\,dZ = d^2 \cos(\delta)\,dd\,d\delta\,d\alpha, \tag{5}$$

having a uniform volume density $V \sim \mathrm{Unif}$ requires drawing uniformly in the transformed variables. Thus the probability density function (PDF) dV must correspond to the PDF of a uniform variable, meaning that we must sample $d^3 \sim \mathrm{Unif}$, $\sin\delta \sim \mathrm{Unif}$, $\alpha \sim \mathrm{Unif}$.

The resulting uniform cartesian distribution is shown in Figure 5b and is denoted "Random Poisson".



(a) DESI Catalog.

(b) Random Poisson.

Figure 5: Spatial (cartesian) distribution of the Luminous Red Galaxies samples.

3.2 Site percolation in 3D

The current section investigates how to formally define the notion of *cluster*. In other words, it is necessary to construct a mathematical definition that identifies a subset of galaxies within a larger set, exclusively based on geometric considerations between the elements of the set. In order to do so, we adopt the following formalism: each galaxy is considered to be a point identified by an index $i \in \{1, \dots, n\}$. At a fixed length L, we say that two points i, j are neighbours if their Euclidean distance is at most L, and in this case we denote $i \sim_L j$. We say that the points i and k are *connected* (or, equivalently, that there exists a *path* between i and k) if there exists a sequence of neighbouring points connecting them, i.e. $i \sim_L j \sim_L \cdots \sim_L k$.

We propose two possible definitions of cluster:

1. A *connected component* (CC), that is, a subset of all nodes such that there exists at least one path connecting all points with a neighbouring distance that is at most L;
2. A subset of points such that all possible couples of points have a distance of at most L.

We now claim that the first definition is the one that better captures the large and filament-like structure of galaxy clusters. As an example, we consider the graph shown in Figure 6. Let's suppose that all the couples of edges connected by a line have a distance of L. In this case, it is evident that following the first definition, all the points are part of a unique connected component, that is, a

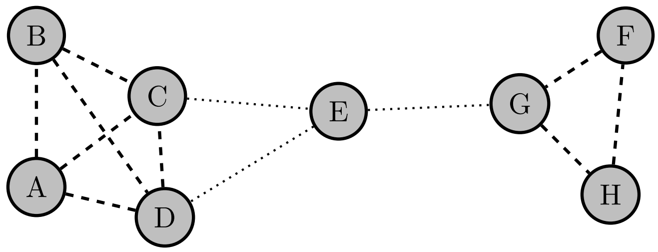


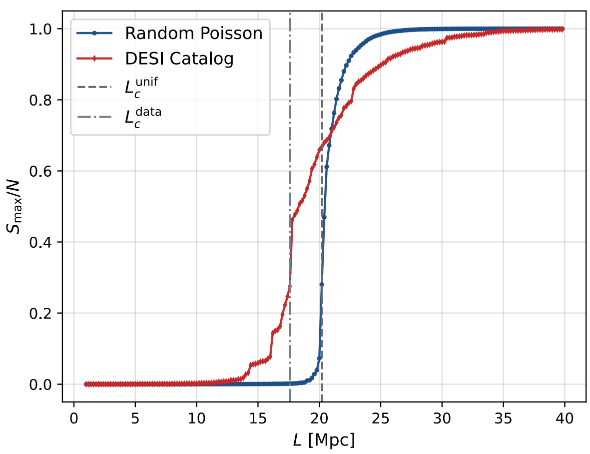
Figure 6: An example of graph representing galaxies.

cluster. As a matter of fact, a path is defined by all possible combinations of dashed lines, combined with the dotted connections E $\sim_L$ G and one between E $\sim_L$ C or E $\sim_L$ D. On the other hand, the second definition would possibly identify three separate clusters: the left and right one, defined by the dashed lines, and the one-element middle one. We thus conclude that the second definition is more adapted to identify local overdensities than large-scale, filament-like patterns. This is why the first definition - borrowed from **percolation theory** - is chosen in the scope of this study.

