Topological Hypothesis Tests for Large-Scale Structure

Mike Wu*[1], Jessi Cisewski[2], Brittany Fasy[3], Wojciech Hellwin[4], Mark R. Lovell[5], Alessandro Rinaldo[6], and Larry Wasserman[6]

[1]Department of Computer Science, Yale University
[2]Department of Statistics, Yale University
[3]Department of Computer Science, Montana State University
[4]ICG, University of Portsmouth
[5]ITF, University of Amsterdam
[6]Department of Statistics, Carnegie Mellon University

May 5, 2016

Abstract

Cosmological simulations allow for visualizing the observable Universe under varying physical assumptions. Employing summaries of topological and geometric features, hypothesis tests are developed using persistent homology to quantify the differences in the resulting large-scale structure due, for example, to the spatial complexity of two or more simulations. A generic hypothesis testing framework is tested with Voronoi simulations, and then applied to a pair of simulations of the Megaparsec cosmic mass with underlying assumptions of warm and cold dark matter respectively. Using such a framework, we demonstrate that while warm and cold dark matter produce similar global topological structure, there are statistically significant differences in both topology and geometry in smaller scale observations.

1 Introduction

Rigorous comparisons of spatially complex data such as the large-scale structure (LSS) of the Universe is notoriously difficult due, in part, to the difficulty in capturing the randomness of geometric and topological structures. However, these comparisons are important as it is becoming apparent that there is potentially information about cosmological parameters in the structure. We propose a framework for constructing topological hypothesis tests using ideas from an emerging area of topological data analysis called persistent homology. Persistent homology offers a novel way to represent, visualize, and interpret complex data by extracting topological features, which can be used to infer properties of the underlying structures, and has already been used as an exploratory tool for some problems in astronomy ([7, 16]) among other areas of science ([1, 9]).

*The authors gratefully acknowledge *omg so many people*

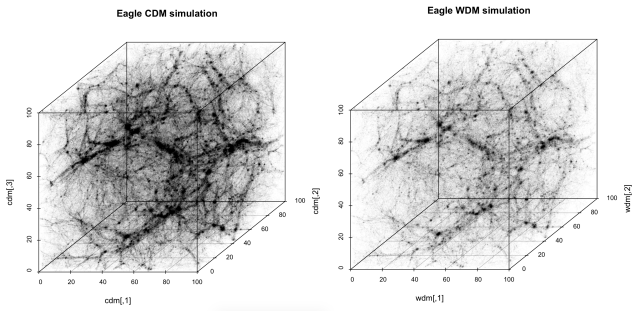

Figure 1: 3D models of the observable Universe under assumptions of cold (left) and warm (right) dark matter. Images generated with the EAGLE dataset [13].

In the field of astronomy, Van de Weygaert et. al. [11] studied the topology of both the simple heuristic Voronoi clustering models and the Megaparsec Cosmic Web (LCDM simulation) using tools from algebraic topology, such as Betti numbers and Alpha shapes. The *Betti numbers* are a topological invariant that counts the number of n-dimensional holes in a space. In this paper, we are considering structures in $\mathbb{R}^3$ such that the holes represent components ($n = 0$), loops ($n = 1$), and voids ($n = 2$). Van de Weygaert et. al [11] further studied the evolution of the Betti numbers at different scales in the context of alternative dark energy assumptions behind LCDM simulations by looking at the persistent homology (a multi-scale view of homology at which topological features arise through "multiple scales" of time and/or space).

In this paper, we move beyond the ideas introduced by [11] towards quantitatively detailing the topological and geometric differences between the observable Universe under respective assumptions of cold and warm dark matter. See Figure 1 for a visual comparison of a segment of the observable Universe under CDM (left) and WDM (right). The CDM image is notable more dense, agreeing with predominant thought that cold dark matter, by having less kinetic energy, moves slower and leads to the formation of more LSS. Provided with the EAGLE simulations [13], we want to provide strict numbers to measure the differences in topological features. Such a framework could inform insights into the effects of dark matter and its properties. To quantify topological differences, we create a hypothesis testing framework to compare functional summaries of persistence diagrams that generalize to any point cloud data set. The hypothesis testing framework is tested and verified on a set of Voronoi foam models where the true distribution of persistence diagrams is known. Using the EAGLE data set [13], we propose a form of bootstrapping called splicing to artificially expand a data set. Using the hypothesis testing framework, we find that while the CDM and WDM assumptions produce similar topological features globally, they contain statistically significant geometric and topological differences in smaller scale splits.

2 Persistent Homology

Homology is the study of the spatial structure of topological spaces (e.g., subsets of $\mathbb{R}^n$ and k-nn graphs constructed over point cloud data). Persistent homology

computes the homology generators of a topological space as the structure of the point cloud data evolves. As the topological space transforms, it causes the homology groups to change as well: generators can appear (or be born) and disappear (or die). More strictly, these generators take the form of connected components, loops, and voids. A *connected component*, or 0-th dimensional homology is a maximal set of a topological space that cannot be covered by two disjoint open sets. Given an arbitrary function f, the *birth* of such an 0-th dimensional homology, or H_0, occurs at the local maximas in f. Respectively, the *death* of an H_0 generator occurs when two such connected components merge, which can happen at a stationary saddle point of f. Additionally, along with the death of an H_0, saddle points can also witness the birth of a one-dimensional homology generator, H_1, known as a *loop*. Similar functional patterns arise for higher dimensional generators, like *voids*.

Various methods can be used in order to transform a discrete point set into a topological space with defined homology generators. For example, points can be connected based on a distance (or a distance-like structure as in [3]), or one can estimate the density from which the points were sampled. In the latter case, one would look at a density function, such as a kernel density estimate (KDE) of a point cloud and study the topological features of super-level sets of that density. Figure 2 shows a density function p created by a circle of radius 5 with 5 smoothed Gaussian peaks of randomly generated height. To derive the persistent homology for p, let there exist a threshold r, represented by a hyperplane that divides p into two separate segments: a super-level set, defined as $\{(x, y, z) \in p \text{ s.t. } z \geq r\}$, and a corresponding sub-level set $\{(x, y, z) \in p \text{ s.t. } z < r\}$. r is initialized at ∞, at which the super-level set is $\emptyset$ and the sub-level set contains all of p. The evolving spatial structure is represented by r approaching $-\infty$. The persistent homology would then track the connected components (H_0), loops (H_1), and voids (H_2) that appear and disappear in the super-level sets $p^{-1}([r, \infty)]$ as r ranges from ∞ to $-\infty$. More specifically, as r intersects p, the super-level set is no longer $\emptyset$ and is instead, composed of disjoint peaks/local maxima (see pane 1 of Figure 2). Each of the intersections of a Gaussian peak with r is a birth of an H_0. Once r decreases sufficiently such that all H_0 merge to form a cyclic structure (see pane 2 of Figure 2), an H_1 is born and all previous connected components die. This first homology continues to exist until potentially further merging with other first homology groups or r decreases past p.

