Topological Hypothesis Tests for the Large-Scale Structure of the Universe

Wu, Mike

Department of Computer Science, Yale University

Cisewski, Jessi*

Department of Statistics, Yale University

Fasy, Brittany T.

Department of Computer Science, Montana State University

Hellwing, Wojciech

ICG, University of Portsmouth

Lovell, Mark R.

ITF, University of Amsterdam

Rinaldo, Alessandro

Department of Statistics, Carnegie Mellon University

Wasserman, Larry

Department of Statistics, Carnegie Mellon University

October 11, 2016

*Corresponding author. The authors gratefully acknowledge Yale Information Technology Services **???any grants to add???**

Abstract

The large-scale structure (LSS) of the Universe is an intricate and spatially complex web. In order to understand the physics of the Universe, theoretical and computational cosmologists develop large-scale simulations that allow for visualizing and analyzing the LSS under varying physical assumptions. In particular, different realizations of dark matter, warm and cold, are thought to lead to contrasting velocities of cosmic structure formation. However, rigorous comparisons and inference on such complicated structures can be problematic. We present a framework for hypothesis testing of LSS using persistent homology. The randomness in the data (due to measurement error or topological noise) is transferred to randomness in the topological summaries, which provides an infrastructure for inference. These tests allow for statistical comparisons between complicated spatial data such as LSS in cosmology, but are also present in other areas of science. We present several possible test statistics using persistence diagrams, carry-out a simulation study to investigate the suitableness of the proposed test statistics, and finally we apply the proposed inference framework to study the topological disparities between assumptions of warm and cold dark matter.

Keywords: persistent homology, voronoi, intensity, euler, silhouette, dark matter

1 Introduction

Rigorous comparisons of spatially complex web-like data such as the large-scale structure (LSS) of the Universe (see Figure 1) 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 (TDA) 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 for some problems in astronomy (Sousbie 2011, Sousbie et al. 2011, Van De Weygaert et al. 2011, Cisewski et al. 2014) among other areas of science (Bendich et al. 2014, Duong et al. 2012).

The large-scale structure (LSS) of the Universe is an important example of a spatially complex structure, and is fittingly referred to as the *Cosmic Web* (Bond et al. (1996), Springel et al. (2006)). The LSS of the Universe is a focus of manifold scientific research because its properties reveal information about the underlying physics and formation of our Universe (Davis et al. (1985)). In order to study theoretical aspects of the formation and evolution of LSS, cosmologists develop large-scale simulations and can adjust the physical inputs and evaluate their effects on the LSS (Cooray & Sheth 2002, Centrella & Melott 1983, Doroshkevich et al. 1980, Schaye et al. 2015). One such input is related to the nature of dark matter (DM). The received position is that the Universe is made up of dark energy, DM, and baryonic matter. The nature of DM is still a mystery, but there are hypotheses regarding its possible particle behavior. Hot DM consists of particles that travel with ultrarelativistic speeds, while cold DM particles move much slower. For an easy introduction to DM, see (Hilbe et al. 2014, p. 61-63).

Though the generally accepted and best supported cosmological model assumes *cold dark matter* (known as ΛCDM), there are some elements of disagreement with observations (Schneider et al. (2012)). Furthermore, it has been demonstrated through cosmological simulations that the nature of DM affects the development and formation of LSS (Schneider et al. (2012)). In Figure 1, two realizations – under the assumption of cold dark matter and

warm dark matter – from the EAGLE cosmological simulation (Schaye et al. (2015)) are displayed. Though there are similarities in shape of the densest regions (called *filaments*), there are differences in the distribution of matter about the filaments.

discuss simulation

(a) Eagle CDM Simulation

(b) Eagle WDM Simulation

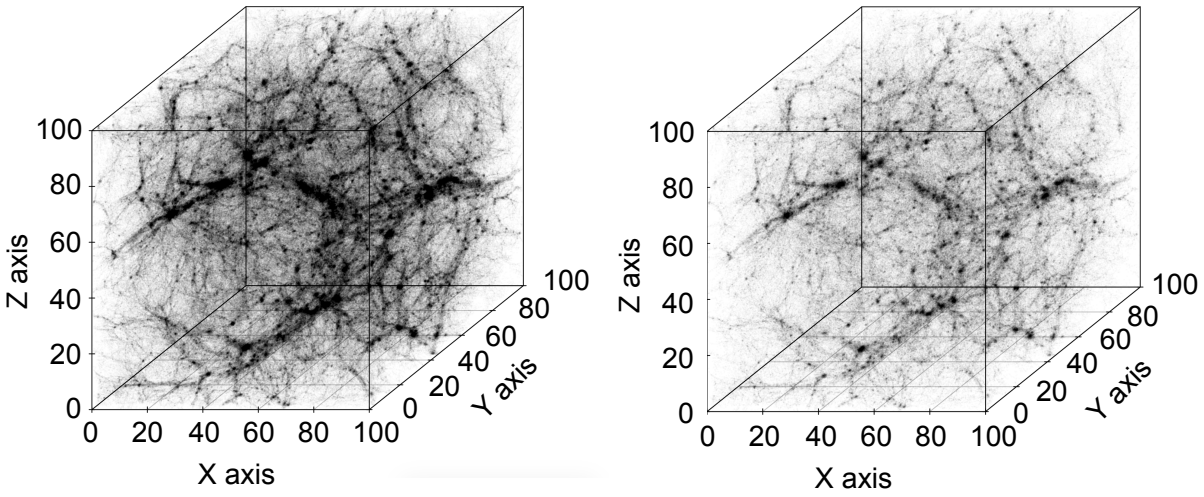


Figure 1: EAGLE simulation under assumptions of (a) cold and (b) warm dark matter Schaye et al. (2015). Note the similaries in structure but the disparaties in density.

The goal of the proposed topological hypothesis tests is to detect differences in the topology between these types of structures using persistent homology. In this paper, we move towards quantitatively detailing the topological and geometric differences between the LSS under respective assumptions of cold and warm dark matter. We begin with an introduction to persistent homology followed by the proposed hypothesis testing framework. Then we carry-out a simulation study to investigate the performance of the proposed test statistics followed by more background on LSS, and then apply the hypothesis tests to quantify the topological disparities between warm and cold DM assumptions. We end with concluding remarks.

2 Tools from TDA Useful for Studying LSS

Homology is the study of certain properties of topological spaces, specifically the number of different ordered holes in the space (e.g. connected components, loops, voids). Persistent homology studies the spacial structure of a parameterized family of topological spaces (e.g., keeping track of the so-called births and deaths of homological features as a topological space changes with the parameter). The type of data that we are investigating is a point cloud, where each point can represent a galaxy or, for cosmological simulation data, a certain mass of DM. With cosmological simulations, we look at cubic regions representing some part of our Universe and analyze the distribution of matter within that cube. We may define a simplicial complex directly on the point cloud, or we may compute a smoothed version of the data using kernel density estimation (KDE). The homological features mentioned above have cosmological interpretations in dimensions zero, one, and two. We discuss this briefly before going further into the algebraic topology.

Clusters A *connected component*, or zeroth-dimensional homology feature (H_0), is a maximal subspace of a topological space that cannot be covered by two disjoint open sets. In words, a connected component is a 'piece' of a topological space. If our topological space is a k-nn graph, then the components are clusters of data points. In cosmology, clusters of galaxies (or other cosmological matter) are an important structure to understand. Persistent homology tracks the appearance of new connected components and the merging of two distinct components into one.

Filaments and Loops A *loop*, or one-dimensional homology feature (H_1), provides information about the connectivity of data. As many H_0 features appear, nearby connected components can merge together. If our topological space is a k-nn graph then the loops are clusters of data points that merge into a fully connected cycle. For LSS, this would appear as filaments joining together in a loop.

Cosmological Voids A *void*, or two-dimensional homology feature (H_2), represents empty areas within the topological space. If our topological space is a k-nn graph, then the voids are the unfilled spaces inside enclosed H_1 features. In cosmology, to fully appreciate

the topology, it is important to understand the high-density regions (connected components and loops) but also where matter is scarce.

2.1 Persistent Homology

Various methods can be used in order to transform a discrete point set into a topological space. For example, points can be connected based on a distance (or a distance-like structure as in Chazal et al. (2011)), or one may estimate the density from which the points were sampled. In the latter case, one can look at a KDE of a point cloud and study the topological features of super-level sets of that density. Below, we summarize some of the key components of persistent homology. See Edelsbrunner & Harer (2010), Hatcher (2002), Munkres (1984) for a more thorough introduction to algebraic and computational topology. Brittany says: check citations

Filtrations 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 $\{(x, y, z) \in p \text{ s.t. } z < r\}$. If r is initialized at ∞, the super-level set is empty and the sub-level set contains all of p. The evolving topological space is characterized by its homology as r decreases to $-\infty$. 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 empty and is instead, composed of disjoint peaks/local maxima. An example of a density with a 2-dimensional domain is presented in Figure 2a, along with the plane representing a threshold for defining super-level sets. Figures 2b and 2c display the upper-level sets for two thresholds. Figure 2d displays the corresponding persistence diagram which is discussed below.

Tracking Homology Generators As the threshold, r, decreases from ∞ to 0, the homology of the upper-level sets change. Between the thresholds corresponding to Figure 2b and Figure 2c, the upper-level set changes from having six connected components (H_0's) and zero loops (H_1's) to having one connected component and one loop. The time in the

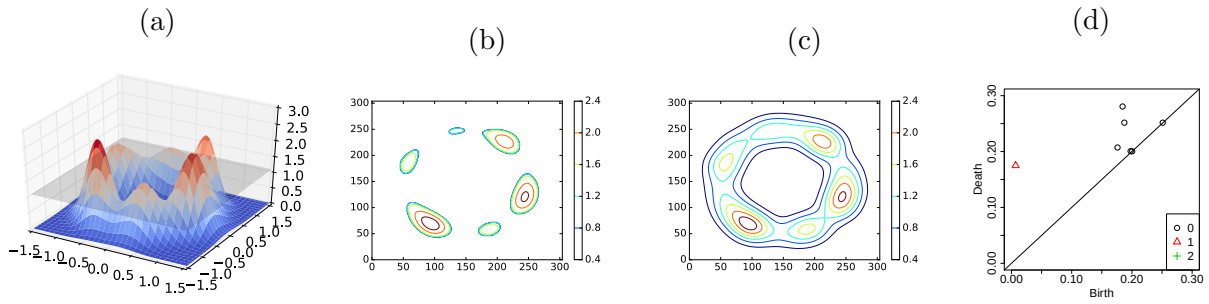


Figure 2: We illustrate persistent homology with a two-dimensional example. The density in (a) shows 3 steep peaks and 3 shallow peaks distributed around a uniform circle. In light gray, we see the hyperplane, shown (b) as $t = 1.0$. (b) and (c) plot the super-level sets $\widehat{p}^{-1}[1.10, \infty)$ and $\widehat{p}^{-1}[0.10, \infty)$ respectively. Notice as the threshold is decreased, the contour more clearly defines a loop-like structure. In (d), we see the persistence diagram for the super-level set filtration of $\widehat{p}$. We highlight the upper left quadrant based at $(0.18, 0.02)$, which contains a one-dimensional persistence point. This corresponds to the loop shown in (c). The remaining 0-dimensional homologies represent the connected components from each of 6 peaks.

filtration when homology features appear, the *birth* of the feature, and the time when a feature joins other features, the *death* of the feature, is captured in a persistence diagram. Figure 2d displays the persistence diagram for the function in Figure 2a, where each point represents the death time (x-axis) and birth time (y-axis) of a homological feature. A diagonal point (x, x) represents a feature that was born and died at the same time x.