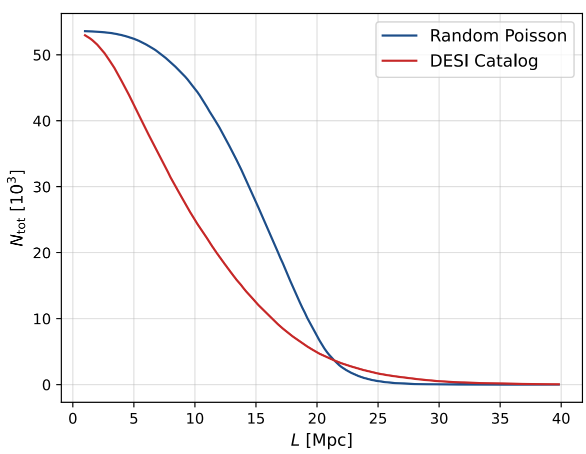
The connection length L is varied between 1 Mpc and 40 Mpc (1 Mpc $\simeq 3.086 \times 10^{22}$ m). We define the size of a cluster as the number of constituent galaxies in the cluster. For each connection length L, both the maximum cluster size S_{max} and the total number of clusters N_{tot} are computed. Let N be the total number of galaxies, such that $N = 53\,592$. Figure 7 shows a typical phase transition curve. In particular, we notice that Figure 7a - which displays the normalised maximum cluster size, representing the fraction of galaxies within the largest cluster - remains flat near zero before rapidly transitioning towards unity in the neighbourhood of a **critical length**, denoted as $L_c^{\text{data}} = 17.60\,\text{Mpc}$ for the DESI catalog and $L_c^{\text{unif}} = 20.40\,\text{Mpc}$ for the uniform distribution.

This phenomenon is analogous to a thermodynamic phase transition, such as boiling water, where a subtle change in temperature causes a shift from liquid to gas. Similarly, a minor increase in the chosen connection length L triggers a global change in the network. Remarkably, the phase transition occurs at a smaller connection length for the DESI dataset than for the uniform sample, confirming that real galaxies are significantly more *clustered* than a random distribution - a result fully consistent with theoretical expectations.

From the physical point of view, the phase transition marks the sudden emergence of a single "giant" cluster (from now on *percolating cluster*) which can be interpreted as the skeleton of the cosmic web. Notably, this phase transition takes place when approximately 50% of all galaxies are incorporated into this giant component. Figure 7b further confirms the dominance of this structure: as the giant cluster forms, the total number of individual clusters drops from an order comparable to the total galaxy count down to nearly one.



(a) Normalised size of the biggest cluster.



(b) Total number of clusters.

Figure 7: Phase transitions of the uniform and the DESI catalogs as a function of the connection length.

A reference value for the average galaxy connection length has not been found. Nevertheless, a critical linking length of $L_c \simeq 17$ Mpc remains physically compatible with the expected connectivity scales of galaxy clusters, given that our dataset is deliberately sparse - comprising exclusively Luminous Red Galaxies (LRGs) restricted to the redshift range $z \in [0.6, 0.8]$. We can therefore state with reasonable confidence that the clear discrepancy in clustering between the uniform catalog and the DESI data is driven by the gravitational attraction of cosmic overdensities. Applying this framework to survey targets with higher numerical densities will naturally scale the critical threshold down to smaller spatial values, thereby mapping cosmic filaments with finer geometric resolution.

3.3 Critical susceptibility

To understand how a complex system responds near a phase transition, we use a quantity called susceptibility (χ). Intuitively, imagine picking a single galaxy at random from the Universe and asking: how large is the cluster it belongs to?

This is fundamentally different from picking a cluster at random. Larger clusters contain vastly more galaxies, meaning a randomly chosen galaxy has a much higher chance of sitting inside a huge group rather than an isolated pair. Susceptibility measures precisely this average cluster size experienced from the perspective of an individual galaxy. To prevent the single largest "super-cluster" from dominating this measure, we exclude it from the computation of both χ and $\langle S \rangle$, isolating the true response of the remaining system.

The detailed mathematical derivation of susceptibility is given in the following paragraph.

We consider a set of $N = \sum s \cdot n(s)$ galaxies and its relative graph at a given connection length L, where s is the cluster dimension and $n(s)$ the number of clusters of dimension s. The set of all the galaxies is given by $G = \{g_1, g_2, \dots, g_N\}$. The probability for each galaxy of being extracted is uniform, i.e. $\mathbb{P}(\{g_i\}) = \frac{1}{N}$. We define the random variable $S : G \to \mathbb{N}^+$ which returns the dimension of the cluster s for a given input galaxy g_i.

