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

Mike Wu

Department of Computer Science, Yale University
Jessi Cisewski*
Department of Statistics, Yale University
Brittany T. Fasy
School of Computing, Montana State University
Wojciech Hellwing
Institute of Cosmology and Gravitation, University of Portsmouth
Mark R. Lovell
Gravitation AstroParticle Physics, University of Amsterdam
Alessandro Rinaldo
Department of Statistics, Carnegie Mellon University
Larry Wasserman
Department of Statistics, Carnegie Mellon University

December 10, 2016

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

*Corresponding author. The authors gratefully acknowledge Yale Information Technology Services

statistical comparisons between complicated spatial data such as LSS in cosmology, but are also relevant to other areas of science. We present several test statistics using persistence diagrams, carry-out a simulation study to investigate the suitableness of the proposed test statistics, and finally apply the inference framework to study topological disparities between assumptions of warm and cold dark matter.

Keywords: Astrostatistics, persistent homology, topological data analysis

1 Introduction

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

Rigorous comparisons of spatially complex web-like data such as the large-scale structure (LSS) of the Universe (see Figure 1) are notoriously challenging due, in part, to the difficulty in capturing the randomness of geometric and topological structures. However, these comparisons are important because there exists 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, as seen 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). For example, LSS can reveal characteristics of dark energy, which is thought to be the driver of the acceleration of the Universe (Sánchez et al. 2012). In order to study theoretical aspects of the formation and evolution of LSS, cosmologists develop large 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 theory

is that the Universe is made up of dark energy, DM, and baryonic matter. However, the nature of DM is still a mystery, but there are hypotheses regarding its possible particle behavior. Hot DM would consist of particles that travel with ultrarelativistic speeds, while cold DM particles would 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, one realization of our Universe under CDM from the EAGLE cosmological simulation (Schaye et al. 2015) is displayed, along with a realization assuming WDM. Though there are similarities in shape of the densest regions (called *filaments*), there are differences in the distribution of matter about these filaments.

(a) WDM Simulation (b) CDM Simulation

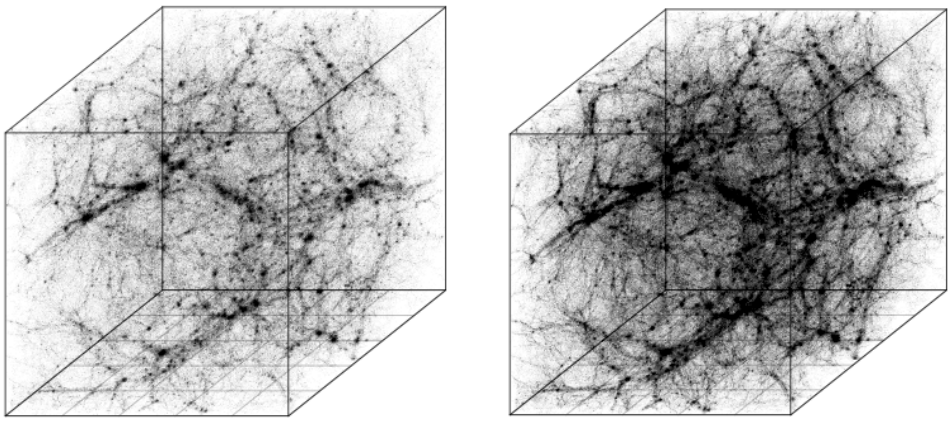


Figure 1: Dark matter-only cosmological simulations assuming (a) warm dark matter (WDM) and (b) cold dark matter (CDM). See Section 5 for details.

In this paper, we wish to better identify differences in the topology between structures using persistent homology. 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 statistics followed by more background on LSS.

And finally, we apply the hypothesis tests to quantify the topological disparities between warm and cold DM using cosmological simulation data. 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 spatial structure of a parameterized family of topological spaces (e.g., keeping track of the so-called births and deaths of the noted homological features as a topological space changes with a varying parameter). The type of data we are investigating is a point cloud, where each point can represent some unit of mass such as a galaxy or, for cosmological simulation data, a certain mass of DM. The homological features mentioned above that get tracked in the filtration have cosmological interpretations in dimensions zero, one, and two. Before providing more details about persistent homology, we explain the interpretation of different ordered holes in the Universe.