Figure 3 (right) shows the resulting persistence diagram for density p, which is a summary of the density's homology generators, plotting the birth time against the death time. The diagonal line in the middle of the Figure, represents the point at which the birth time is equivalent to the death time, suggesting non-existence. Therefore, the farther away homology generators are from the diagonal line, the more *persistent* they were. Further referencing Figure 3, the black dots, red triangles, and green pluses (none shown) respectively represent H_0, H_1, and H_2. The black dots and red triangles in the Figure represent a one-to-one correspondence between the homology groups and the local extremas in Figure 3. For example, notice that peak number 1, being the tallest one, has the H_0 with the earliest birth. Number 6 represents the location at which all

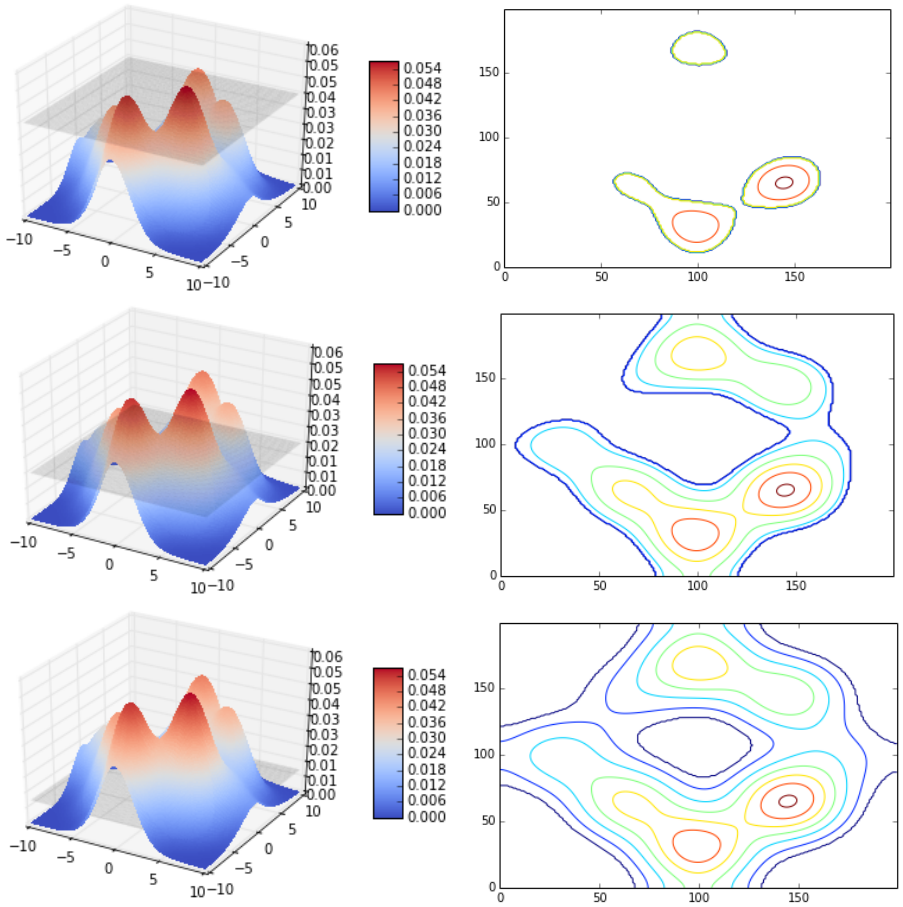

Figure 2: On the left is the KDE of points sampled from a circular loop of radius 10 with 5 Gaussian peaks as the hyperplane lowers through the density. The corresponding figures on the right display the super-level sets defined by the hyperplane on the left. As the hyperplane, r intersects with each of the Gaussian peaks, an H_0 homology is born. The corresponding contour shows disjoint circles that represent each H_0 birth. As r continues to decrease, many of H_0's merge, symbolized by the merging of the contour circles in the figures to the right. Finally, as r is small enough such that the set of existing H_0's merge into a loop, an H_1, or first-order homology, is born, symbolized by the full cycle generated in the third contour image.

H_0's have merged into an H_1, denoted in the density by the location of the hyperplane, r. Generally, the birth of an H_0, or a black point in the persistence diagram, represents local peaks in the density being intersected by the hyperplane. Additionally, there exists a single (non-trivial) H_1, red triangle, that was born later. This represents the *true* circle/loop from which the points were sampled. Because persistent homology is also subject to noise, there may be additional homology groups close to the diagonal line that likely represent perturbations in the topological summary. Furthermore, the point labeled N is an artifact of the persistent homology calculation and should be disregarded when conducting further analysis.

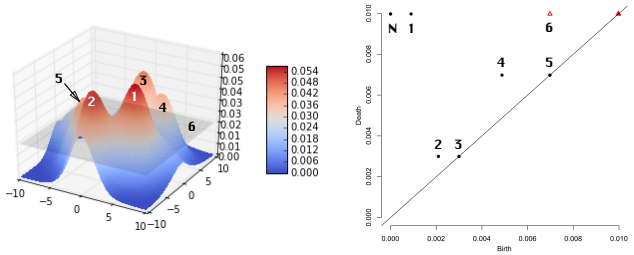

Figure 3: (left) Kernel density estimation of an arbitrary function with varying peaks around a circle. The area above the hyperplane is referred to as the superlevel set. (right) Persistent diagram showing births and deaths of homology cycles in the example density function. Numbers are provided to relate the H_0 and H_1 objects to the density. N is an artifact of persistent homology and should be disregarded.

2.1 Distance to Measure

Distance to measure [3, 4] (DTM) is an alternative way to compute a distance-based topological summary and is often used in persistent homology as another method besides KDE of computing a persistence diagram. Given a set of points $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ sampled from some distribution P, DTM approximates topological features of the underlying space. It is defined for each $y \in \mathbb{R}^d$ as

$$d_{m_0}(y) = \sqrt{\frac{1}{m_0}\int_0^{m_0}(G_y^{-1}(u))^2 du}$$

where $G_y(t) = P(\|X - y\| \leq t)$ and $0 < m_0 < 1$ is a smoothing parameter similar to h in the KDE definition. Given the points in X, the empirical version of DTM is

$$\hat{d}_{m_0}(y) = \sqrt{\frac{1}{k}\sum_{x_i \in N_k(y)} \|x_i - y\|^2}$$

where $k = \lceil m_0 n \rceil$ and $N_k(y)$ is the set containing the k nearest neighbors of $y \in X$. One should best think of DTM as a smoothed version of the distance function.