The *persistence* of a persistence point (d, b) is the length of the interval of the persistence parameter that support that feature: $b - d$. In the persistence diagram, the distance from (b, d) to the diagonal is proportional to this value; in fact, the (Euclidean) distance to the diagonal is $(b-d)/\sqrt{2}$. Sometimes the persistence of a feature is indicative of the significance of the feature, which means that points close to the diagonal are indistinguishable from noise.

2.2 Derivatives of Persistence Diagrams

While persistence diagrams can summarize the topology of a data set, it is not straightforward to compare two different persistence diagrams. Often distances like the bottleneck distance or the q-Wasserstein distance are used in this setting, but are computationally expensive. Below are several examples of methods to further summarize a persistence diagram that are used to develop hypothesis tests.

Landscapes and Silhouettes Weighted silhouette functions are formed by weighting a particular functional summary of persistence diagrams called *landscape functions* (Bubenik (2015)). More details and theoretical properties of landscapes and silhouettes are provided in Chazal et al. (2014).

Landscape functions are defined as follows. 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 $t \in [t_{\min}, t_{\max}]$. For a given h, the persistence landscape is then defined as the following collection of functions

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

where kmax is the kth largest value in D_h. An example of a landscape function is displayed in Figure 3.

Rather than working with each k of $\lambda_{D_h}(k, t)$ individually, silhouettes provide a way of combining the information in the collection of landscape functions. Silhouettes are weighted

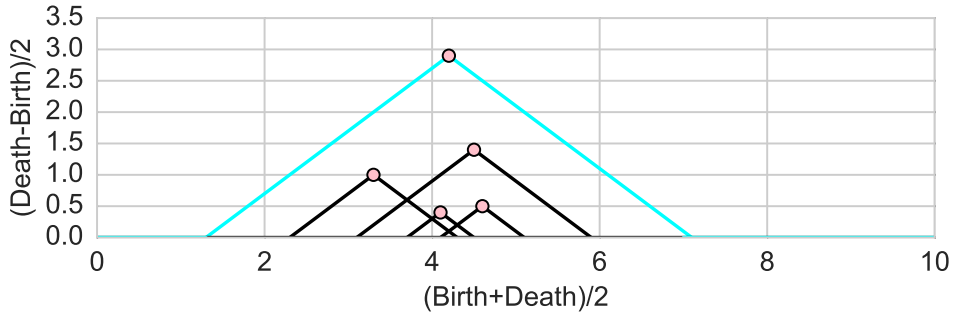


Figure 3: The shaded circles in pink represent the point coordinates in a persistence diagram D. The landscape $\lambda(k, \cdot)$ is the k-th largest of the arrangement of the graphs of $\{\Lambda_p\}$. In particular, the cyan curve is the landscape $\lambda(1, \cdot)$.

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. As suggested in Chazal et al. (2014), we use $w_{hi} = |d_{hi} - b_{hi}|^p$, where p is a tuning parameter that needs to be selected.

Euler Characteristic Function The Euler characteristic is a topological invariant and defined as: $\chi = \beta_0 - \beta_1 + \beta_2 - \beta_3 + ... \pm \beta_N = \sum_{i=0}^{N} (-1)^i \beta_i$, where β_i represents the i-th Betti number (the rank of the i-th homology group) and N is the number of dimensions. When analyzing persistence diagrams of LSS, since there exist only three dimensions of data, the only non-trivial homology groups will be in dimensions 0, 1, and 2. Given the Betti numbers β_0, β_1, and β_2, the Euler characteristic equation simplifies to: $\chi = \beta_0 - \beta_1 + \beta_2$. The topological space is parameterized by the threshold, t (ranging from $t_{\min}$ to $t_{\max}$), defining the upper-level sets. The Euler Characteristic function, $\chi(t)$, captures the Euler characteristic for each threshold t.

3 Methods

Provided data as a point cloud, persistent homology can be used to summarize its topological information into a persistence diagram. Next, we describe several frameworks for

using persistence diagrams in two-sample hypothesis tests in search of a way to compare and test differences between sets of data in topological structure.

Suppose we have two sets of persistence diagrams, $\{\mathcal{P}_1^{(1)}, \ldots, \mathcal{P}_n^{(1)}\}$ and $\{\mathcal{P}_1^{(2)}, \ldots, \mathcal{P}_m^{(2)}\}$. These samples can be used to test $H_0 : \mathcal{P}^{(1)} = \mathcal{P}^{(2)}$ vs. $H_2 : \mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$, where $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are the true underlying distributions of persistence diagrams for group 1 and 2, respectively. We would like the framework to test the hypothesis that the two samples are drawn from a population with the same random topology. However, we note that without incorporating a scaling adjustment on the space of the data or the space of the diagrams, geometrical differences can also lead to a rejection of the null hypothesis. This is discussed in more detail in Section 4.2.1.

Given two samples of persistence diagrams, there are a number of possible ways to derive test statistics. We select four possible approaches with nine individual tests. The first approach is based on functional summaries derived from the sampled persistence diagrams: Euler characteristic function (EC), Silhouette function (Sil), and a Silhouette-Euler characteristic function (SilEC). Given that each observed dataset will have a corresponding function, a p-value can be derived from a two-sample T-test based on the integral of the absolute value of the functional summaries. The next two approaches use variations on smoothed persistence diagrams called *intensity functions* Chen et al. (2015) with p-values derived from a two-sample kernel test Gretton et al. (2012) and an asymptotic argument through permutation (KC, WKC, PI). Additionally, we consider a test using the two-point correlation function (CORR) in order to have a comparison with a summary capturing the spatial behavior of the LSS. Similar to the functional summaries above, p-values for the CORR test will be carried out with a T-test based on the integral of the absolute value of the correlation function. The proposed test statistics are discussed in more detail below.

Among the variations to test statistics considered (EC, Sil, SilEC, KC, WKC, PI, and CORR), we are seeking the summary that best captures differences in distributions of persistence diagrams produced from structures like the LSS of the observable Universe.

Euler Characteristic Test (EC) To use the Euler characteristic (EC) function, $\chi(t)$, in a hypothesis testing framework, the absolute value of the EC function is integrated with respect to the threshold t to produce an Euler test statistic, $\widehat{EC}$:

$$\widehat{EC} = \int_{t_{\min}}^{t_{\max}} |\chi(t)| \, \mathrm{d}t.$$

Given the Euler statistics for two sets of persistence diagrams (where a set contains persistence diagrams drawn from the same topological distribution), a T-test is used to calculate a p-value for the null hypothesis that the two sets of persistence diagrams are sampled from the same distribution.

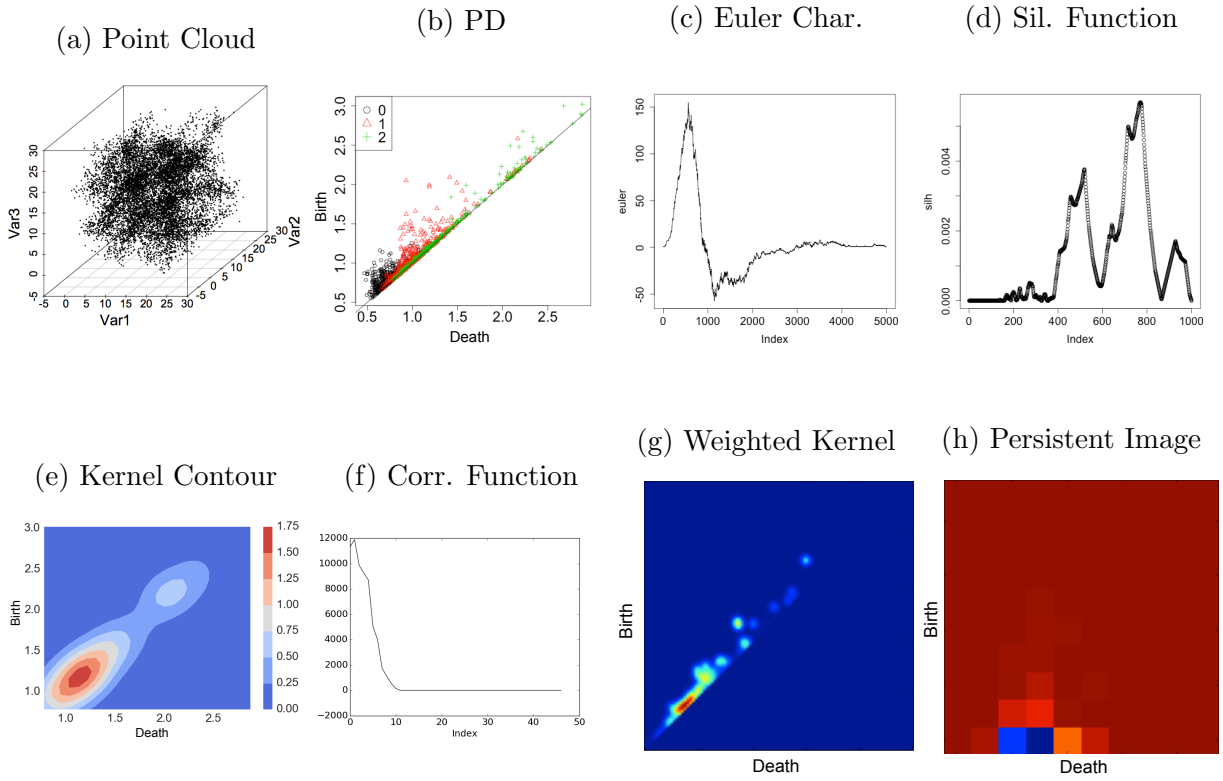


Figure 4: Examples of test statistics used in the hypothesis tests. Figure 4a shows the point cloud, and Figure 4b shows the persistence diagram used to generate the functional summaries shown in Figure 4b-Figure 4h.