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, these 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 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. Again, if our topological space is a k-nn graph, then the voids are the unfilled spaces inside enclosed H_1 features. In cosmology, these would be the low-density regions typically away from the filamentary, LSS.

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.

Filtrations To derive the persistent homology for some density function p over $\mathcal{X}$, consider a threshold r, represented by a hyperplane that divides the domain of p into two separate segments: a super-level set, defined as $\{x \in \mathcal{X} \in p \text{ s.t. } p(x) \geq r\}$, and a sub-level set $\{x \in \mathcal{X} \text{ s.t. } p(x) < r\}$. If r is initialized at ∞, then the super-level set is empty and the sub-level set contains the domain of p. The evolving topological space is characterized by its homology as r decreases to $-\infty$ (or zero when p is a density). 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 $p^{-1}([r, \infty)]$. More specifically, as r intersects p, the super-level set is no longer empty and is instead, composed of disjoint 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 shows the persistence diagram which is discussed below.

Tracking Homology Generators Figure 2b and Figure 2c show the threshold, r decreasing from 1.1 to 0.1. In that interval, the upper-level set changed from having six connected components (H_0's) and zero loops (H_1's) to having one connected component

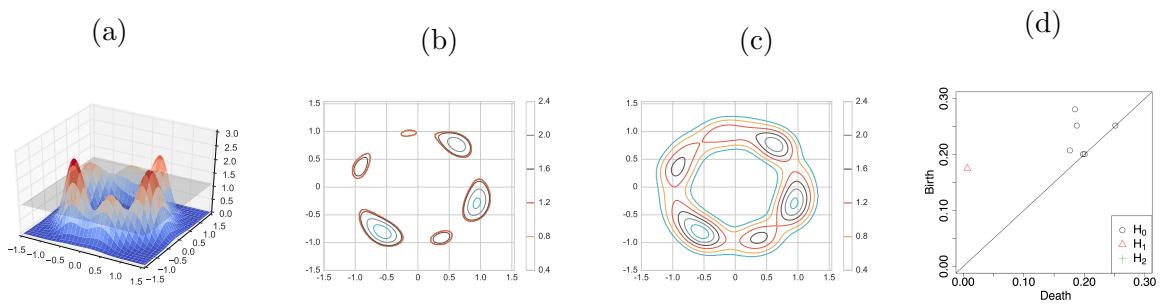


Figure 2: We illustrate persistent homology with this two-dimensional example. The density, p, in (a) shows 3 steep peaks and 3 shallow peaks distributed around a circle. The gray hyperplane defines the threshold for super-level sets, and (b) and (c) plot the super-level sets $p^{-1}[1.10, \infty)$ and $p^{-1}[0.10, \infty)$ respectively. As the threshold decreases, the contour more clearly defines a loop-like structure. In (d), the persistence diagram for the super-level set filtration of p. We highlight the upper left quadrant based at $(0.18, 0.02)$, which contains an H_1 generator as a red triangle. The remaining 0-dimensional H_0 generators represent the connected components from each of the 6 peaks and are displayed as black circles. There are no H_2 generators (green pluses).

and one loop. The time in the filtration when homology features appear, the *birth* of the feature, and the time when a feature joins other features, the *death* of the feature, are captured in a persistence diagram. Figure 2d displays the persistence diagram for the function in Figure 2a, where each point represents the birth time (y-axis) and death time (x-axis) of a homological feature. A point (x, x) on the diagonal represents a feature with a lifespan of 0. The *persistence* of a point (b, d) is the length of the interval of the persistence parameter that supports 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 $\frac{(b-d)}{\sqrt{2}}$. Sometimes it is desired to interpret features with longer lifetimes as topological signal, and the features with shorter lifetimes (closer to the diagonal) as topological noise.