3 Methods

As described, given a single point cloud, persistent homology provides an an algorithm to summarize topological information from data. However, there is not an existing method to compare the topological summaries of a set of two or more point clouds. In this paper, we propose a hypothesis testing framework built on top of the multi-scale topological summaries. Given any persistence diagram x extracted from a point cloud, x represents a sample from some latent distribution χ. For example, given a simulation of the Megaparsec cosmic mass, the (x, y, z) points are discrete samples from the continuous, observable Universe, which represents the true distribution. Using that definition, suppose there are two persistence diagrams, x and y, each produced from a separate point cloud. Given that $x \sim \chi$, and $y \sim \gamma$, where χ and γ represent the true distributions, an interesting question is whether χ and γ are identical to some error ϵ. In other words, are the samples x and y sampled from the same latent distribution? One might consider the following hypothesis test, where H_0 is the null and H_A is the alternative.

$$H_0 : \chi = \gamma$$
$$H_A : \chi \neq \gamma$$

However, given that the parameters of the distributions *chi* and γ are unknown, one cannot directly compare them. Instead, one much infer the relationship between *chi* and *gamma* by comparing the diagrams sampled from χ and γ respectively. Given n samples, a hypothesis test should be designed to compare $\{x_1, ..., x_n \sim \chi\}$ and $\{y_1, ..., y_n \sim \gamma\}$. However, comparing two persistence diagrams, x and y, is non-trivial. Naive distance calculations, like bottleneck or Wasserstein, are easily perturbed by randomness and noise, and are computationally expensive. Instead, we propose to further summarize a persistence diagram by a test statistic. To find such a test statistic, we must define a function f_T that takes an input x and produces a statistic T_x, $f_T(x) = T_x$. Provided such a function exists, using the further summarized statistics, we can test the hypothesis that two persistence diagrams, x and y are sampled from the same latent distribution with the following framework:

$$H_0 : f_T(x) = f_T(y)$$
$$H_A : f_T(x) \neq f_T(y)$$

One main contribution of this paper is studying the effectiveness of different functions f_T when the true topological relationship between x and y are known. The best functions f_T are then used to analyze a data set in which the topological relationships is unknown. We now proceed to describe different methods for defining f_T for hypothesis testing in the setting of large-scale structure (LSS). Five tests are considered: EC, Sil., Sil.(EC), GKD, and GC.

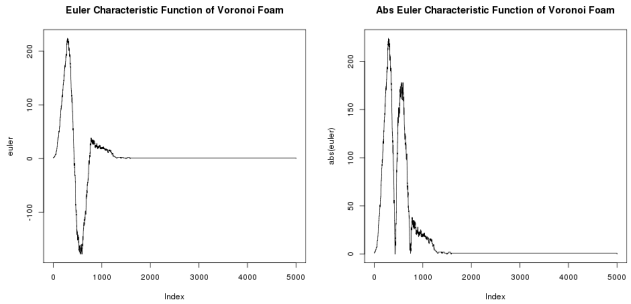

Figure 4: (left) An Euler characteristic function created by plotting the Euler characteristic of a Voronoi foam model at each point beginning at time 0 until the last birth time. (right) The absolute value of the characteristic function that is integrated to produce the test statistic E_t.

Euler Characteristic Function (EC). The Euler characteristic (EC) is a topological invariant and defined as an alternating sum of Betti numbers, where N is the number of dimensions:

$$\chi = \beta_0 - \beta_1 + \beta_2 - \beta_3 + ... \pm \beta_N = \sum_{i=0}^{N} (-1)^i \beta_i$$

where β_i represents the i-th Betti number, or the rank of the i-th homology group. Therefore, when analyzing persistence diagrams of LSS, there exist only three dimensions of data, which means that the only non-trivial homology groups will be in dimensions zero, one, and two. Given the Betti numbers β_0, β_1, and β_2, the Euler characteristic equation simplifies to:

$$\chi = \beta_0 - \beta_1 + \beta_2.$$

In the persistent homology framework described above, the topological space is parameterized by the threshold (hyperplane), and the homology groups (and hence Betti numbers) can be computed using H_0, H_1, and H_2. Plotting the Euler characteristic against t, we obtain the Euler Characteristic function, where t represents threshold at an observation. See Figure 4 for an example of an characteristic function on a Voronoi foam.

$$E_t = \int_t |\chi| \, \mathrm{d}t$$

The absolute value of the resulting Euler function is integrated with respect to time using trapezoidal summation to produce an Euler statistic, E_t. Given two sets of Euler statistics (one set from each sample in the hypothesis test), a T-test is used to calculate a p-value.

Silhouette Test (Sil.) The persistence diagram is a set of points in the upper-half plane. Transforming this diagram into a continuous one-Lipschitz function allows us to use more tools from statistics. In particular, the weighted silhouette functions are formed by weighting a particular functional summary

of persistence diagrams called *landscape functions* [2]. Details and theoretical properties of landscapes and silhouettes are provided in [5].

Let the finite birth and death intervals of a persistence diagram with n_h points, for homology dimension $h = 0, 1, 2, \ldots$, be defined as $\{(b_{hi}, d_{hi})\}_{i=1}^{n_h}$. Next consider rotating the persistence diagram such that a given point is $p_{hi} = \left(\frac{b_{hi}+d_{hi}}{2}, \frac{d_{hi}-b_{hi}}{2}\right) \in D_h, \quad i = 1, \ldots, n_h$. Equilateral triangles are formed from each p_{hi} to the base as

$$\Lambda_{p_{hi}}(t) = \begin{cases} t - b_{hi} & t \in [b_{hi}, \frac{d_{hi}+b_{hi}}{2}] \\ d_{hi} - t & t \in [\frac{d_{hi}+b_{hi}}{2}, d_{hi}] \\ 0 & \text{otherwise} \end{cases}$$

where $t \in [0, T]$ for a positive real number $T \geq \max\left\{\frac{b_{hi}+d_{hi}}{2}\right\}$.

For a given h, the persistence landscape is then defined as the following collection of functions

$$\lambda_{D_h}(k, t) = \underset{p_{hi} \in D_h}{\text{kmax}} \Lambda_{p_{hi}}(t), \quad t \in [0, T], k = 1, \ldots, n_h$$

where kmax is the kth largest value in D_h.

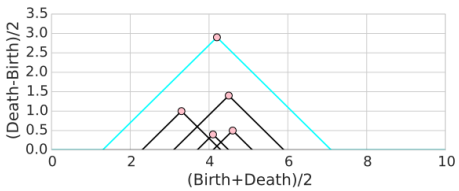

Figure 5: The pink circles are the points in a persistence diagram D. The landscape $\lambda(k, \cdot)$ is the k-th largest of the arrangement of the graphs of $\{\Lambda_p\}$. In particular, the cyan curve is the landscape $\lambda(1, \cdot)$.

Rather than work with each k of $\lambda_{D_h}(k, t)$ individually, silhouettes are weighted averages of the individual functions for homology dimension h defined as

$$\phi_h(t) = \frac{\sum_{i=1}^{m} w_{hi} \Lambda_{hi}(t)}{\sum_{i=1}^{m} w_{hi}}$$

where the weights w_i can be defined to give more emphasis or less emphasis to features with longer lifetimes. We use $w_{hi} = |d_{hi} - b_{hi}|^p$, where p is a tuning parameter that needs to be selected.

For hypothesis testing, as with the EC function, a weighted silhouette summarizes a persistence diagram through some continuous function. However, unlike the EC, the silhouette must be calculated per dimension (Sil_h). In order to carry out a hypothesis test, each data set was summarized by a weighted silhouette, and a T-test was performed on the areas under the weighted silhouettes. A silhouette can be calculated for each dimensional homology. Unfortunately, the silhouette calculated through the only the connected components may be

very different from the silhouette computed from the loops. As an attempt to stabilize the silhouette statistic across multiple dimensions of homology, for each persistence diagram, a parallel silhouette test ($\text{Sil}_{0:2}$) is used to evaluate all dimensions at once using a Hotelling T^2 test.