We consider the problem of randomly extracting *a cluster*. The mean dimension of clusters is given by the weighted average of dimension w.r.t. their appearances:

$$\langle S \rangle = \frac{\sum s \cdot n(s)}{\sum n(s)} = \frac{N}{\sum n(s)}. \tag{6}$$

This corresponds to the expected value of the probability of cluster of having dimension s when picked randomly, as a matter of fact:

$$\mathbb{P}(s) = \frac{n(s)}{\sum n(s)} \quad \Rightarrow \quad \mathbb{E}[s] = \sum s \cdot \mathbb{P}(s) \equiv \langle S \rangle. \tag{7}$$

We now consider the problem of extracting *a galaxy*. We're interested in computing the probability that a randomly picked galaxy belongs to a cluster of size s. The total number of galaxies belonging to clusters of dimension s is $sn(s)$, whereas the total number of galaxies is $N = \sum s \cdot n(s)$. Thus the probability we aim to compute - which corresponds to the probability mass function (PMF) of the random variable S - is:

$$\mathbb{P}(S = s) = \frac{s \cdot n(s)}{\sum k \cdot n(k)} = \frac{s \cdot n(s)}{N}. \tag{8}$$

This allows us to compute the expected value of the random variable S, which we define as the *susceptibility* of the system:

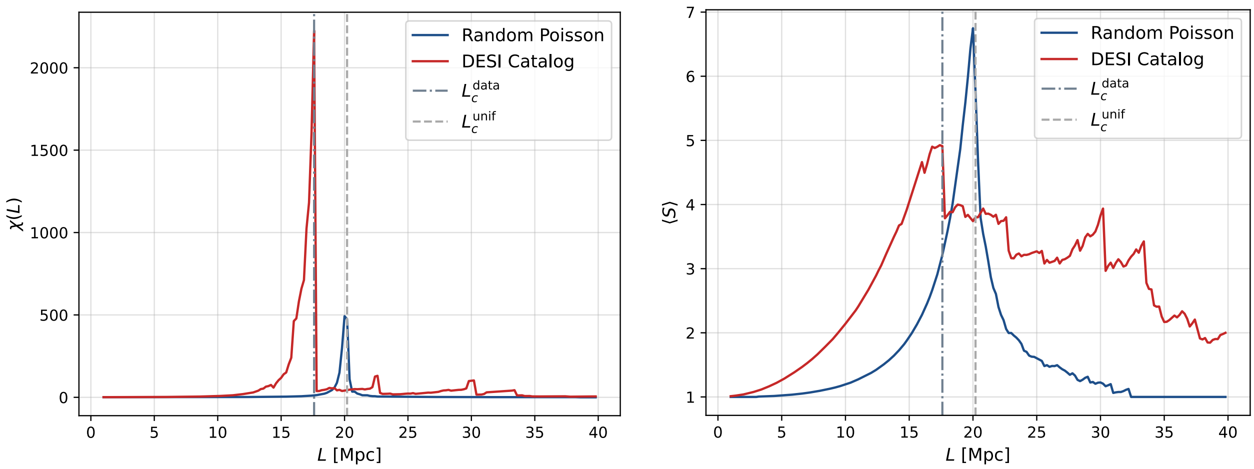
$$\chi(L) \equiv \mathbb{E}[S] = \sum_{s < s_{\max}} s \cdot \mathbb{P}(s) = \frac{\sum_{s < s_{\max}} s^2 \cdot n(s)}{\sum_{s < s_{\max}} s \cdot n(s)}. \tag{9}$$

Formally, we notice that the susceptivity is simply the fraction between the second and the first moment of S:

$$\chi = \frac{\mathbb{E}[S^2]}{\mathbb{E}[S]}. \tag{10}$$

The reason why we exclude the maximal percolating cluster $s_{\max}$ is that, acting as an order parameter, it would force χ to mirror its own behaviour rather than reflecting the response of the system. Excluding it thus allows us to quantify the "response" of the system via the variable χ.