Silhouette Test (Sil) As with the Euler Characteristic (EC) function, a weighted silhouette summarizes a persistence diagram through a continuous function. Given a set of persistence diagrams (each with N homology groups), each diagram is summarized by N weighted silhouettes. For each silhouette, we can derive a statistic $\widehat{Sil}$ using the area under

the respective weighted silhouette, resulting in N sets of statistics. In this manner, each dimensional homology of a persistence diagram has its own test statistic, $\widehat{Sil}_h$, where h represents a dimensionality of the homology group. This is preferred since the silhouette molded from connected components may be very different from the silhouette created from loops. We consider T-tests on each dimension ($\widehat{Sil}_0$, $\widehat{Sil}_1$, $\widehat{Sil}_2$), along with considering all dimensions in parallel ($\widehat{Sil}_{0:2}$) using a Hotelling T^2 test.

Silhouette-Euler Characteristic (SilEC). Another method for considering all the individual silhouettes, $S_h(t)$, across dimensions $h = 0$, 1, and 2, simultaneously is a modified Euler characteristic function. Instead of calculating the alternating sum of Betti numbers, the Silhouette-Euler characteristic function (SilEC) is defined as the alternating sum of the individual silhouette functions across the threshold parameter, t. The statistic derived, $\widehat{SilEC}$, is the integral of the absolute value of the summed function,

$$\widehat{SilEC} = \int_{t_{\min}}^{t_{\max}} |S_1(t) - S_2(t) + S_3(t)| \, \mathrm{d}t$$

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

Permutation Method. Although not one of the nine approaches, the permutation method is frequently used in our tests and provides a very generic approach for calculating a p-value from any arbitrary test statistic. Assuming we observe two independent samples $X_1, \cdots, X_n \sim P$ and $Y_1, \cdots, Y_n \sim Q$, the two sample test problem is to test the hypothesis $H_0 : P = Q$ versus the alternative $H_1 : P \neq Q$. Assuming there is some arbitrary method of calculating a test statistic T as a function of the data, we reject H_0 if $T > t$ where t is a critical value. We choose t such that if H_0 is true then $W(T > t) \leq \alpha$ where W is the distribution of T when H_0 is true. Permutation testing attempts to help us choose t and find W of T under H_0.

In the permutation method, the data is concatenated as a vector in the order

$$X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_n$$

and presented initial labels of 0 or 1 where all X_i receive label 0 and all Y_i receive label 1. If H_0 is true, then the entire data vector is an i.i.d sample from P and the group labels

are arbitrary. To test this, the group labels are randomly permuted and the test statistic is recalculated. This changes the values of T but (under H_0), it should not change the distribution of T. The labels are permuted N times and the p-value is

$$p = \frac{1}{N} \sum_{j=1}^{n} I(T_j \geq T)$$

where I is the indicator function. Therefore, the p-value is the fraction of times T_j is larger than T. As N approaches inf, p approaches the exact value.

Kernel Contour Test (KC). Rather than working with the raw persistence diagrams, the Kernel Contour Test (KC) uses a test statistic derived from a kernel two-sample test on the smoothed persistence diagram called the *intensity function* Chen et al. (2015). The kernel two-sample test was first introduced for analyzing and comparing distributions with a maximum mean discrepancy (MMD) statistic Gretton et al. (2012). The KC 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, $\widehat{KC}$, is defined as

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

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 $||X-Y|| = \int \left(X(t_1, t_2) - Y(t_1, t_2)\right)^2 dt_1 dt_2$ where $X(t_1, t_2)$ and $Y(t_1, t_2)$ are two intensity functions. The h is a hyperparameter that sets the standard deviation of the Gaussian distribution used in the kernel 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 0.1 ± 0.04 using grid search from 0 to 5. A permutation test is used to calculate a p-value from the statistic.

Weighted Kernel Test (WKC) Two disadvantages of the Kernel Contour Test (KC) shown above are (1) the tendency of the test to heavily weight dense clusters near the diagonal rather than sparse persistent features with longer lifespans, and (2) the possibility of the smoothed intensity function to "spill over" the diagonal as a product of the Gaussian kernel. The Weighted Kernel Test (Chen et al. (2015)) addresses both concerns with an adjusted kernel function.

Let b and d denote the birth and death of a persistent feature, such that

$$\mathcal{D} = (b_j, d_j) : j = 1, \cdots, K$$

represents a persistence diagram where K is the cardinality of persistent features. Given $\mathcal{D}$, we can estimate its intensity, κ_τ by

$$\hat{\kappa}_\tau(x, y) = \sum_j (d_j - b_j) \frac{1}{\tau^2} K(\frac{x - b_j}{\tau}) K(\frac{y - d_j}{\tau})$$

such that the sum is over all persistent features, K is a symmetric kernel function (such as 1D Gaussian where the mean and standard deviation are estimated by the sample), and τ is a smoothing parameter. This statistic is calculated over a two-dimensional grid

$$\{(x, y) \text{ s.t. } b_{min} \leq x \leq b_{max} \wedge d_{min} \leq y \leq d_{max}\}$$

where $b_{min} = \min(\{b_j\}_{j=1}^K)$, $d_{min} = \min(\{d_j\}_{j=1}^K)$, $b_{max} = \max(\{b_j\}_{j=1}^K)$, and $d_{max} = \max(\{d_j\}_{j=1}^K)$. A step size h_x between adjacent pairs (x_i, y_j), (x_{i+1}, y_j), and h_y between pairs (x_i, y_j), (x_i, y_{j+1}) control the size of the grid ($h_x = h_y = 0.01$ used in practice). This Weighted Kernel Contour statistic (Chen et al. (2015)) uses the difference between the birth and death of each persistent feature as the weight for that feature. Therefore, features near the diagonal will have very little effect, reducing spill over, and sparse features with long lifespans (generally more interesting topologically) will contribute more.

To calculate the p-value, the grid $\{x, y\}$ is flattened into a vector where each index represents an abstracted topological feature, and as with the KC test, an asymptotic argument (permutation test) is used. The WKC test is performed independently for each homology, resulting in three tests: WKC(0), WKC(1), and WKC(2).

Persistent Image Test (PI) The Persistent Image test Adams et al. (2015) is a variation of the WKC test with an alternative kernel function and sub-sampling. Let $\mathcal{D} = (b_j, d_j) : j = 1, \cdots, K$ be a persistence diagram where K is the number of persistent features. Unlike WKC, the PI test first transposes the persistent diagram using the linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ where $T(x, y) = (x, y - x)$. Therefore, let $T(\mathcal{D})$ represent the transformed diagram with the new birth-persistence coordinates. Let $K_u : \mathbb{R}^2 \to \mathbb{R}$ be a differentiable probability distribution with mean $u = (u_x, u_y)$. In our applications, we choose this distribution to be the normalized symmetric 2D Gaussian with mean u and variance σ^2.

$$K_u(x, y) = \frac{1}{2\pi\sigma^2} e^{-[(x-u_x)^2+(y-u_y)^2]/2\sigma^2}$$

Finally, let $f : \mathbb{R}^2 \to \mathbb{R}$ be a non-negative weighting function that is zero along the horizontal axis, continuous and differentiable. f is chosen to only depend on the rotated vertical persistence coordinate y, and like WKC, weight points of higher persistence (lifespan) more heavily. In our applications, we use a piecewise linear weighting function that was found to be stable. Given $b > 0$, define f_b as

$$f_b(t) =\to \begin{cases} 0 & \text{if } t \leq 0 \\ \frac{t}{b} & \text{if } 0 < t < b, \text{ and} \\ 1 & \text{if } t \geq b \end{cases}$$

Given f, K, and $T(\mathcal{D})$, we can define a persistence surface (Adams et al. (2015)), $\rho_{\mathcal{D}} : \mathbb{R}^2 \to \mathbb{R}$ as

$$\rho_{\mathcal{D}}(x, y) = \sum_{u \in T(\mathcal{D})} f(u) K_u(x, y)$$

Like the WKC test, ρ is calculated over the two-dimensional grid

$$\{(x, y) \text{ s.t. } b_{min} \leq x \leq b_{max} \wedge d_{min} \leq y \leq d_{max}\}$$

. In our applications, the step sizes h_x and h_y were chosen to make the grid size 30 by 30. Given $\rho_{\mathcal{D}}$, a persistence image (Adams et al. (2015)) $I(\rho_{\mathcal{D}})_p$ is defined as a grid of pixels calculated from convolving the persistent surface with a uniform matrix of size n by n of

1's. In order words, $I(\rho_{\mathcal{D}})_p = \int \int_p \rho_{\mathcal{D}} \, \mathrm{d}y\mathrm{d}x$. This matrix is then flattened into a 1D vector. The accuracy of the p-value using this PI framework is found to be robust to the choice of the convolution filter size. In our applications, a filter size of 3 is used, resulting in a 10 by 10 matrix, or 100 length vector when flattened.

Unlike the WKC test, the PI test combines all homologies into a single statistic. Suppose the homologies $H_0, \cdots, H_k$ are computed. One can concatenate the PI vectors for $H_0, \cdots, H_k$ into a single vector representing all dimensions simultaneously. Like the WKC test, the p-value is calculated using the permutation method, but only once for all 3 homologies.

Two-point Correlation Function Test (CORR) Jessi: revise section, explain why to consider this The Two Point Correlation Test Landy & Szalay (1993), unlike previous methods, quantitatively measures large scale structure through tracing the amplitude of clustering as a function of scale, directly summarizing the point cloud instead of a persistence diagram. Such a trace is determined by the correlation function, $\xi(r)$. $\xi(r)$ is defined as the measure of the excess probability $\mathrm{d}P$, above the expectation for an unclustered random Poisson distribution of finding a cluster in a given volume element $\mathrm{d}V$ at a separation radius r from another cluster such that

$$\mathrm{d}P = n[1 + \xi(r)] \, \mathrm{d}V$$

where n is the mean number density of the sample dataset.