2.2 Derivatives of Persistence Diagrams

While persistence diagrams are useful summaries of the topology of a data set, they are not easy to work with directly. For example, the distance between two persistence diagrams can be calculated using, for example, the bottleneck distance or the q-Wasserstein distance, but both are computationally expensive. Fréchet means and medians have been defined for spaces of persistence diagrams (Turner et al. 2014), but are also computationally expensive and not necessarily unique. Instead we consider transformations and summaries of persistence diagrams that make computations more tractable. Below are several approaches that further summarize a persistence diagram and will be used in §3 to develop hypothesis tests.

Weighted silhouette functions Weighted silhouette functions are formed by weighting a particular functional summary of persistence diagrams called *landscape functions* (Bubenik 2015). 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 persistence landscape function is displayed in Figure 3.

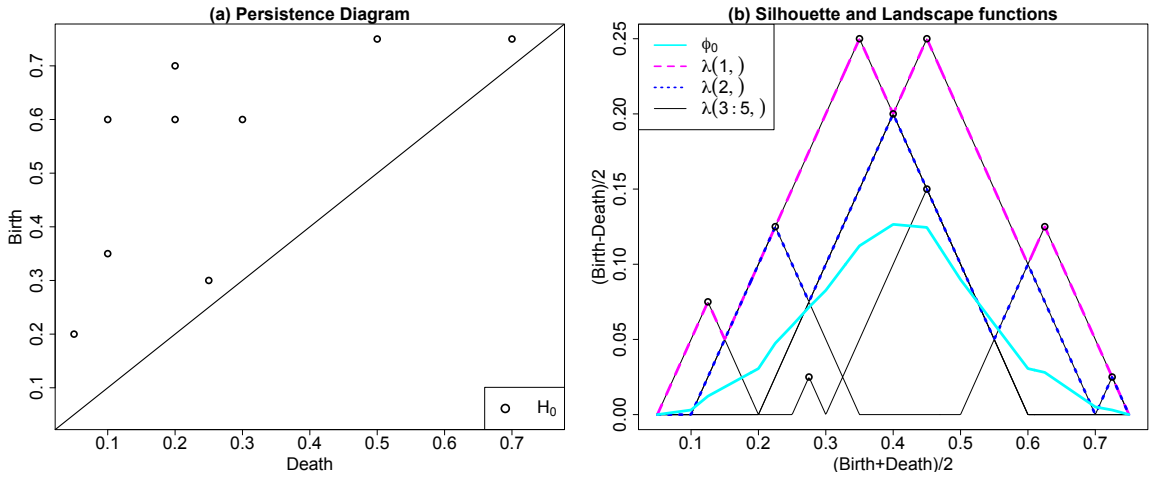


Figure 3: A persistence diagram (a) along with its silhouette and landscape functions (b) for H_0. The solid cyan curve is the silhouette function with $p = 1$; the magenta dashed and blue dotted lines in (b) are landscape functions $\lambda_{D_0}(1, \cdot)$ and $\lambda_{D_0}(2, \cdot)$, respectively.

Rather than working with each k of $\lambda_{D_h}(k, t)$ individually, weighted silhouettes provide a way of combining the information in the collection of landscape functions. 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 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. An example of a weighted silhouette function is provided in Figure 3. More details and theoretical properties of landscapes and silhouettes are provided in (Chazal et al. 2014).

Euler Characteristic Function The Euler characteristic is a topological invariant and defined as: $\chi = \sum_{i=0}^{N} (-1)^i \beta_i$, 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 equation simplifies to: $\chi = \beta_0 - \beta_1 + \beta_2$. As the filtration threshold t decreases and new features are born or old ones die, the Euler Characteristic changes so we consider

the Euler Characterisitic function, $\chi(t)$, $t \in [t_{\min}, t_{\max}]$.