Figure 8 shows the susceptibility χ and the mean cluster size $\langle S \rangle$ as a function of the connection length L. For both catalogs, the susceptibility reaches its maximum value around the critical length, with slightly asymmetric curves plummeting back to zero shortly after the critical length. This behaviour directly reflects the rapid merging of massive sub-clusters immediately prior to the phase transition. The DESI catalog reaches values an order of magnitude larger than the uniform one, confirming the highly clustered nature of the catalog with respect to the uniform distribution. Additionally, the secondary peaks observed in the DESI susceptibility curve at larger connection scales (e.g. around $L \simeq 22.5$ Mpc and $L \simeq 30$ Mpc) highlight the hierarchical and inhomogeneous nature of the cosmic web. These features correspond to sub-percolation events among secondary structures which temporarily merge into sub-graphs before eventually being absorbed into the primary percolating component. These secondary peaks are completely absent in the uniform catalog.



(a) Susceptibility.

(b) Mean cluster size.

Figure 8: Susceptibility and mean cluster size as a function of L for the DESI and uniform catalogs.

For the DESI catalog, $\langle S \rangle$ rises steadily up at L_c^{data} before exhibiting a broader, plateau-like feature with secondary fluctuations at larger linking scales. This behaviour is once again due to the hierarchical topology of the cosmic web, where isolated systems residing in low-density cosmic voids are progressively absorbed into the main network at larger connection thresholds. Conversely, the uniform Poisson distribution displays a smooth, symmetric curve centred at L_c^{unif}, highlighting the absence of clustered substructures in the random sample.

3.4 Characterisation of the system at critical length

We recall that s indicates the size of a cluster, whereas $n(s)$ indicates the number of appearances of that size (i.e. the number of clusters having size s). At the critical linking length L_c, the cluster size distribution $n(s)$ follows a power law:

$$n(s) \propto s^{-\tau} \quad \Leftrightarrow \quad n(s) = A \cdot s^{-\tau}. \tag{11}$$

To easily extract the critical exponent τ from our data, we transform this relationship into a linear form by taking the logarithm ($\log_{10}$) of both sides:

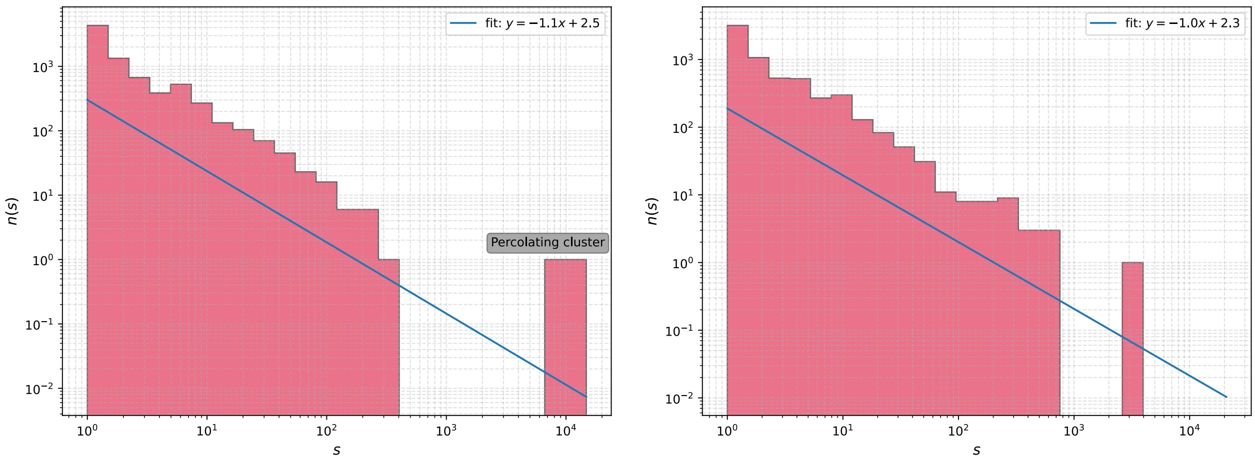
$$\log_{10}(n) = \log_{10}(A) - \tau \log_{10}(s). \tag{12}$$

In a log-log plot ($\log_{10}(n)$ versus $\log_{10}(s)$), this relation yields a straight line $y = mx + q$, where:

- the slope gives the critical exponent directly, $m = -\tau$;
- the y-intercept gives the amplitude, $q = \log_{10}(A)$.