$\xi(r)$ is measured by counting pairs of clusters as a function of the separation radius compared to the count for an unclustered distribution. To do this, one must construct a *random catalog* that has similar three dimensional coverage as the data but is populated with randomly distributed points. Define DD, DR, RR as counts of pairs of clusters (in bins of varying separation radii) in the data only, between the data and the random catalog, and in the random catalog only; let n_D, n_R respectively define the mean number of densities of clusters in the data and random catalogs. The correlation function, defined by Landy & Szalay (1993), is:

$$\xi = \frac{1}{RR}\left[DD\left(\frac{n_R}{n_D}\right)^2 - 2DR\left(\frac{n_R}{n_D}\right) + RR\right]$$

Provided a point cloud, clusters are derived and compared to a random catalog that is generated within the same dimensional space. Unlike other statistics, because $\widehat{CORR}$ is not dependent on persistence diagrams, only a parallel test is applicable in which clustering occurs along all three dimensions. A correlation function, ξ can then be calculated. The test statistic, $\widehat{CORR}$ is defined by the area under the absolute value of ξ.

$$\widehat{CORR} = \int_r |\xi(r)| \ \mathrm{d}r$$

To generate the random catalog $|\mathcal{P}|$ points were drawn from a Uniform distribution $\mathrm{U}(0, max(\mathcal{P}))$ where $\mathcal{P}$ is a point cloud. If $\mathcal{P}$ was standardized, points were sampled from $\mathrm{U}(0, 1)$.

4 Simulation Study

To evaluate the performance of the proposed test statistics for the two-sample hypothesis tests, we carried out a simulation study by generating realizations of web-like spatial structures. The simulation model is discussed in detail below.

4.1 Simulation model

Motivated by LSS, we developed our simulation model to approximate the Cosmic Web, though these structures appear in other areas of science as well. In particular, we drew from ideas that use Voronoi tesselations to model the filament structure of the Universe, known as *Voronoi Foam* (Icke & van de Weygaert (1987, 1991), Van De Weygaert (2007)). The Voronoi Foam model offers an approximation to the distribution of matter in the Universe at large scales (e.g. galactic clusters, filaments, walls), but not scales (e.g. small groups of galaxies) (Icke & van de Weygaert (1991)).

The cells of the Voronoi tesselation become the cosmological voids, the outline of the cells are the filaments and walls, and the points of intersection are the superclusters (large clusters of galaxies). Once the tesselation is defined, points are added according to several parameters - the points can represent individual galaxies, clusters of galaxies, or dark matter halos (which would host gravitationally-bound galaxies or galactic clusters). The

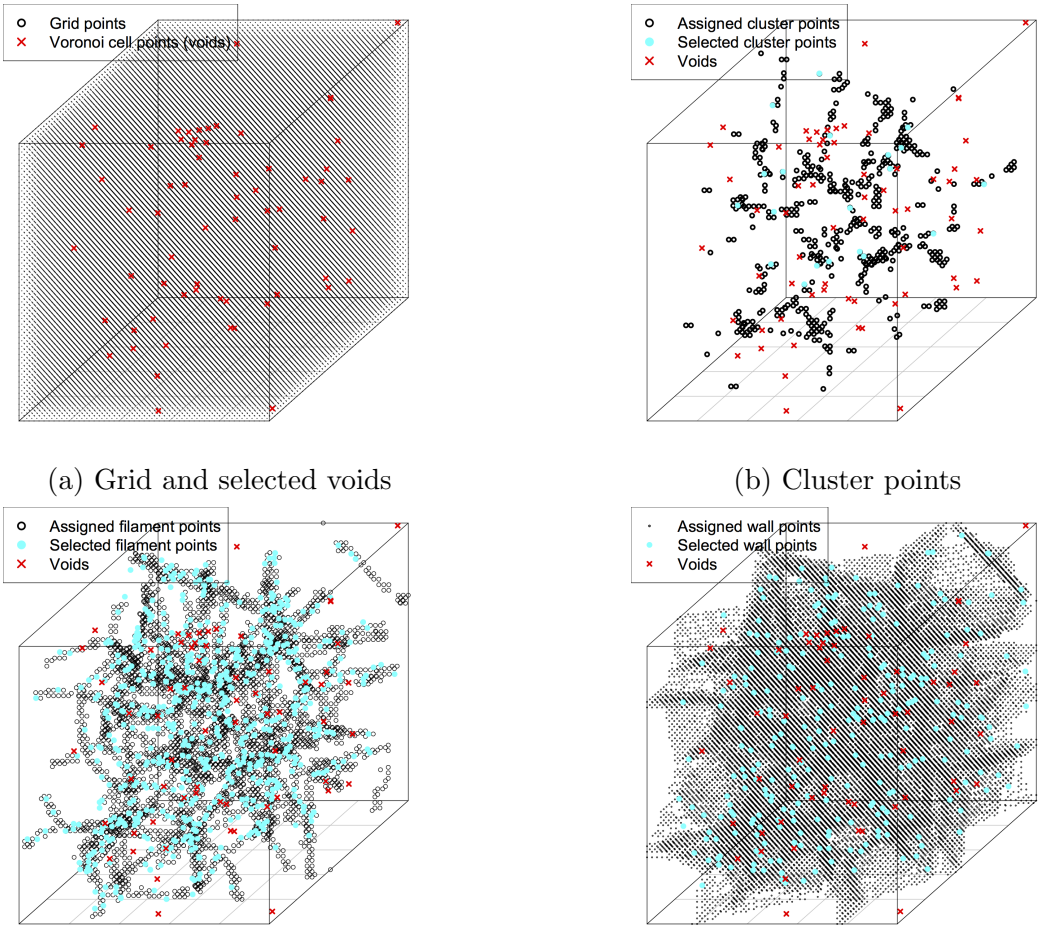
elements of our approximate Voronoi Foam model include (i) the number of voids (the number of cells in the Voronoi tesselation), (ii) the number of galaxies/clusters/halos (the number of points to generate), and (iii) the percentage of the points that should fall on the cluster, filaments, and walls, see Table 1. In this simulation study, we varied the filament percentage (percFil) from 0.1 to 0.3 by a 0.05 step size. See the Appendix for additional tests for filament percentages from 0.1 to 0.9.

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 in the clusters	0.02

Table 1: Parameters of LSS model. For the simulation study, p_f will vary from 0.1 to 0.9.

Figure 5 displays the construction procedure of one realization of our simulation model: (i) First a grid is defined at a specified resolution within a specified volume; (ii) then a specified number of points are randomly selected within the volume - these will be used to define the Voronoi tesselation and will act as voids (these will be called *void points*; (iii) the nearest void point to each grid point is found and stored, call this the *void label* of a grid point; (iv) the void labels of the eight nearest neighbors of each grid point is noted - if there are more than three unique void labels among the eight then that grid point is assigned to be a cluster point, if there are exactly three unique void labels among the eight nearest neighbors then that grid point is assigned to be a filament point, and if there are exactly two unique void labels among the eight nearest neighbors then that grid point is assigned to be a wall point. The black, empty circles in Figures 5b, 5c, 5d display the grid points that were selected to be cluster points, filament points, and wall points, respectively. Depending on the parameter assignments in Table 1 and the total desired sample size of dataset, the number of points are randomly selected among the cluster, filament, and wall points. Specified Gaussian noise is also added to the selected points so they do not fall exactly on the defined grid.

Examples of three separate Voronoi foam models of percFil 0.1, 0.5, 0.9 are displayed



(a) Grid and selected voids

(b) Cluster points

(c) Filament points

(d) Wall points

Figure 5: Simulation model construction. (a) A grid is defined and points are randomly selected to define the Voronoi tesselation - the Voronoi cells are the voids. (b) - (d) Based on the location of the voids and the grid, points are defined to be cluster points, filament points or wall points - these are the empty black circles. Based on the values assigned from Table 1, a number of points are randomly selected from the assigned points to be in the dataset.

in Figure 6 along with their corresponding persistence diagrams. One can see that the as the percFil (the percentage of points that are part of filaments) increases, the web-like structure becomes more pronounced, changing the distribution of topological features.

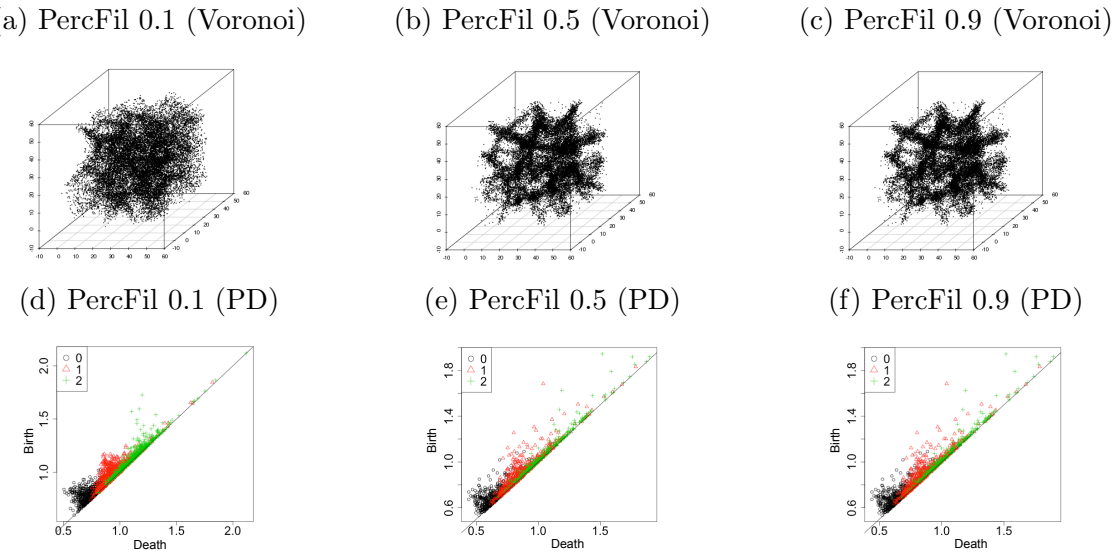


Figure 6: (a) PercFil 0.1; (b) PercFil 0.5; (c) PercFil 0.9; (d-f) Corresponding persistence diagrams (PD) for point clouds (a-c). All other parameters are defined based on Table 1.