Silhouette-Euler Characteristic (Sil (EC)). A method for combining the individual silhouettes, S_i, across dimensions 0, 1, and 2 is to define a modified Euler characteristic function. Instead of calculating the alternating sum of Betti numbers b_0, b_1, b_2, the Silhouette-Euler characteristic function is defined as the alternating sum of the individual silhouette functions over time. The statistic returned T is the integral of the absolute value of the summed function,

$$T = \int_t |S_1 - S_2 + S_3| \, \mathrm{d}t_1 t_2$$

As with EC, a p-value is derived using a paired T-test.

Global Contour Test (GC). Rather than working with the raw persistence diagrams, the Global Contour Test (GC) uses a test statistic derived from a kernel two-sample test on the smoothed persistence diagram called the *intensity function* [6]. The kernel two-sample test was first introduced for analyzing and comparing distributions with a maximum mean discrepancy (MMD) statistic [10]. The GC statistic used in this paper is computed for two sets of persistence diagrams and a hypothesis test is carried out using a permutation test. The discrepancy between persistence diagrams is calculated as the integrated squared difference between two intensity functions instead of points directly from the persistence diagrams. The two-sample test statistic, T, is defined as

$$T = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(X_i, X_j) - \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} K_h(X_i, Y_j) + \frac{1}{m^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(Y_i, Y_j),$$

where n and m are the sizes of the two samples, and $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ are the two sets of intensity functions. $K_h(x, y)$ can be thought of as a similarity measure between intensity functions x and y, and in this case is a Gaussian kernel $K_h(X, Y) = exp(-\frac{||X-Y||^2}{h^2})$ with $||X - Y|| = \int \left(X(t_1, t_2) - Y(t_1, t_2)\right)^2 dt$ where $X(t_1, t_2)$ and $Y(t_1, t_2)$ are two intensity functions. h is a hyperparameter that sets the standard deviation of the Gaussian distribution used to model K_h. A larger h will reduce sensitivity to small differences between X and Y, while a smaller h will heighten sensitivity. The optimal h value was found to be 1 ± 0.2 using greedy search from 0 to 5, where optimality is measured by both best preserving a high p-value when X and Y are very similarly and best returning a low p-value when X and Y are significantly different. A permutation test is used with N permutations, and the p-value represents the fraction of times T_j is larger than the observed T for any j,

$$\text{p-value} = \frac{1}{N}\sum_{j=1}^{N} I(T_j \geq T).$$

Global Kernel Density Test (GKD). The Global Kernel Density Test (GKD) also uses a kernel density estimate of the persistence diagrams, but relies on a test statistic which asymptotically follows a normal distribution thus providing *approximate* p-values. More details on this test can be found in [8, 9].

4 Results of Simulation Study

4.1 Topological Analysis of Voronoi Universes

Prior to analyzing the EAGLE simulations, we performed a simulation study using Voronoi foam models. Icke and Van de Weygaert [11, 15] introduced the Voronoi foam as a packing of polyhedral units with walls representing pancakes, edges representing filaments, and vertices representing clusters in the galaxy. Icke showed that the Voronoi foam is an appropriate model for a 10-500 Mpc scale Universe with pressure-free Newtonian gravitational collapse. Statistical study showed that the spatial two-point correlation of Voronoi foams has a power law behavior with close to identical amplitude and slope as that of actual Abell clusters [15].

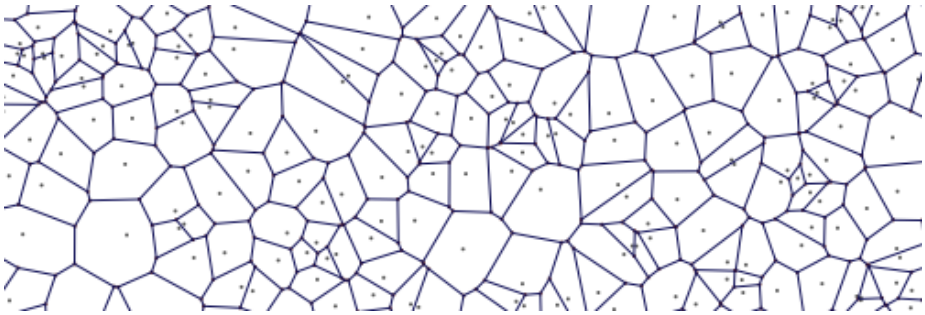

Figure 6: Example of a large Voronoi tessellation from which points can be sampled from to generate a desired dataset. The number of faces and cells are variable and can be tuned.

A Voronoi foam model, shown in Figure 6, is generated from a tessellation, where the edges of each cell represent filaments and the faces of the cell, enclosed voids. Given a plane with fixed size, such a tessellation partitions the plane into cellular regions with nuclei. Every cell territory is defined as the set of points equal or closer to that cell's nuclei than any other. To produce a simulation in polynomial time, each point in the plane is compared to the k closest nuclei using a nearest neighbor algorithm where k is chosen to be a small integer. Gaussian noise is added to perturb the plane and inject randomness. Because the nuclei number and the plane size are variable, the points in the simulation representing filaments, clusters, and walls are variable as well. By choosing a reasonable percentage of filaments and related structures, Voronoi simulations are to approximate the topology and LSS of true simulations of the observable Universe. By varying these percentages and repeating simulations, one can quickly generate a large, labeled data sets for hypothesis testing. Our interest primarily was gauging the effect of changing the percent filament in the Voronoi tessellation on the ability of the hypothesis tests to distinguish two foam models

sampled from different tessellations. In this paper, we varied only the filament percentage (percFil) from 0.1 to 0.9. All other parameters, see table 1, were kept constant. Figure 7 plots three separate Voronoi foam models of percFil 0.1, 0.5, 0.9 respectively. One can see that the higher the percFil, the more clusters appear in the foam.

Abbrev	Definition	Value
percWall	Percentage of particles on the walls	$0.98 - p_f$
percFil	Percentage of particles on the filaments	p_f
percClust	Percentage of particles on the clusters	0.02

Table 1: Table of parameters in definiting a Voronoi foam model. p_f represents a variable taking the range from 0.1 to 0.9.

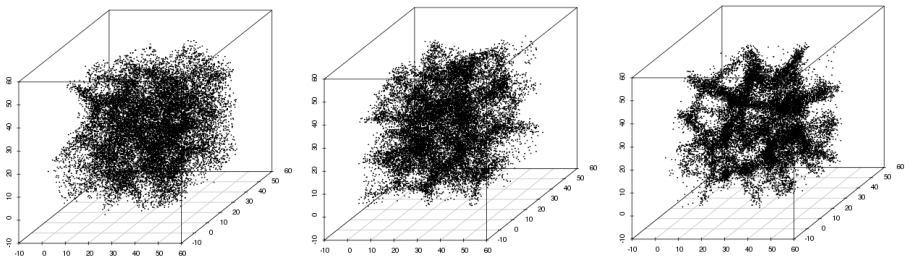

Figure 7: (left) PercFil 0.1; (middle) PercFil 0.5; (right) PercFil 0.9. All other parameters were fixed to constants.