By performing a linear fit to the cluster distribution in a neighbourhood of $L = L_c$, we can measure τ and compare the scaling behaviour of the DESI catalog against the uniform simulation.

Figure 9 shows the resulting distributions at critical length and the fit performed to verify the power law.



(a) DESI Catalog. **(b)** Random Poisson.

Figure 9: Distribution of cluster sizes at critical length.

As expected near a phase transition, the finite clusters follow a straight line in log-log space, confirming power-law behaviour. Crucially, both plots show a prominent isolated peak far to the right, representing the giant percolating cluster $s_{\max}$, which has decoupled from the continuous distribution of finite sub-clusters. Comparing the two panels in the figure, highlights two key structural differences between the LRG dataset and the random baseline, detailed below.

- Exponent τ: the DESI fit yields a 10%-steeper slope compared to the uniform sample. In a power law $n(s) \propto s^{-\tau}$, a higher value for τ means the line drops off faster. Mathematically, this tells us there are way more small clusters compared to medium-sized ones. Physically, this is gravity doing its job. Gravity works from the bottom up: it pulls matter together locally first, creating a huge population of dense, small galaxy groups.
- Percolating cluster: the isolated percolating peak $s_{\max}$ in DESI is roughly three times larger ($\sim 10^4$ galaxies) than in the Poisson catalog ($\sim 3 \times 10^3$). Cosmic filaments act as high-density "highways", enabling the main network to capture far more galaxies as soon as the threshold L_c is crossed.

3.5 Entropy

In order to somehow measure the "unpredictability" of the system, we rely on Shannon's entropy. Imagine flipping a coin: if the coin is double-sided, the outcome is completely predictable, offering zero element of surprise and thus zero entropy. Conversely, a fair coin yields maximum uncertainty, corresponding to high entropy.

Given a *discrete* probability $p(s)$ (just like in our case), the formal definition of Shannon's entropy is:

$$\mathcal{H} = -\sum_s p(s)\log_2 p(s). \tag{13}$$

Figure 10 shows Shannon's entropy as a function of the connection length for both catalogs. It can be remarked that in both cases, the peak is reached before the respective critical length. This happens because the maximum uncertainty happens when the cluster sizes have a great variability, and this cannot correspond to the instant of phase transition, as big clusters start emerging to later on merge into a giant one. A major difference between the two catalogs can be observed: while the

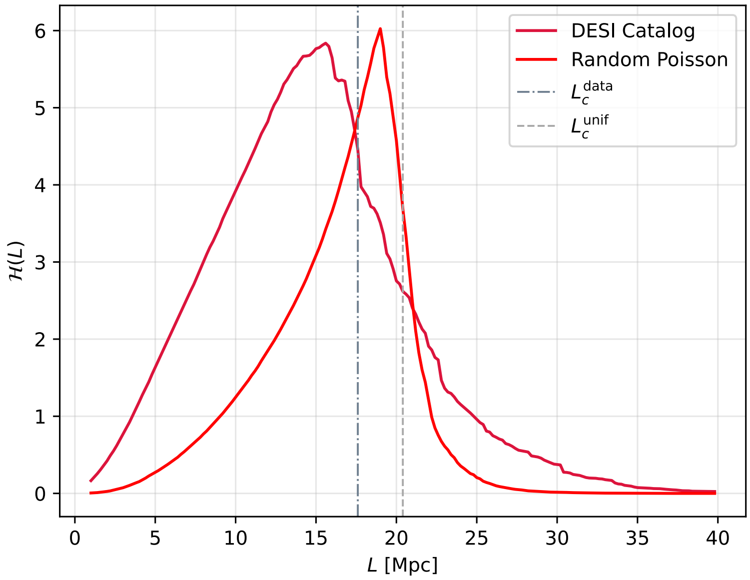


Figure 10: Shannon's entropy as a function of connection lenght.

random sample drops off super fast after its peak, the DESI catalog shows a much slower, "bumpy" decline. This long tail happens because the LRG sample isn't uniform. Even after the main cosmic network connects, smaller groups tucked away in low-density voids keep merging at larger distances, keeping the entropy alive a bit longer. This can be thus seen as a further compatible proof of the filament-like structure of galaxies, forming the cosmic web.