4.2 Simulation Study Results

The simulation models used in this paper were generated under 1.25×10^5 box volume, 0.1 resolution, 1×10^4 points, 64 cells (voids), 0.02 percClust, and $[0.1, 0.3]$ percFil and $[0.68, 0.88]$ percWall. Persistent diagrams are generated using distance-to-measure (DTM) with a 0.01 tuning parameter. The 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 ∞ (with exception to the Euler characteristic function in which the artifact is preserved). The hypothesis tests were performed on 100 independent iterations of 15 independent realizations from each of the two populations. The 15 independent datasets are each generated using a percFil setting from 0.1 to 0.3 (0.05 step size); we also include a control model per iteration with a percFil of 0.1. (Hence 5 sets of 15 datasets, repeated 100 times.) Using the proposed hypothesis tests, each of the variable models will be compared to the control. More specifically, given the data drawn from a model with percFil p, each of the proposed test statistics are used to compute a p-value for the test $H_0 : \mathcal{P}^{(1)} = \mathcal{P}^{(2)}$ vs. $H_1 : \mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$, based on two samples of persistence diagrams, $\{\mathcal{P}_1^{(1)}, \ldots, \mathcal{P}_{15}^{(1)}\}$ drawn from the model with percFil = 0.1, and $\{\mathcal{P}_1^{(2)}, \ldots, \mathcal{P}_{15}^{(2)}\}$ drawn from the model with percFil $= p$, $p = 0.1, 0.15, \ldots, 0.3$. (Recall that $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are

the true underlying distributions of persistence diagrams for group 1 and 2, respectively.) Similar tests were also completed against a control model with 0.3 percFil, and similar results were found. The simulation study results are displayed in Figure 13, which shows the median p-values along with a 95% confidence interval across the 100 iterations of the following test statistics: (i) Euler-based tests, (ii) Silhouette-based tests, (iii) Kernel-based tests, (iv) Weighted Kernel-based tests, and (v) Correlation-based tests.

From Figure 7, we see that the EC test, the CORR test, and the SilEC test are the most effective (in that order) in distinguishing differences between the Voronoi simulations when the percent filaments between the control and the test differ only slightly. All other tests are not as sensitive, eventually finding differences between the test and control once percfil is around 0.7 (see Appendix). Additionally, all of the tests derived from Euler characteristics perform relatively well compared to the other tests, suggesting that the Betti numbers, by being topologically invariant, are much better functional summaries of the persistence diagrams than intensity functions, silhouettes, and landscapes. More interestingly, it is possible that the alternating linear function by which the Betti numbers are combined may better preserve topological information given that the SilEC test, which combines silhouettes through a similar linear fashion, performed better than any individual silhouette counterpart and the naive cross-dimension combination (Sil (0:2)). Finally, the CORR test, being independent of persistence diagrams, acts as an powerful alternative to the EC test, being nearly as sensitive as percfil increases.

4.2.1 Standardization of persistence diagrams

Since the goal is to discern topological differences between two samples, issues with the proposed method is that scaling differences can result in rejection when, in fact there are not statistically significant topological differences. For example, suppose we are considering two datasets - one has points randomly sampled from the perimeter of a circle with a radius of 1 and the other set has points randomly sampled from the perimeter of a circle with a radius of 10. Depending on one's goals, it may or may not be desirable to conclude that the two datasets come from different persistence diagram generators (i.e. conclude

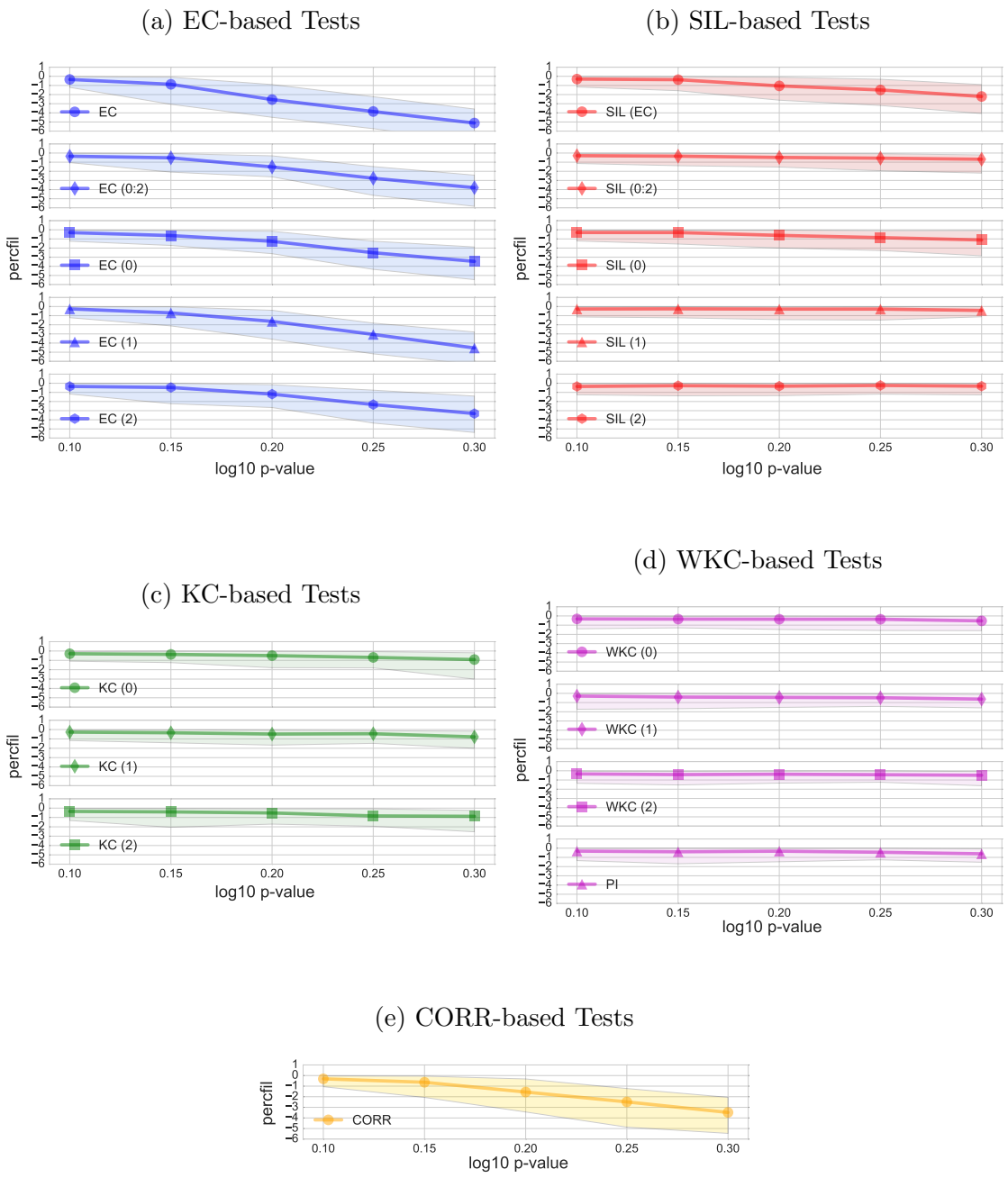


Figure 7: $\log_{10}$ p-values for the proposed test statistics in the simulation study: (a) EC-based, (b) SIL-based, (c) Kernel-based (KC), (d) Weighted Kernel-based (WKC), (e) CORR-based. X-axis represents the percent filament of the set being compared to the baseline set (0.1 to 0.3 by 0.05 increments); y-axis shows the p-values in $1\log_{10}$ space. Each plot contains a line representing the median $\log_{10}$ p-value of the 100 iterations, while the shaded region represents a 97.5 confidence interval.

$\mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$) since the inference would be based on geometrical (scaling) differences rather than topological differences. If the desire is to focus on topological differences (and remove the geometrical differences), we propose a possible preprocessing step. Specifically, we standardize the persistence diagrams so that all the homological features are re-scaled to $[0,1] \times [0,1]$. This simple standardization takes the persistence diagram window and shrinks it or expands it to fill the $[0,1] \times [0,1]$ window, maintaining the same relationship among all the homological features. If there is concern about outliers, then other quantiles (rather than the minimum and maximum) could be used for the standardization. The exception is with the Correlation test, in which the point cloud itself is standardized to $[0,1] \times [0,1] \times [0,1]$.

We repeated the simulation study from Section 4.2 except preprocessing with the standardization step; the results are displayed in Figure 14. Standardization had an appreciable impact on hypothesis test results, decreasing the effectiveness and sensitivity of all test statistics. Notice, however, that the best-performing test statistics remain the same as those from the unstandardized setting: EC, CORR, SilEC tests, confirming that the properties underlying those three tests are better able to capture purely topological differences than any other test presented in this paper. Notably after standardization, many of the KC, WKC, and other Sil tests have essentially constant p-values across the percfil variation, losing almost all effectiveness in distinguishing differences. One interpretation is that the differences captured by these tests in the unstandardized version were all due to geometric dissimilarities.

5 Application: Cosmological Simulation Data

Similar to the simulation models from Section 4.1, the Megaparsec cosmic mass distributions are likewise characterized by intricate multiscale configuration of web-like filaments and voids. We apply the proposed methodology to cosmological simulation data.