Smooth Persistent Diagrams Another variation of persistent diagrams are smoothed persistence diagrams. Chen et al. (2015) introduced *persistent intensity functions*, which is a weighted kernel density estimate of a persistence diagram with weights as a function of a point's persistence. Let $\mathcal{D} = (b_i, d_i) : i = 1, \cdots, m$ be a persistence diagram where m is the number of persistent features for dimension h. Then the weighted intensity function for the persistence diagram is

$$X(t_1, t_2) = \sum_{j=1}^{m} (d_j - b_j) \frac{1}{\tau^2} K\left(\frac{t_1 - d_j}{\tau}\right) K\left(\frac{t_2 - b_j}{\tau}\right)$$

for $t_{\min} \leq t_1 \leq t_2 \leq t_{\max}$ with symmetric kernel function K, and smoothing parameter τ. An example of a persistent intensity function is displayed in Figure 4c.

Persistent Images are similar to persistent intensity functions (Adams et al. 2015) except that the persistence diagram is rotated and the smoothed diagram is pixelized into a uniformly spaced 2-dimensional grid of a chosen resolution, and finally vectorized. The persistence diagram is transformed using the linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ where $T(x, y) = (x, y - x)$; therefore $T(\mathcal{D})$ represent the transformed diagram with birth-persistence coordinates. As with persistent intensity functions, the now transformed diagram is smoothed using weighted kernel density estimation, with weights that are zero on the horizontal axis. Adams et al. (2015) use a piecewise, linear weighting function that assigns a weight of 0 to points with 0 persistence. A 2-dimensional grid is then defined over the smoothed $T(\mathcal{D})$ to turn it into a matrix of a user-selected resolution. The matrix is then vectorized and easily used as inputs into a statistical analysis. See Figure 4d for an example of a persistent image.

3 Methods

Using the variations of persistence diagrams described in §2.2, we describe several options for test statistics for two-sample hypothesis tests for topological structure. For this setup, suppose we have two sets of persistence diagrams, $\{\mathcal{P}_1^{(1)}, \ldots, \mathcal{P}_{n_1}^{(1)}\}$ and $\{\mathcal{P}_1^{(2)}, \ldots, \mathcal{P}_{n_2}^{(2)}\}$. These samples can be used to test $H_0 : \mathcal{P}^{(1)} = \mathcal{P}^{(2)}$ vs. $H_1 : \mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$, where $\mathcal{P}^{(1)}$ and

(a) Point Cloud (b) Persistence Diag (c) Intensity Func. (d) Persistent Image

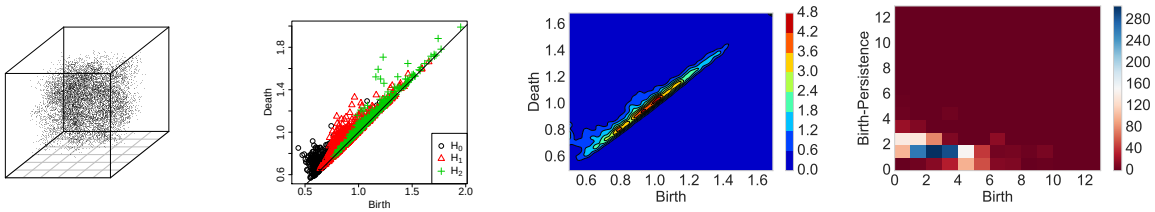


Figure 4: Examples of smoothed persistence diagrams. A 3-dimensional point cloud is provided in (a) along with its persistence diagram in (b). The Persistence Intensity function in (c) and Persistence Image in (d) are both for H_0, See §3 for more details.

$\mathcal{P}^{(2)}$ are the true underlying distributions of persistence diagrams. Probability measures can be defined on a space of persistence diagrams (with a Wasserstein metric) as presented in (Mileyko et al. 2011). We would like the framework to test the hypothesis that the two samples are drawn from populations with different random topologies. 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 §4.2.1.