For reference, Voronoi foam models used in this paper were generated under 1.25×10^5 box volume, 0.1 resolution, 1×10^4 points, 64 cells, 0.02 percClust, 0.00 percClutter, and $[0.1, 0.9]$ percFil and $[0.08, 0.88]$ percWall. Persistent homology is calculated using DTM with a 0.01 tuning parameter. Unless using an Euler-based metric, persistence diagrams are preprocessed to remove the known 0-dimensional artifact wherein persistent homology algorithms produce a vestigial H_0 element with birth time of 0 and a death time of 1. See figure 3 for an example of such an artifact, as labeled N in right subfigure. Optionally, persistence diagrams are standardized by the maximum birth or death value after removal. The hypothesis tests were performed on 100 independent iterations on 15 sets of Voronoi foam realizations. Each set includes 10 individual Voronoi foam models, 9 variable models and 1 control. The variable foam models are each generated using a percFil setting from 0.1 to 0.9; the control model takes a fixed percFil setting of 0.1. Using the hypothesis tests, each of the variable models will be compared to the control. More specifically, given the variable model with percFil p, each of the hypothesis testing frameworks computes a final p-value as the median p-value comparing variable model p with the control model across the 15 sets. Doing so for each of the 9 variable models provides a sense of how increasing the filament percentage changes the statistical confidence in rejecting or failing to reject whether two foam models are sampled

from the same topological distribution. To vary the control, similar tests were also completed against a control foam model of 0.9 percFil, and identical results were found. Figure 8 shows line plots of the resulting final p-values from four sets of hypothesis testing frameworks: (1) euler-based tests, (2) silhouette-based tests, (3) smoothing-based tests, and (4) best-of-the-best tests, where the best test from each of the previous 3 categories are compared.

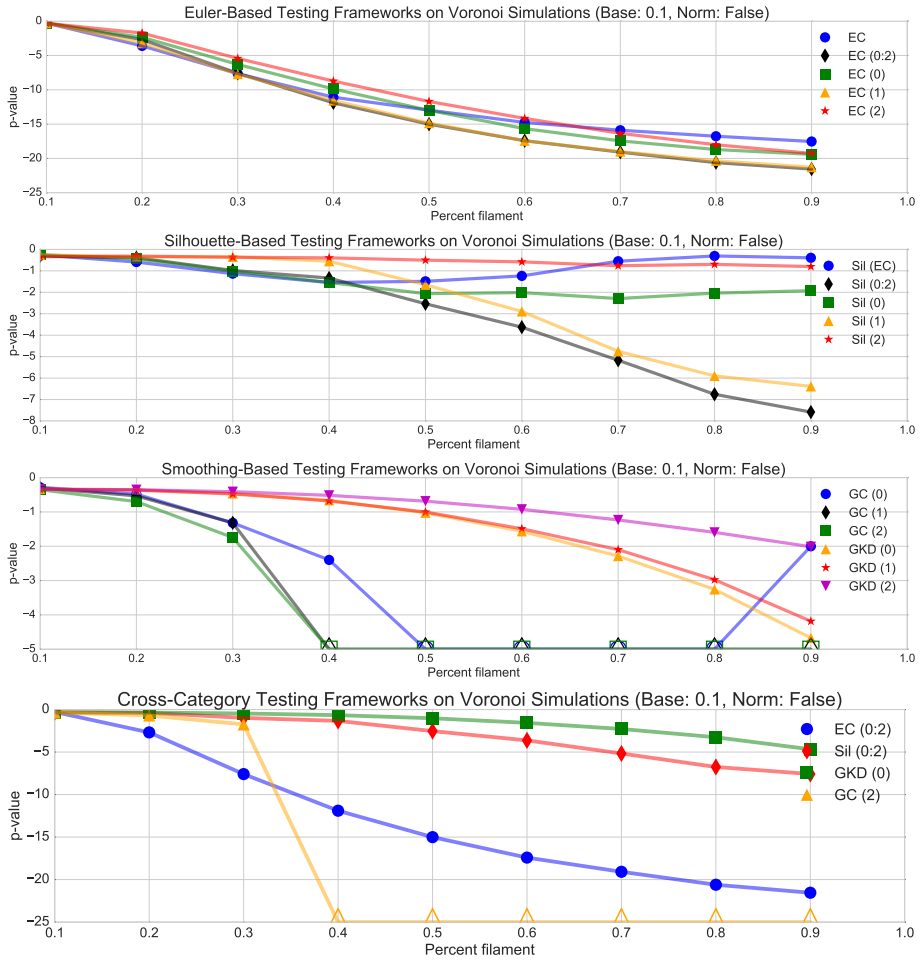

Figure 8: Line plots of three classes of hypothesis testing frameworks: Euler-based, Silhouette-based, Smoothing-based (top to bottom). The fourth plot (bottom) includes the best hypothesis test from each of three frameworks. X-axis represents the percent filament of the set being compared to the baseline set; y-axis shows the p-values. The lines plot the median p-value of the 100 iterations.

The best three testing frameworks are the [1] $EC_{0:2}$, [2] $Sil_{0:2}$, and [3] GC_2. These three frameworks produce significantly lower p-values as the percFil of the test set differs more from the percFil of the control; this is seen by the increasingly negative slope in figure 8, shown in log_{10} space. $EC_{0:2}$ and GC_2 are able to uncover significant differences early (between 0.1 and 0.5 percFil)

whereas other tests were less differentiable. Between these two tests, GC_2 expressed a more sudden and drastic shift, returning a p-value approaching $-\infty$ in log_{10} space, while $EC_{0:2}$ displayed a more gradual and predictable decay. Another interesting observation is that the set of GKD hypothesis tests are uniquely convex, making it better suited to answer questions of "how different" for higher filaments compositions. Contrastingly, many tests consistently perform inferior: Sil_0, Sil_2, Sil_{EC} show no discernible pattern among different filament percentages, most likely because most, if not all, of the differentiating information is captured by Sil_1.

4.1.1 Paired T-tests

In the prior methods, all T-tests were two-sampled. However, after studying the WDM and CDM data sets, we found that it's highly likely that the sampled WDM and CDM persistence diagrams are correlated and thereby, not independent. Therefore, the hypothesis tests should instead check whether the mean of the differences between two paired test statistics $f_T(x)$ and $f_T(y)$ differ from 0 given persistence diagrams x and y. All methods, when applied to the EAGLE data set include paired T-tests and paired Hotelling T^2 tests.

4.1.2 Standardization of Persistence Diagrams

A possible preprocessing step to hypothesis testing is standardizing the persistence diagrams, in which all the homology coordinates for (birth,death) of H_0, H_1, H_2 are re-scaled to $[0, 1]$. Without standardizing, certain model parameters, like box size or spatial resolution, produce differently scaled axes on the persistence diagrams. Standardizing these axes may increase comparability between simulation sets prior to hypothesis testing.

As shown in figure 9, standardization has a drastic impact on hypothesis test results, decreasing the effectiveness of several hypothesis tests. Notice, however, that the "best" tests remain the same as those from the unstandardized setting. Primarily, it seems that the GKD and Sil_1 have significantly reduced effectiveness, remaining roughly the same across the spectrum of percent filaments. All of the Euler characteristics are also less significant in their ability to discern differences, shown by the overall higher p-values. Remarkably, $EC_{0:2}$, $Sil_{0:2}$, and GC_2 remain superior. Geometrically, standardizing the persistence diagrams is equivalent to disregarding geometry in favor of structure. For example, a circle of radius 5 is considered topologically equal to a circle of radius 500 after standardization. The results possibly indicate that the GKD and Sil (1) might be more primed to detect geometric differences (size) while the $EC_{0:2}$, $Sil_{0:2}$, and GC_2 are more proficient at detecting structural differences.