- Discuss LSS

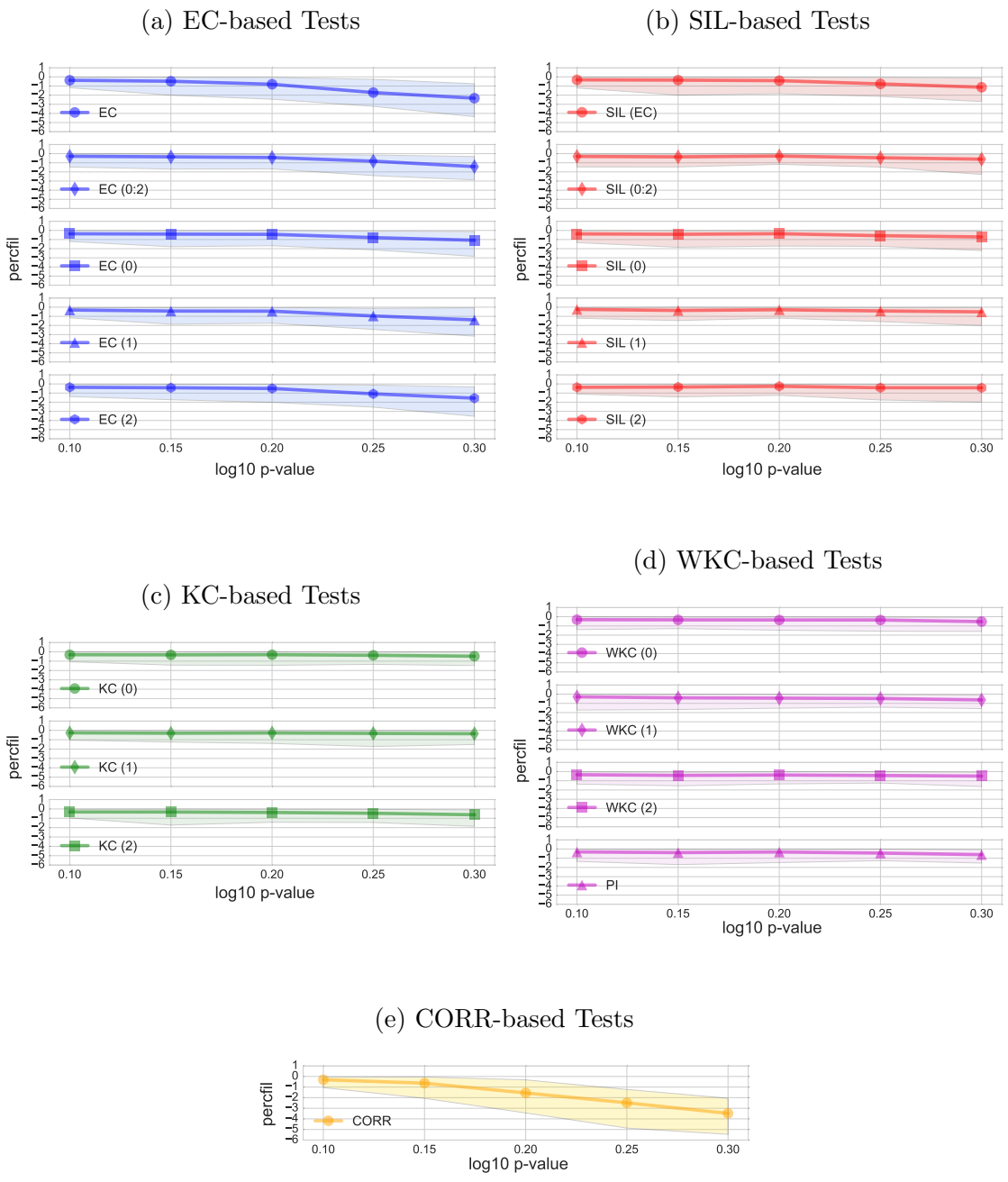


Figure 8: A similar set of results to Figure 7 except that all persistence diagrams were standardized prior to hypothesis testing. In the CORR test, the Voronoi dataset itself was standardized prior to testing.

- Discuss Cosmological simulations
- Discuss Warm vs. Cold DM

Real observations of cosmic web: Great Wall Geller & HUCHRA (1989), Sloan Great Wall Gott III et al. (2005), gas Cantalupo et al. (2014).

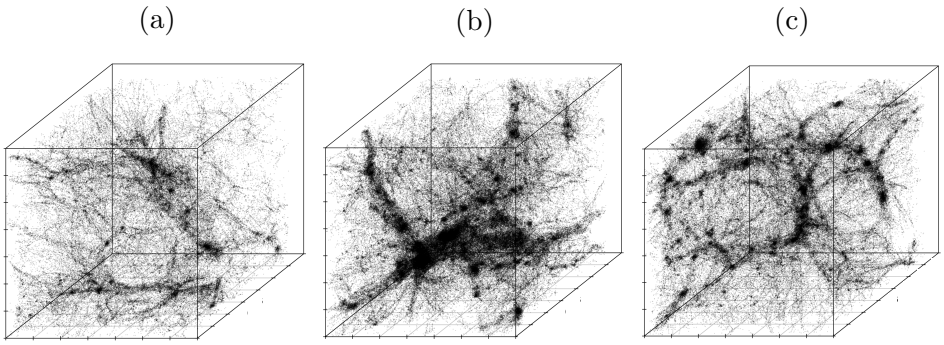


Figure 9: Examples of three quadruple-split samples from a CDM simulation.

One parameter of interest is the state of the dark matter in the LSS. 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. 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 hypothesis testing framework would offer a method to *quantify* the topological differences between simulations with underlying WDM and CDM assumptions.

In the simulation study, all test were standard two-sample T-tests using the varying proposed test statistics derived from the persistence diagrams. For our two cosmological simulation cubes, one might consider cutting up the cubes into sub-cubes to produce similar sets of samples as used in the Voronoi experiments. Let a *double-split* be defined as splitting the simulation cube along both the x and y axis into 2 equal groups, creating 4 equally sized cubes in total. Similarly, a *quadruple-split* produces 64 equally sized cubes, further splitting each of the cubes in a *double-split* set into four sub-cubes. Using this splitting technique on both the CDM and WDM simulations, we end up with two usable data sets.

By doing this, however, the WDM and CDM data sets are correlated due to their identical initial conditions. That is, the differences between the two cosmological simulation cubes is due to differences in the physics of warm vs. cold DM. Because of this, the skeleton structure of the warm and cold DM simulations are nearly identical, allowing for paired T-tests. Note that there is also correlation between the samples due to the large-scale structure crossing the boundaries of the sub-cubes. We investigate this by changing the sizes of the sub-cubes, but consequently by doing so, decreases our sample size.

5.1 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 assumed to be less than 10Myr, to the present day (13.8 Gyr). The cosmological parameters are consistent with the 7-year results from the WMAP satellites: matter density $\Omega_0 = 0.272$, dark energy density $\Omega_\Lambda = 0.728$, Hubble parameter $h_0 = 0.704$, spectral index $n_s = 0.967$, and power spectrum normalization $\sigma_8 = 0.81$. The mass of the simulation particle is $8.8x10^6\ M_{sun}$. Haloes and subhaloes were identified using the SUBFIND algorithm (Springel et al. (2001)), 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 Schaye et al. (2015); we stress that the runs used in this paper use gravity alone.

We use two simulations in this study, one cold dark matter (CDM) and the other warm dark matter (WDM) (recall Figure 1). They make use of the same initial phases, and differ in that the latter has wave amplitudes rescaled using the transfer function of a 3.3keV thermal relic, the relic mass chosen to be in agreement with the Lyman-alpha constraints of Viel et al. (2013). This results in the suppression of structure on the scale of dwarf galaxies. Spurious subhaloes have been removed using the algorithm of Lovell et al. (2014). 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

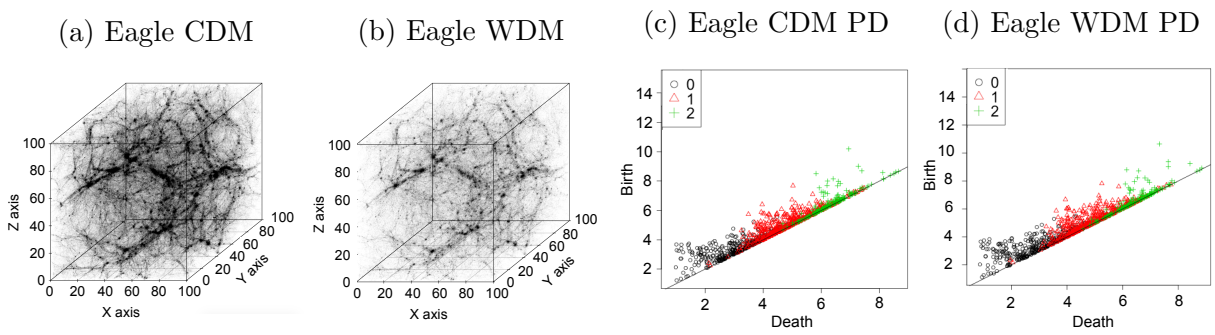


Figure 10: (a, b) Visualization of the complete CDM and WDM simulations. (c, d) Their corresponding persistence diagrams. Although the CDM structure is denser, the persistence diagrams appear comparable.

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.

5.2 Results

Table 2 shows both the standardized (right) and unstandardized (left) results of the five categories of hypothesis tests. The unstandardized results suggest that with a higher number of splits, we are able to focus on smaller-scale topology, finding more significant differences. Through standardizing the persistence diagrams, we see that all the frameworks produce higher p-values on average and are less confident in the difference between double-split and quadruple-split datasets. This seems to explain that similar to the Voronoi foams, a large degree of the difference between WDM and CDM assumptions in the LSS are due to geometrical properties such as size.

5.3 Localizing Differences

A natural question after discovering that topological differences exist in both geometry and topology is how are these differences distributed? One hypothesis might be that the differences are grouped among clustered sections of the observable Universe while other

Unstandardized			Standardized		
Test	Double Split	Quadruple Split	Test	Double Split	Quadruple Split
EC	1.156e-06	7.379e-26	EC	1.102e-04	7.345e-12
$EC_{0:2}$	2.104e-05	7.379e-28	$EC_{0:2}$	1.327e-05	3.972e-14
EC_0	3.273e-08	3.811e-30	EC_0	1.387e-06	1.614e-12
EC_1	1.849e-05	0.935	EC_1	0.0235	0.00143
EC_2	0.340	0.0877	EC_2	0.00385	1.977e-07
Sil_{EC}	7.709e-08	2.455e-20	Sil_{EC}	0.0122	0.00500
$Sil_{0:2}$	1.892e-06	1.114e-33	$Sil_{0:2}$	4.295e-04	8.072e-12
Sil_0	2.958e-08	1.489e-34	Sil_0	6.471e-06	1.762e-11
Sil_1	1.169e-05	2.884e-23	Sil_1	0.0282	3.899e-07
Sil_2	0.925	0.0345	Sil_2	0.00294	1.545e-04
KC_0	0.442	0.000	GC_0	0.986	0.856
KC_1	0.192	0.000	GC_1	0.962	0.956
KC_2	0.248	0.00199	GC_2	0.962	0.260
WKC_0	0.084	0.000	WKC_0	0.092	0.000
WKC_1	0.051	0.000	WKC_1	0.066	0.000
WKC_2	0.496	0.000	WKC_2	0.459	0.001
PI	0.923	0.281	PI	0.999	0.306
CORR	6.656e-04	7.355e-16	CORR	0.0289	0.918

Table 2: P-values from hypothesis tests on the unstandardized (left) and standardized (right) WDM & CDM simulations by double, and quadruple splits. A p-value of 0.000 comes from a permutation test with no positive examples.

galaxoes are practically identical. It is also possible that the differences are uniformly distributed across our cosmic data set. To explore these questions more, we looked at the Bottleneck distances between different cubic splits as the number of splits increased. Because a higher number of splits focuses on smaller parts of the CDM/WDM simulation, we expect higher resolution into the topological features specific to that split's region, and more comparability to other neighboring regions. Figure 11 shows 3 sets of heatmaps for each of the 3 persistent homologies. Each set includes another 3 heatmaps, detailing the Bottleneck distances between congruent CDM and WDM unsplit, double-split, and quadruple-split partitions.