Given two samples of persistence diagrams, there are a number of possible ways to derive test statistics. We consider functional versions of persistence diagrams as test statistics: the Euler characteristic function, Silhouette function, and a Silhouette-Euler characteristic function (which will be defined below). Given that each observed dataset will have a corresponding functional summary, a *p-value* can be derived from a two-sample T-test by integrating the absolute value of the functional summary. We consider persistent intensity functions and persistent images with p-values derived from a two-sample kernel test statistic (Gretton et al. 2012), and use a permutation test to get a p-value. Additionally, we consider a test using the two-point correlation function. The correlation function is used to capture the spatial behavior of LSS, and we include it to see if the differences in LSS can be attributed to spatial differences rather than topological differences. As with the functional summaries above, p-values are calculated with a T-test using the integral of the absolute value of the correlation function.

The proposed test statistics are discussed in more detail below. Among the proposed test

statistics, we are seeking the summary that is most sensitive to differences in distributions of persistence diagrams produced from LSS.

Euler Characteristic Test (EC) Each persistence diagram in the two samples result in an individual Euler characteristic function, $\chi^{(J)}(t)$, $J = 1, 2$. The two-sample T-test is based on the sample means $\bar{EC}_J = \frac{1}{n_J} \sum_{i=1}^{n_J} \int_{t_{\min}}^{t_{\max}} |\chi_i^{(J)}(t)| \, \mathrm{d}t$. Note that the Euler characteristic function combines all homology dimensions into a single functional summary.

Silhouette Test (SIL) As with EC, the Silhouette Test (SIL) is a two-sample T-test is carried-out by integrating the weighted silhouette functions for each sample $J = 1, 2$ and each homology dimension $h = 0, 1, 2$: $\bar{SIL}_{hJ} = \frac{1}{n_J} \sum_{i=1}^{n_J} \int_{t_{\min}}^{t_{\max}} |SIL_i^{(J)}(t)| \, \mathrm{d}t$. The SIL tests for homology dimension $h = 0, 1, 2$ are denoted, SIL_0, SIL_1, and SIL_2, respectively. In order to combine information across homology dimension, we use a Hotelling's T^2 test with all three dimensions, denoted $SIL_{0:2}$.

Silhouette-Euler Characteristic (SILEC). Another method for simultaneously considering individual silhouettes, $S_h(t)$, across dimensions $h = 0$, 1, and 2, is a modified Euler characteristic function. Instead of calculating the alternating sum of Betti numbers, the Silhouette-Euler characteristic function (SILEC) computes the alternating sum of silhouette functions across the threshold parameter, t, $SILEC(t) = S_0(t) - S_1(t) + S_2(t)$. A p-value is calculated using a T-test in the same fashion as done for EC.

Intensity Kernel Test (IK). The Intensity Kernel Test (IK) uses persistent intensity functions rather than the raw persistent diagrams. The IK statistic used in this paper is computed for two sets of persistence diagrams, and the discrepancy between the diagrams is calculated as the integrated squared difference between two *unweighted* intensity functions instead of points directly from the persistence diagrams. Unweighted intensity functions assigns uniform weights to points on the diagram and therefore does not account for the diagonal boundary on persistence diagrams where birth = death; we considered the usual weighted intensity functions as defined in Chen et al. (2015) in the Weighted Intensity Kernel Test discussed below.

The two-sample test statistic for $X = X_1, \ldots, X_{n_1}$ and $Y = Y_1, \ldots, Y_{n_2}$ is defined using the two-sample Kernel Test from Gretton et al. (2012):

$$\widehat{T}_{IK}(X,Y) = \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} K_\sigma(X_i, X_j) - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_\sigma(X_i, Y_j) + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} K_\sigma(Y_i, Y_j), \quad (1)$$

where n_1 and n_2 are the sizes of the two samples, and $\{X_1, \ldots, X_{n_1}\}$ and $\{Y_1, \ldots, Y_{n_2}\}$ are the two sets of intensity functions. $K_\sigma(X,Y)$ can be thought of as a similarity measure between intensity functions X and Y, and in this case is a Gaussian kernel $K_\sigma(X,Y) = exp(-\frac{||X-Y||^2}{\sigma^2})$ with $||X-Y|| = \sqrt{\int \left(X(t_1,t_2) - Y(t_1,t_2)\right)^2 dt_1 dt_2}$. The σ is a hyperparameter that sets the standard deviation of the Gaussian distribution used in the kernel K_σ: a larger σ will reduce sensitivity to small differences between X and Y, while a smaller σ will heighten sensitivity. The optimal σ 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 for each homology dimension, IK_0, IK_1, and IK_2.