4.2 Topological Analysis of the LCDM Universe

Similar to Voronoi foam models, the Megaparsec cosmic mass distributions are likewise characterized by an also intricate multiscale configuration of web-like filaments and voids. Using the tools and testing frameworks developed from the Voronoi simulations, we study simulations of the Megaparsec cosmic mass and evaluate similarities and differences in LSS based on variations in the parameters of the simulation. One particular parameter that we are interested in is the

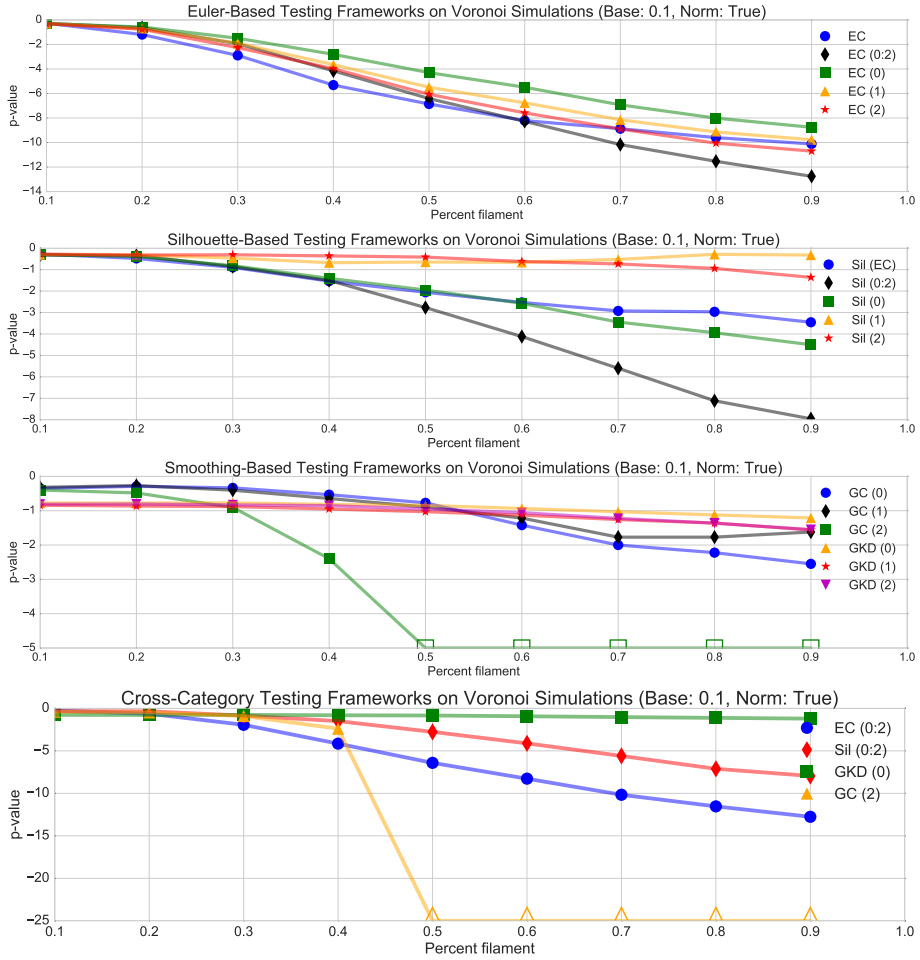

Figure 9: Line plots of three classes of hypothesis testing frameworks: Euler-based, Silhouette-based, Smoothing-based (top to bottom). The fourth plot (bottom) includes the best hypothesis test from each of three frameworks. All axes were standardized prior to hypothesis testing.

state of the dark matter in the simulation. Warm dark matter (WDM) and cold dark matter (CDM) are traditionally believed to produce very different realizations of the observable Universe, with the latter involving more slowly moving particles prone to produce a more lumpy distribution of galaxies and clusters in the LSS. The former, containing more kinetic energy, has higher resistance to formation of global structure, and is theorized to render a less topologically-interesting cosmic mass. Applying the proposed hypothesis testing framework would offer a method to *quantify* the topological differences between simulations with underlying WDM and CDM assumptions.

4.2.1 The EAGLE project

We analyze N-body simulations of structure formation. The simulation box is 100 co-moving Mpc on a side, and the numerical integration of the gravitational forces is run from redshift 127, when the age of the Universe is much less than 10 Myr, to the present day (13.8 Gyr). The cosmological parameters are consistent with the seven year results from the WMAP satellites: matter density $\Omega_0 = 0.272$, dark energy density $\Omega_\Lambda = 0.728$, Hubble parameter $h0 = 0.704$, spectral index $n_s = 0.967$, and power spectrum normalization $\sigma_8 = 0.81$. The mass of the simulation particle is 8.8×10^6 Msun. Haloes and subhaloes were identified using the SUBFIND algorithm [14], and the smallest halo that can be resolved has 20 particles. These runs were performed to be dark matter-only counterparts to the the hydrodynamical runs of the EAGLE project [13]; we stress that the runs used in this paper use gravity alone.

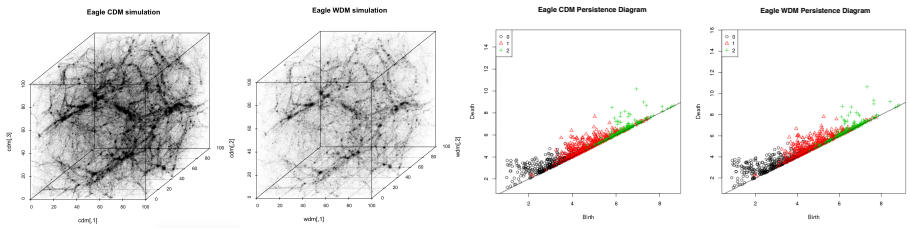

Figure 10: (left) Visualization of the complete CDM and WDM simulations. (right) Their corresponding persistence diagrams. Although the CDM structure is denser, the persistence diagrams appear comparable.

We use two simulations in this study, one cold dark matter (CDM) and the other warm dark matter (WDM). They make use of the same initial phases, and differ in that the latter has wave amplitudes rescaled using the transfer function of a 3.3 keV thermal relic, the relic mass chosen to be in agreement with the Lyman-alpha constraints of [17]. This results in the suppression of structure on the scale of dwarf galaxies. Spurious subhaloes have been removed using the algorithm of [12]. Figure 10 shows a scatter plot of both the CDM and WDM simulations along with their respective persistent diagrams. Visually, we can see that the CDM scatter plot is far more dense than WDM but share similar internal structure; the persistence diagrams also share a general structure but we can identify smaller differences in homology groups that we hope to quantify using the hypothesis testing framework. These diagrams were generated under a volume of 1×10^6, resolution of 2 and a DTM distance function with hyper-

parameter 0.001. Because the WDM and CDM simulations are each only a single data set, most of the aforementioned hypothesis tests do not apply directly. Only the global kernel density test (GKD) is immediately appropriate, which produced a 0.694 p-value, failing to reject the null hypothesis that the two simulations are sampled from the same data set. The GKD test suggests that there does not exist a statistically significant difference between the topology of the observable Universe under WDM and CDM assumptions.