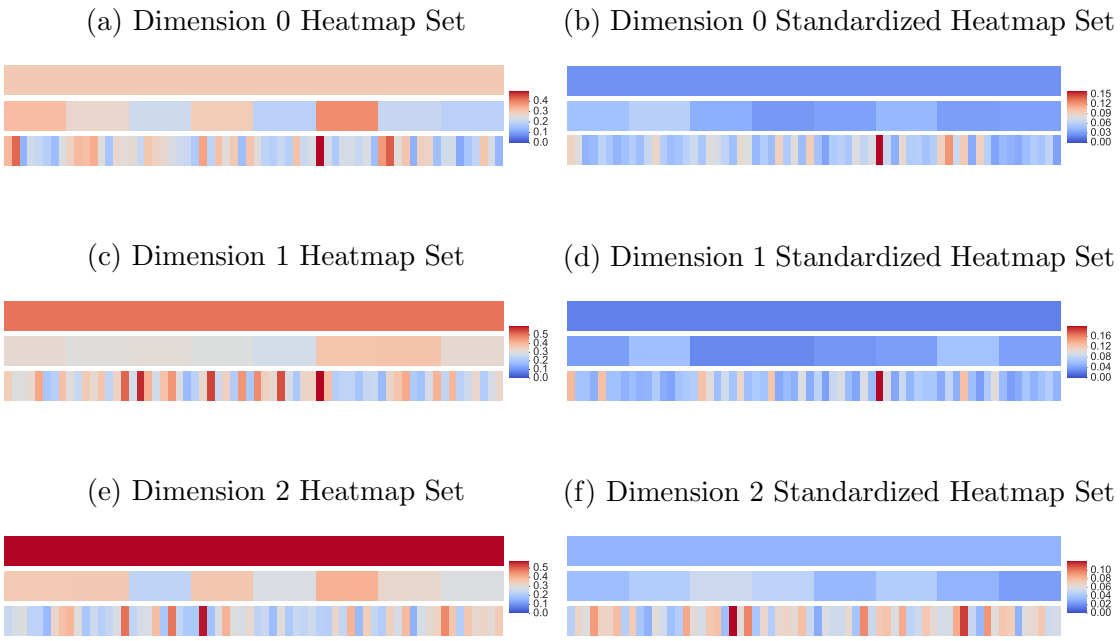


Figure 11: Each subfigure contains three horizontal heatmaps respectively representing (from top to bottom) the Bottleneck distances between each slice of unsplit, double split, and quadruple split pairs from respective WDM and CDM data sets by dimension. Subfigures (a, c, e) are unstandardized persistence diagrams while the subfigures (b, d, f) are standardized. Each individual heatmap is a vectorization of the cubic slices produced from the WDM and CDM Eagle simulations.

From Figure 11, we can infer that the topological and geometric differences are certainly

not uniform across different sections of the observable Universe. For each of the three homologies, introducing a greater number of splits uncovers greater variance in magnitudes of Bottleneck distances between CDM and WDM cubes. This suggests that it is possible for topological differences to be masked by local similarities, and that there exist partitions of topologically similar areas and partitions of topologically dissimilar areas within the EAGLE simulation. Further work should explore what areas in the Universe occupy these two partitions and if there exist physical evidence of differences.

(a) EC (High) (b) CORR (High) (c) EC (Low) (d) CORR (Low)

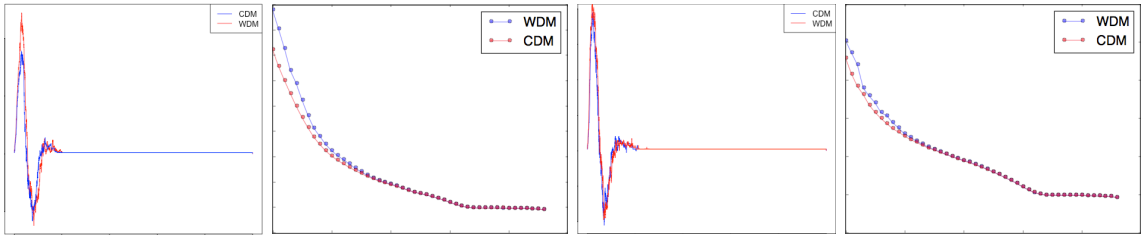


Figure 12: A comparison of the EC Function and the CORR Function for Dimension 1 between the cube with the highest bottleneck distance and the cube with the lowest.

Provided that the Euler Characteristic Function and the Correlation Function were the two most effective hypothesis tests, to confirm the validity of these heatmaps, we compared these two functions for the unstandardized cubes with the highest Bottleneck distance and the unstandardized cubes with the lowest Bottleneck distance. As expected, Figure **??** confirms that the (High) EC and (High) CORR functions have far more discrepancies than their (Low) counterparts, also suggesting that certain partitions are more similar topologically than others. The same pattern arises for Dimension 0 and Dimension 2.

We can do the same heatmap experiment with the standardized persistence diagrams of the CDM and WDM simulations prior to splitting. As confirmed in the hypothesis testing of both the Voronoi and the EAGLE simulations, standardizing removes geometric differences, greatly reducing the Bottleneck distances, as seen in Figure 11(b,d,f). Notably, for dimensions 0 and 1, the cube with the largest bottleneck distance remained the same as in the unstandardized heatmaps, suggesting that those differences are largely not due to size, but true topological distinctions. Dimension 2, however, was less consistent, possibly

indicating that voids in CDM and WDM are more aberrant in terms of size, not topological shape. Despite the differences standardizing induced, we are still confident that differences in topology alone are not uniform across the EAGLE Universe: certain splits are consistently indicating differences in dark matter make-up between warm and cold assumptions.

6 Conclusion

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, and the distinction between the two. 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.

SUPPLEMENTARY MATERIAL

Additional voronoi simulations: Hypothesis testing p-values for voronoi simulations of percent filament varying from 0.1 to 0.9. See Figure 13.

Additional voronoi simulations: Standardized hypothesis testing p-values for voronoi simulations of percent filament varying from 0.1 to 0.9. See Figure 14.

References

Adams, H., Chepushtanova, S., Emerson, T., Hanson, E., Kirby, M., Motta, F., Neville, R., Peterson, C., Shipman, P. & Ziegelmeier, L. (2015), 'Persistent images: A stable vector representation of persistent homology', *arXiv preprint arXiv:1507.06217* .

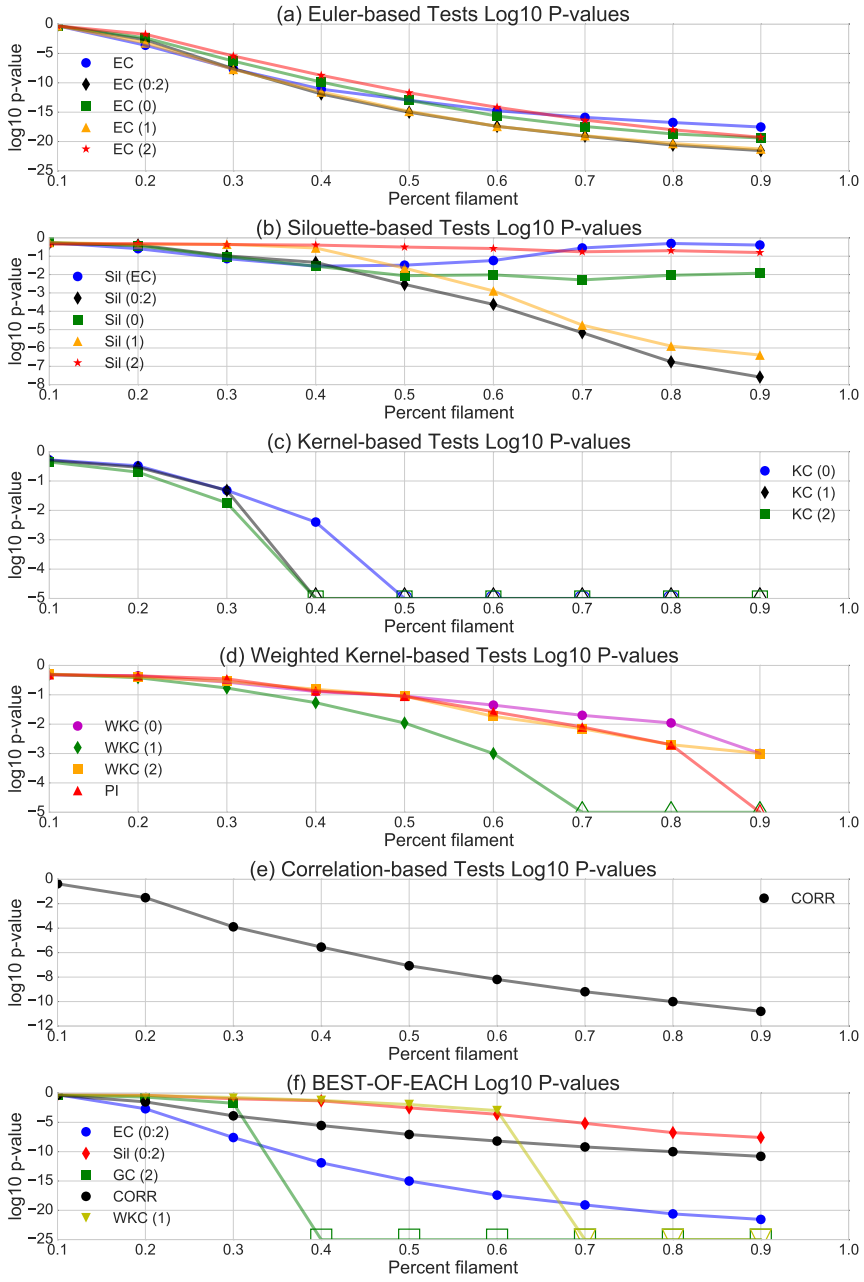


Figure 13: $\log_{10}$ p-values for the proposed test statistics in the simulation study: (a) Euler-based, (b) Silhouette-based, (c) Kernel-based, (d) Weighted Kernel-based, (e) Correlation-based. The fifth plot (f) includes the best hypothesis test from each of three frameworks. X-axis represents the percent filament of the set being compared to the baseline set (from 0.1 to 0.9 by 0.1 increments); y-axis shows the p-values in $1\log_{10}$ space. The lines plot the median $\log_{10}$ p-value of the 100 iterations.