Weighted Intensity Kernel Test (WIK) The intensity function in WIK, unlike IK, uses weighted kernel density estimates (Chen et al. 2015) where the weights are a function of a feature's persistence. Using weighted intensity functions, a p-value is calculated by considering the same kernel test statistic as Equation (1). A permutation test for each homology dimension produces three analogous statistics: WIK_0, WIK_1, and WIK_2.

Persistent Image Test (PI) The Persistent Image Test (PI) uses the Persistent Image transformation of persistent diagrams introduced in Adams et al. (2015). The dimension of the persistent image is 10×10 resulting in a vector of length 100. Unlike the WIK test, the PI test combines all homologies into a single statistic by concatenating the vectors for each homology dimension. A p-value is calculated using the two-sample kernel test statistic from Equation (1) along with permutation test.

Two-point Correlation Function Test (CORR) The Two-Point Correlation Function Test (CORR) considers to the two-point correlation function of the raw data in order to compare the persistent homology tests with a test based on the spatial distribution of the LSS. The two-point correlation function was selected due to its ubiquitous employment

in the astronomy literature (e.g. Baugh 2006, Sánchez et al. 2012), and provides a measure of the degree of clustering in a dataset. The implementation used to get the two-point correlation function is from the Python `TreeCorr` library (Jarvis et al. 2004, Jarvis 2015). As with the other function-based test statistics (where the two-point correlation function is a function that takes distances between points in the raw data as the input), the two-sample T-test is based on the sample means is used for CORR.

4 Simulation Study

To evaluate the performance of the proposed test statistics, 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. In particular, we drew from ideas that use Voronoi tessellations to model the filament structure of the Universe, known as *Voronoi Foam* (Icke & Van de Weygaert 1987, Icke & van de Weygaert 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) (Icke & van de Weygaert 1991).

The cells of the Voronoi tessellation 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 tessellation 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 tessellation), (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 10% to 30% by a 5% step size. See the Appendix for additional

tests for filament percentages ranging from 10% to 90%.

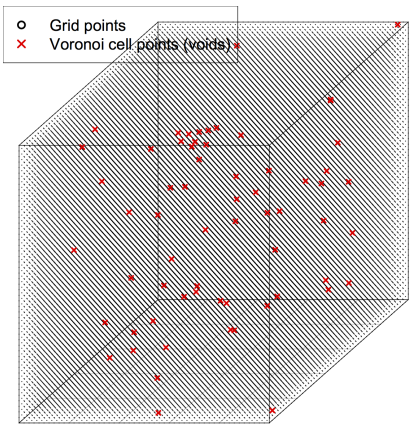
Abbrev	Definition	Value
percWall	Percentage of particles on the walls	$0.98 - p_f$
percFil	Percentage of particles on the filaments	p_f
percClust	Percentage of particles in the clusters	0.02

Table 1: Parameters of LSS model. For the simulation study, p_f will vary from 0.1 to 0.3 by 0.05 increments. See Figure 5 for a visual representation of walls, filaments, and clusters.

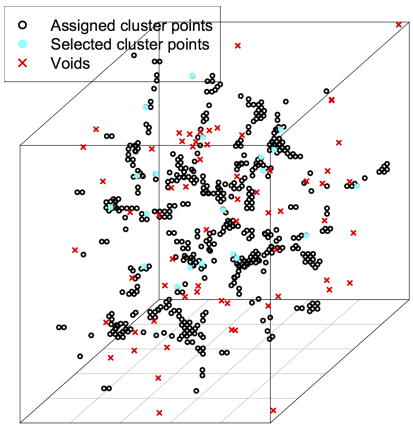
Figure 5 displays the construction procedure of one realization of our simulation model: (i) First a grid is defined at a specified resolution within a specified volume; (ii) then a specified number of points are randomly selected within the volume - these will be used to define the Voronoi tesselation 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 Voronoi Foam models with percFil 10%and 90% are shown in Figure 6 along with their persistence diagrams. One can see that the as the percFil increases, the web-like structure becomes more pronounced, changing the distribution of topological features.