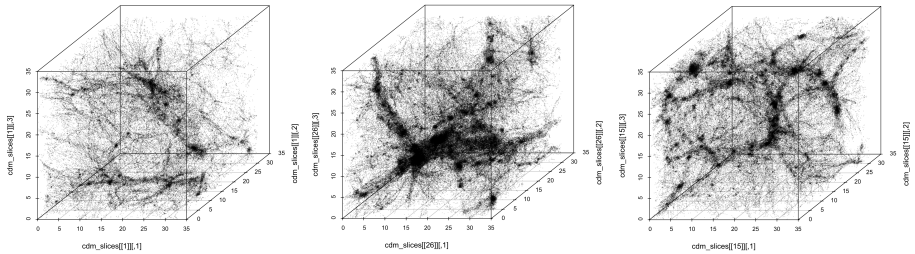

Figure 11: Examples of three triple-split samples from CDM simulation.

4.2.2 Cubic Slicing

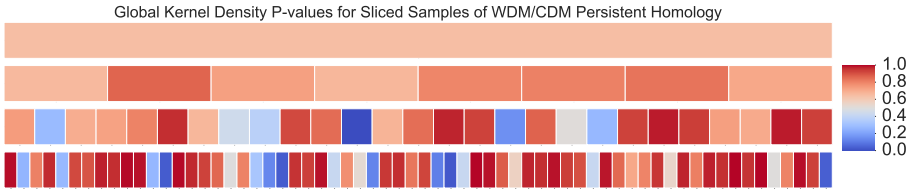

Figure 12: Four horizontal heatmaps respectively representing the p-values returned by GKD per slice for unsplit, double split, triple split, and quadruple split (top to bottom) slices of respective WDM and CDM data sets.

Since theoretically and visually, the persistence diagrams suggest that differences do exist, a more in-depth examination of the WDM and CDM simulations is worthwhile. In order to make use of the entire set of hypothesis testing frameworks, a corresponding set of simulations is required. To artificially create such a simulation set, one can bootstrap the single simulation into smaller cubes evenly, as shown in Figure 11. For example, a *double split* would produce a set of size 8 equally sized simulations while a *triple split* would produce a set of size 27. The CDM and WDM datasets can then be compared topologically slice by slice, providing greater resolution in smaller scale differences as the number of slices is increased. Figure 12 displays the distribution of GKD p-values given a number of slices. The uppermost heatmap represents the raw, unsplit version of the WDM/CDM datasets, where seemingly no large-scale differences arise. However, by increasing the number of splits and focusing on smaller scale structure, one can see that while it is true that most of the large scale Universe under WDM and CDM assumptions are similar (high p-values), there exist regions that suggest large differences in topology (low p-values) that are masked

at lower resolutions.

To confirm the validity of splitting procedure, the slice with the lowest p-value (slice 34) and the slice with the highest p-value (slice 57) were taken from the quadruple split WDM and CDM datasets. Respective persistence diagrams were generated and both the bottleneck and 2-Wasserstein distance metrics were used to categorize the distance between homologies of the two persistence diagrams. As Table 2 shows, slice 34 has much higher distance measurements for the 0th and 1st homologies than slice 57 with comparable measurements for 2nd homologies, suggesting larger topological differences in the former slice. Additionally, Figure 13 shows a comparision of the CDM and WDM Euler characteristic functions, known to be a good topological summary, of slices 34 (left) and 57 (right). It is apparent that there are far greater discrepancies between the Euler characteristic for slice 34 than 57, again suggesting greater topological variance in the former.

Distance Metric	N-th Homology	Slice 34	Slice 57
Bottleneck	0	0.812508	0.170569
Bottleneck	1	0.273844	0.164631
Bottleneck	2	0.232097	0.392844
2-Wasserstein	0	1.068576	0.301319
2-Wasserstein	1	0.566911	0.360367
2-Wasserstein	2	0.207662	0.293671

Table 2: Distance measurements between slice 34 and 57 of the WDM and CDM quadruple split datasets.

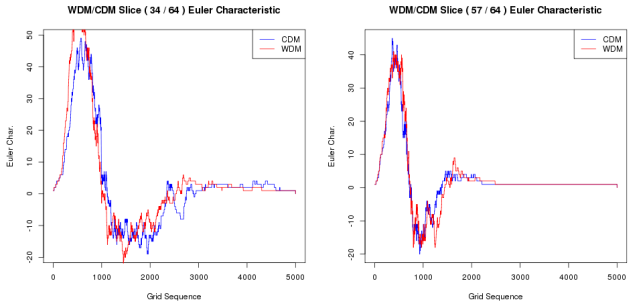

Figure 13: (left) Euler characteristic function for WDM (red) and CDM (blue) in slice 34; (right) Euler characteristic function for slice 57.

Provided that standardization acts as a way to separate topology and geometry, we repeated the same cubic slicing procedure with standardized persistence diagrams for each split. Figure 14 represents a similar heatmap created from the standardized persistence diagrams from the differently split WDM and CDM data sets where the sliced cubes are identical regardless regardless of standardization. Comparing the two heatmaps, one can notice:

(1) The standardized heatmaps are more biased towards the extreme p-values in the double and triple split, represented by darker red and blue colors;

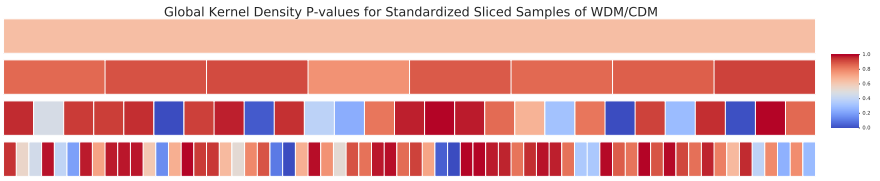

Figure 14: GKD p-values per slice for unsplit, double split, triple split, and quadruple split (top to bottom) slices of **standardized** WDM/CDM data sets.

however, in the quadruple split set, the standardized heatmap shows GKD p-values that are much more biased towards 0.5, suggesting that in smaller scale analysis, geometry plays a large role in the different homologies. See Figure 15.

(2) The slices with significant p-values are different, sometimes completely opposing, between the standardized and unstandardized heatmaps. For example, consider the sixth slice of the triple-split data sets: for the unstandardized data sets, the GKD test on this slice returned a p-value of 0.96, the highest p-value within the triple-split slices; however, for the standardized data sets, the same slice returned a p-value of $1.93e-54$, the lowest p-value within the triple-split slices. This suggests that the sixth slice of the triple-split set interestingly shares similar homologies between WDM and CDM assumptions when considering both geometry and topology, but when solely considering topology and normalizing size, the homologies differ greatly. Similarly, certain slices, like slice 13 of the quadruple-split set, have low p-values in both the standardized and unstandardized setting. In the case of slice 13, the standardized p-value is less significant (0.15 vs 0.04), suggesting that there are important differences in structure and size. Thorough analysis of Figures 14 and 12 allow us to quantify and categorize the magnitude and type of differences between WDM and CDM assumptions on the EAGLE data set.