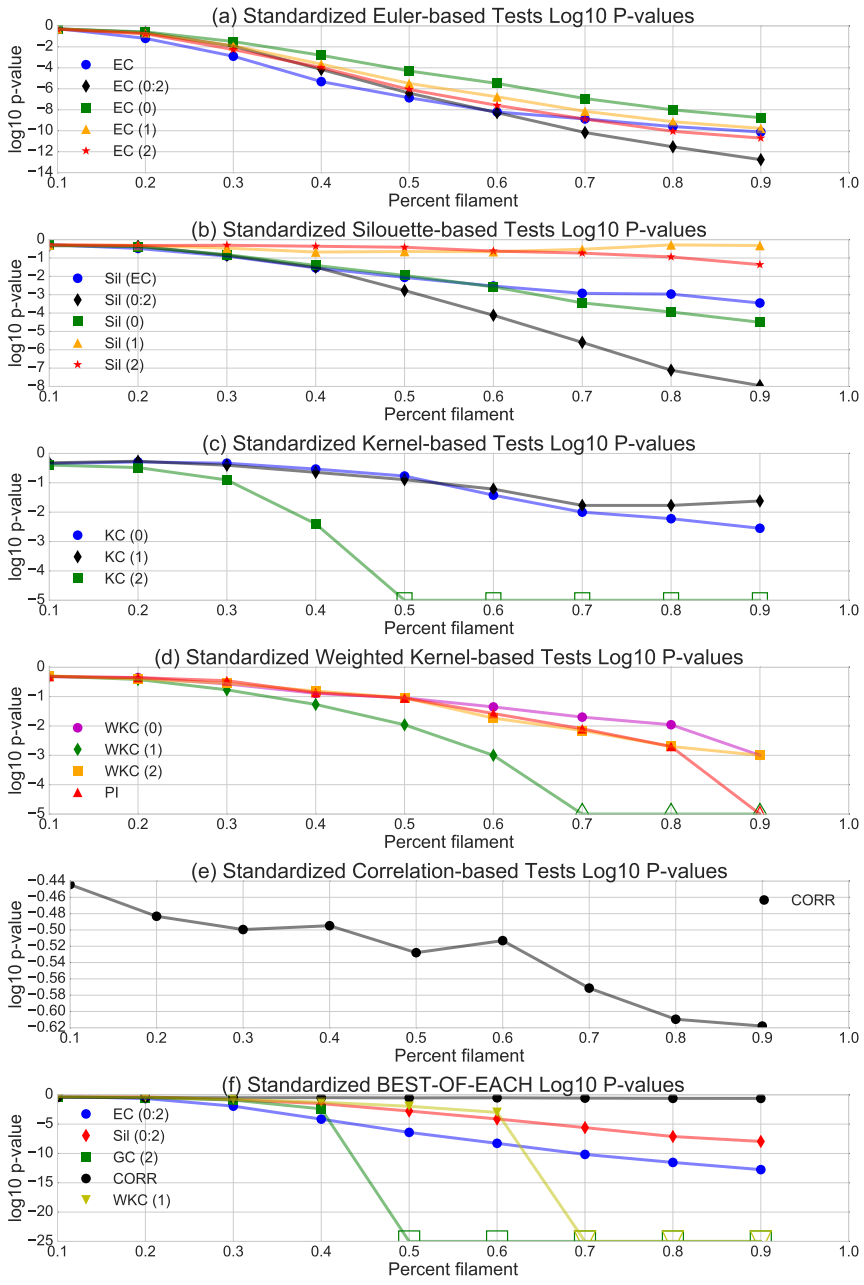


Figure 14: A similar set of results to Figure 13 (percfil 0.1 to 0.9) except that all persistence diagrams were standardized prior to hypothesis testing. In the CORR test, the Voronoi dataset itself was standardized prior to testing.

Bendich, P., Marron, J., Miller, E., Pieloch, A. & Skwerer, S. (2014), 'Persistent homology analysis of brain artery trees', *arXiv preprint arXiv:1411.6652* .

Bond, J. R., Kofman, L. & Pogosyan, D. (1996), 'How filaments of galaxies are woven into the cosmic web', *Nature* **380**, 603–606.

Bubenik, P. (2015), 'Statistical topological data analysis using persistence landscapes', *Journal of Machine Learning Research* **16**(1), 77–102.

Cantalupo, S., Arrigoni-Battaia, F., Prochaska, J. X., Hennawi, J. F. & Madau, P. (2014), 'A cosmic web filament revealed in lyman-[agr] emission around a luminous high-redshift quasar', *Nature* **506**(7486), 63–66.

Centrella, J. & Melott, A. L. (1983), 'Three-dimensional simulation of large-scale structure in the universe'.

Chazal, F., Cohen-Steiner, D. & Mérigot, Q. (2011), 'Geometric inference for probability measures', *Foundations of Computational Mathematics* **11**(6), 733–751.

Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A. & Wasserman, L. (2014), Stochastic convergence of persistence landscapes and silhouettes, *in* 'Proceedings of the thirtieth annual symposium on Computational geometry', ACM, p. 474.

Chen, Y.-C., Wang, D., Rinaldo, A. & Wasserman, L. (2015), 'Statistical analysis of persistence intensity functions', *arXiv preprint arXiv:1510.02502* .

Cisewski, J., Croft, R. A., Freeman, P. E., Genovese, C. R., Khandai, N., Ozbek, M. & Wasserman, L. (2014), 'Non-parametric 3d map of the intergalactic medium using the lyman-alpha forest', *Monthly Notices of the Royal Astronomical Society* **440**(3), 2599–2609.

Cohen-Steiner, D., Edelsbrunner, H. & Harer, J. (2007), 'Stability of persistence diagrams', *Discrete & Computational Geometry* **37**(1), 103–120.

Cooray, A. & Sheth, R. (2002), 'Halo models of large scale structure', *Physics Reports* **372**(1), 1–129.

Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. (1985), 'The evolution of large-scale structure in a universe dominated by cold dark matter', *The Astrophysical Journal* **292**, 371–394.

Doroshkevich, A., Kotok, E., Novikov, I., Polyudov, A., Shandarin, S. & Sigov, Y. S. (1980), 'Two-dimensional simulation of the gravitational system dynamics and formation of the large-scale structure of the universe', *Monthly Notices of the Royal Astronomical Society* **192**(2), 321–337.

Duong, T., Goud, B. & Schauer, K. (2012), 'Closed-form density-based framework for automatic detection of cellular morphology changes', *Proceedings of the National Academy of Sciences* **109**(22), 8382–8387.

Edelsbrunner, H. & Harer, J. (2010), *Computational topology: an introduction*, American Mathematical Soc.

Geller, M. & HUCHRA, J. (1989), 'Mapping the universe', *Science* **246**(4932), 897–903.

Gott III, J. R., Schlegel, D., Hoyle, F., Vogeley, M., Tegmark, M., Bahcall, N., Brinkmann, J. et al. (2005), 'A map of the universe', *The Astrophysical Journal* **624**(2), 463.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012), 'A kernel two-sample test', *The Journal of Machine Learning Research* **13**(1), 723–773.

Hatcher, A. (2002), *Algebraic topology*, .

Hilbe, J. M., Riggs, J., Wandelt, B. D., de Souza, R. S., Ishida, E. E., Cisewski, J., Surdin, V., Killedar, M., Trotta, R., Bassett, B. et al. (2014), 'Life, the universe, and everything', *Significance* **11**(5), 48–75.

Icke, V. & van de Weygaert, R. (1987), 'Fragmenting the universe', *Astronomy and Astrophysics* **184**, 16–32.

Icke, V. & van de Weygaert, R. (1991), 'The galaxy distribution as a voronoi foam', *Quarterly Journal of the Royal Astronomical Society* **32**, 85–112.

Landy, S. D. & Szalay, A. S. (1993), 'Bias and variance of angular correlation functions', *The Astrophysical Journal* **412**, 64–71.

Lovell, M. R., Frenk, C. S., Eke, V. R., Jenkins, A., Gao, L. & Theuns, T. (2014), 'The properties of warm dark matter haloes', *Monthly Notices of the Royal Astronomical Society* **439**(1), 300–317.

Munkres, J. R. (1984), *Elements of algebraic topology*, Vol. 2, Addison-Wesley Menlo Park.

Schaye, J., Crain, R. A., Bower, R. G., Furlong, M., Schaller, M., Theuns, T., Dalla Vecchia, C., Frenk, C. S., McCarthy, I., Helly, J. C. et al. (2015), 'The eagle project: simulating the evolution and assembly of galaxies and their environments', *Monthly Notices of the Royal Astronomical Society* **446**(1), 521–554.

Schneider, A., Smith, R. E., Macciò, A. V. & Moore, B. (2012), 'Non-linear evolution of cosmological structures in warm dark matter models', *Monthly Notices of the Royal Astronomical Society* **424**(1), 684–698.

Sousbie, T. (2011), 'Ther persistent cosmic web and its filamentary structure - I. Theory and implementation', *Monthly Notices of the Royal Astronomical Society* **414**(1), 350 – 383.

Sousbie, T., Pichon, C. & Kawahara, H. (2011), 'The persistent cosmic web and its filamentary structure – II. Illustrations', *Monthly Notices of the Royal Astronomical Society* **414**(1), 384 – 403.

Springel, V., Frenk, C. S. & White, S. D. (2006), 'The large-scale structure of the universe', *Nature* **440**(7088), 1137–1144.

Springel, V., White, S. D., Tormen, G. & Kauffmann, G. (2001), 'Populating a cluster of galaxies–i. results at z= 0', *Monthly Notices of the Royal Astronomical Society* **328**(3), 726–750.

Van De Weygaert, R. (2007), Voronoi tessellations and the cosmic web: Spatial patterns and clustering across the universe, *in* 'Voronoi Diagrams in Science and Engineering, 2007. ISVD'07. 4th International Symposium on', IEEE, pp. 230–239.

Van De Weygaert, R., Vegter, G., Edelsbrunner, H., Jones, B. J., Pranav, P., Park, C., Hellwing, W. A., Eldering, B., Kruithof, N., Bos, E. et al. (2011), Alpha, betti and the megaparsec universe: on the topology of the cosmic web, *in* 'Transactions on Computational Science XIV', Springer-Verlag, pp. 60–101.

Viel, M., Becker, G. D., Bolton, J. S. & Haehnelt, M. G. (2013), '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.