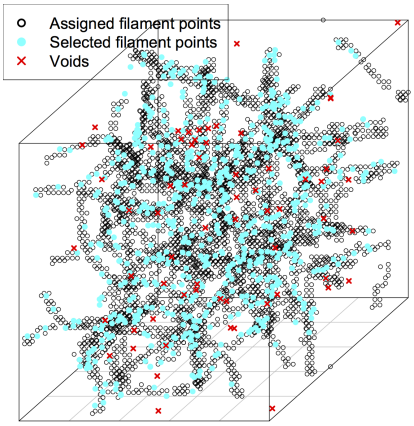
(a) Grid and selected voids



(b) Cluster points



(c) Filament points



(d) Wall points

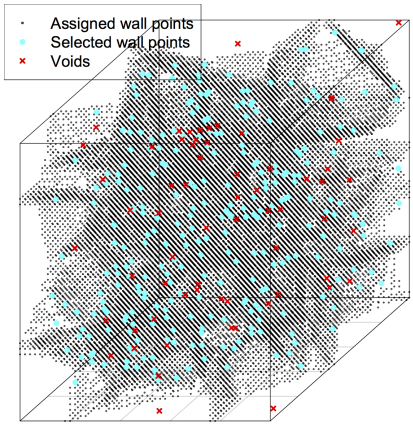


Figure 5: Simulation model construction. (a) A grid is defined and points are randomly selected to define the Voronoi tesselation - the Voronoi cells are the voids. (b) - (d) Based on the location of the voids and the grid, points are defined to be cluster points, filament points or wall points - these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

4.2 Simulation Study Results

The simulated LSS data were generated under 1.25×10^5 box volume, 0.1 resolution, 1×10^4 points, 64 cells (voids), percClust = 2%, percFil = $[10\%, 30\%]$, and percWall = $[68\%, 88\%]$. Persistence 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, a vestigial H_0 element with birth time of 0 and a death time of ∞ (with exception to the

(a) PercFil=10% (b) PercFil=10% (PD) (c) PercFil=90% (d) PercFil=90% (PD)

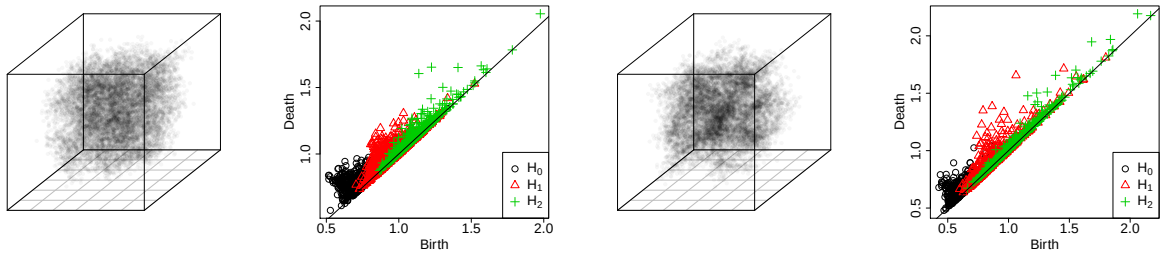


Figure 6: Simulated point clouds and their corresponding persistence diagrams (PD) using PercFil of 10% and 90%. As the PercFil increases, the point clouds tend to co-agulate into dense and sparse regions; the PDs also form more features with longer lifespans. The other parameters used the generate the data are as defined in Table 1.

EC test in which the artifact is preserved). The hypothesis tests were performed on 1000 independent iterations of 15 independent realizations from each of the two populations. The 15 independent datasets are each generated using a percFil setting from 10% to 30% (5% step size).