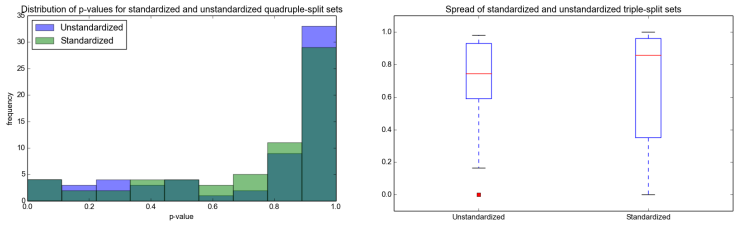

Figure 15: (left) Histogram of the p-values for standardized and unstandardized quadruple-split sets. The standardized p-values are more frequent in the middle while the unstandardized p-values are more biased towards extremes. (right) Boxplots of the spreads for standardized and unstandardized triple-split sets. The standardized p-values have more significantly small values while the unstandardized values are more left skewed.

4.2.3 Hypothesis testing

Having created datasets of multiple slices, the suite of hypothesis testing frameworks explored on Voronoi foams are applicable to the WDM and CDM data.

Tables 3, 4 shows both the standardized and unstandardized results of four categories of hypothesis testing frameworks. The unstandardized results suggest that with the exception of less powerful testing frameworks, shown through the Voronoi simulations to include all GKD tests and most Sil. tests, that the more splits there are, the more differences found in the topology. More interestingly still, a similar effect is observed through standardizing the persistence diagrams prior to hypothesis testing: all the frameworks produce higher p-values on average and are less confident in the difference between double split, triple split and quadruple split datasets. This seems to explain that like the Voronoi foams, a large degree of the topological difference between warm dark matter and cold dark matter assumptions in the large-scale universe are due to geometrical properties such as size.

Test	Double Split	Triple Split	Quadruple Split
EC	-5.937	-13.589	-25.132
$EC_{0:2}$	-4.677	-16.261	-27.132
EC_0	-7.485	-16.939	-29.419
EC_1	-4.733	-15.071	-0.0292
EC_2	-0.469	-20.086	-1.057
Sil_{EC}	-7.113	-12.046	-19.610
$Sil_{0:2}$	-5.723	-17.307	-32.953
Sil_0	-7.529	-19.152	-33.827
Sil_1	-4.932	-13.260	-22.540
Sil_2	-0.0339	-0.0515	-1.462
GC_0	-0.355	-0.860	$-\infty$
GC_1	-0.717	-1.027	$-\infty$
GC_2	-0.606	-1.495	-2.699
GKD_0	-0.254	-0.211	-0.213
GKD_1	-0.242	-0.254	-0.253
GKD_2	-0.217	-0.266	-0.222

Table 3: P-values from hypothesis tests on the **unstandardized** WDM and CDM simulations by double, triple and quadruple splits. The p-values of $-\infty$ arise from a permutation test with no positive examples.

5 Conclusions

In this paper, we presented a hypothesis testing framework, build on persistent homology, to compare topological summaries of two sets of point cloud data. We showed empirically that such a framework is able to infer differences in the true distribution of topology by comparing Voronoi tessellations with controlled hyperparameters. Additionally, we presented the application of this framework on the EAGLE data set to analyze the topology of the cosmic mass given assumptions of warm and cold dark matter, resulting in the discovering of locally significant spatial differences in geometry and topology. We believe this framework may provide a standard method for evaluating hypothesis regarding topology in a diverse array of fields that greatly improve over currently existing methods.

Test	Double Split	Triple Split	Quadruple Split
EC	-3.958	-3.369	-11.134
$EC_{0:2}$	-4.877	-12.855	-13.401
EC_0	-5.858	-5.284	-11.792
EC_1	-1.629	-0.834	-2.844
EC_2	-2.414	-5.573	-6.704
Sil_{EC}	-1.915	-0.197	-2.301
$Sil_{0:2}$	-3.367	-6.884	-11.093
Sil_0	-5.189	-4.096	-10.754
Sil_1	-1.549	-1.253	-6.409
Sil_2	-2.531	-4.132	-3.811
GC_0	-0.00612	-0.202	-0.0675
GC_1	-0.0168	0.000	-0.0195
GC_2	-0.0168	-0.007	-0.585
GKD_0	-0.226	-0.261	-0.250
GKD_1	-0.189	-0.260	-0.229
GKD_2	-0.179	-0.238	-0.196

Table 4: P-values from hypothesis tests on the **standardized** WDM and CDM simulations by double, triple and quadruple splits.

Acknowledgement

This research was supported by *list all relevant stuff.*

References

[1] Paul Bendich, JS Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *arXiv preprint arXiv:1411.6652*, 2014.

[2] Peter Bubenik. Statistical topology using persistence landscapes. 2012.

[3] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.

[4] Frédéric Chazal, Brittany T Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.

[5] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 474. ACM, 2014.

[6] Yen-Chi Chen, Daren Wang, Alessandro Rinaldo, and Larry Wasserman. Statistical analysis of persistence intensity functions. *arXiv preprint arXiv:1510.02502*, 2015.

[7] Jessi Cisewski, Rupert AC Croft, Peter E Freeman, Christopher R Genovese, Nishikanta Khandai, Melih Ozbek, and Larry Wasserman. Nonparametric 3d map of the intergalactic medium using the lyman-alpha forest. *Monthly Notices of the Royal Astronomical Society*, 440(3):2599–2609, 2014.

[8] Tarn Duong. Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics*, 25(3):635–645, 2013.

[9] Tarn Duong, Bruno Goud, and Kristine Schauer. Closed-form density-based framework for automatic detection of cellular morphology changes. *Proceedings of the National Academy of Sciences*, 109(22):8382–8387, 2012.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[11] Vincent Icke and Rien van de Weygaert. The galaxy distribution as a voronoi foam. *Quarterly Journal of the Royal Astronomical Society*, 32:85–112, 1991.

[12] Mark R Lovell, Carlos S Frenk, Vincent R Eke, Adrian Jenkins, Liang Gao, and Tom Theuns. The properties of warm dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 439(1):300–317, 2014.

[13] Joop Schaye, Robert A Crain, Richard G Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S Frenk, IG McCarthy, John C Helly, et al. The eagle project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554, 2015.

[14] Volker Springel, Simon DM White, Giuseppe Tormen, and Guinevere Kauffmann. Populating a cluster of galaxies–i. results at z= 0. *Monthly Notices of the Royal Astronomical Society*, 328(3):726–750, 2001.

[15] Rien van de Weygaert and Vincent Icke. Voronoi vertices as abell clusters.

[16] Rien Van De Weygaert, Gert Vegter, Herbert Edelsbrunner, Bernard JT Jones, Pratyush Pranav, Changbom Park, Wojciech A Hellwing, Bob Eldering, Nico Kruithof, EGP Bos, et al. Alpha, betti and the megaparsec universe: on the topology of the cosmic web. In *Transactions on Computational Science XIV*, pages 60–101. Springer-Verlag, 2011.

[17] Matteo Viel, George D Becker, James S Bolton, and Martin G Haehnelt. Warm dark matter as a solution to the small scale crisis: New constraints from high redshift lyman-α forest data. *Physical Review D*, 88(4):043502, 2013.