4 Conclusions

The visual analysis of 1015 spectra coming from the 4MOST Survey Validation dataset formed a first, reliable benchmark to assess the overall quality of the pipeline and of the telescope. This represents a valid starting point for the Laboratory of Astrophysics of EPFL, and potentially the 4MOST collaboration in order to produce further results. The accuracy of the pipeline has been assessed to be of 83.45% and 86.70% for the redshift and the object type classification, respectively, leaving some space for improvement bur already forming a solid basis. Hopefully, this will help the data analysis resulting from future 4MOST's data realeases.

On the other hand, we explored the spatial network of the cosmic web using a three-dimensional percolation approach. By analysing a select population of Luminous Red Galaxies from the DESI survey within the redshift range $0.6 < z < 0.8$, we tracked how individual galaxies connect to form larger structures compared to a purely random distribution of points. Using key physical diagnostics - such as system susceptibility, Shannon's entropy, and cluster size distributions - we mapped out the connectivity and scaling properties of the Universe on cosmological scales.

Our findings show that the real galaxy catalog reaches global connectivity at a noticeably smaller connection distance ($L_c^{\mathrm{data}} = 17.60$ Mpc) than the random baseline ($L_c^{\mathrm{unif}} = 20.40$ Mpc). Furthermore, the presence of secondary peaks in susceptibility and a lingering tail in Shannon's entropy demonstrate the complex, hierarchical nature of cosmic structure. Combined with a steeper critical exponent ($\tau = 1.1$), these findings provide a compelling case that the observed galaxy distribution is consistent with a structured framework of cosmic filaments and nodes, rather than a uniform spatial pattern.

Overall, this research gave some interesting insights on how tools borrowed from statistical physics and computation can help treat galaxies as a complex sub-system of the universe. The Universe is immensely vast and intricate, and unravelling its large-scale structure requires moving beyond traditional spatial metrics toward methods capable of capturing its underlying complexity. While these luminous galaxies act as excellent beacons for tracing the primary framework of the cosmic web, extending this percolation method to denser galaxy populations will be vital for capturing its fainter, low-density bridges. Finally, while simple distance-based linking provides a strong foundational picture, combining percolation analysis with advanced pattern-recognition frameworks - such as topological algorithms like DisPerSE [5] - will allow future research to more clearly separate potential filaments from surrounding cosmic walls and voids [6], deepening our understanding of the cosmic network.

The environmental impact of this research

As climate change challenges the Earth's wellbeing and our daily lives, it's important to assess the environmental impact of our actions, including our work. Because this work is primarily theoretical and computational, its environmental impact remains minimal. All code was executed locally on a personal laptop, avoiding high-performance computing clusters or carbon-intensive data centres. Additionally, Large Language Models (LLMs) were queried at a low-intensity level, restricting their usage, via free plans, to code optimisation, debugging, text or mathematical review, and the browsing of sources and scientific literature. This kept the associated computational footprint limited.

Bibliography

[1] 4MOST et al. "4MOST: Project overview and information for the First Call for Proposals". In: *ESO Messenger* 175 (2019). ISSN: 0722-6691. arXiv: 1903.02464 [astro-ph.IM].

[2] DESI Collaboration et al. "Data Release 1 of the Dark Energy Spectroscopic Instrument". In: *The Astronomical Journal* 171.5 (2026). arXiv: 2503.14745 [astro-ph.CO].

[3] Paul Martini et al. "Overview of the Dark Energy Spectroscopic Instrument". In: *Ground-based and Airborne Instrumentation for Astronomy VII*. Vol. 0702. 2018. ISBN: 9781510619579. arXiv: 1807.09287 [astro-ph.IM].

[4] Planck Collaboration et al. "Planck 2018 results. VI. Cosmological parameters". In: *Astronomy & Astrophysics* 641.A6 (2021). arXiv: 1807.06209 [astro-ph.CO].

[5] Thierry Soubsie. "DisPerSE: robust structure identification in 2D and 3D". In: *Astrophysics Source Code Library* (2013). arXiv: 1302.6221 [astro-ph.CO].

[6] E. Tempel et al. "Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the SDSS". In: *Monthly Notices of the Royal Astronomical Society* 438.4 (2014), pp. 3465–3482. ISSN: 0035-8711. arXiv: 1308.2533 [astro-ph.CO].