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 $= 10\%$, and $\{\mathcal{P}_1^{(2)}, \ldots, \mathcal{P}_{15}^{(2)}\}$ drawn from the model with percFil $= p$, $p = 10\%, 15\%, \ldots, 30\%$. (Recall that $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are the true underlying distributions of persistence diagrams.) Similar tests were also completed against a control model with percFil $= 30\%$, and similar results were found.

The simulation study results are displayed in Figure 7, which shows the median $\log_{10}$ p-values along with interval from the $25th$ to the $75th$ quantiles of the 1000 iterations of the proposed test statistics. We see that EC, CORR, and SILEC test are the most effective in distinguishing differences between the models. Additionally, all tests derived from Euler characteristic functions 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 than the SILEC test, which combines silhouettes through a similar linear fashion, performed better than any individual silhouette counterpart and the naive cross-dimension statistic ($SIL_{0:2}$). Finally, CORR, though not as powerful as EC, performs well as percFil increases.

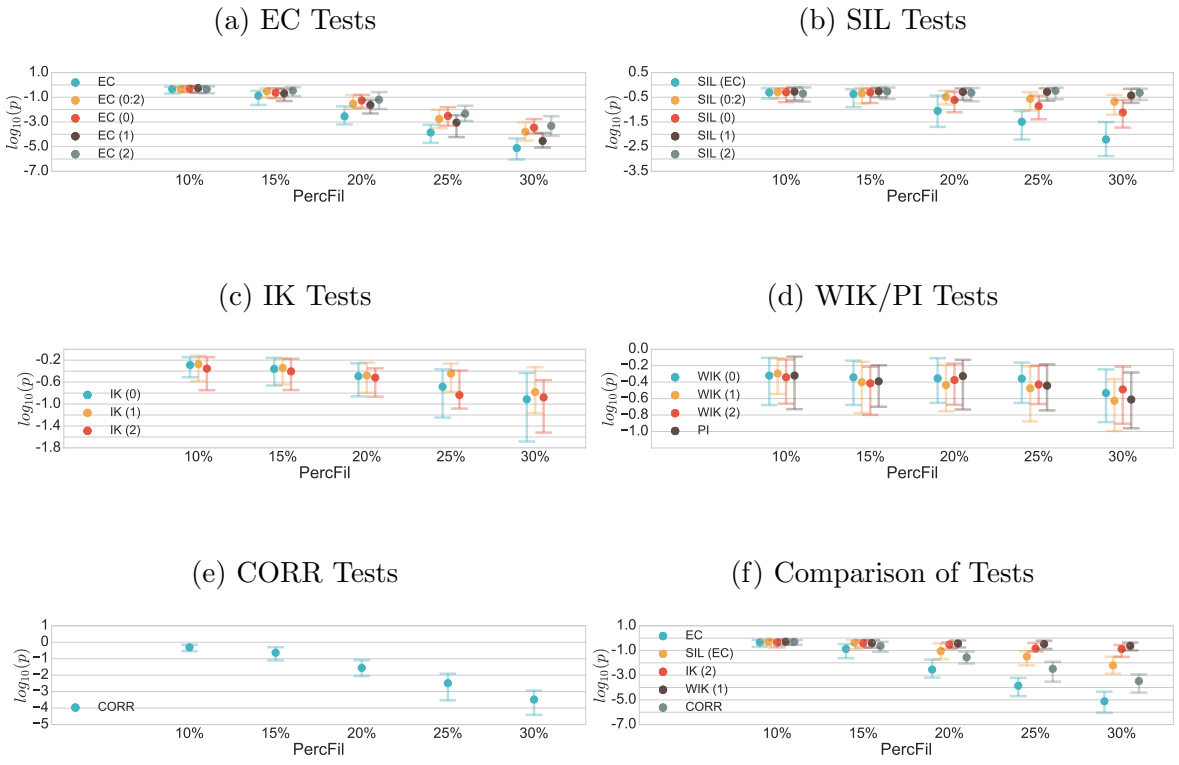


Figure 7: Results for EC, SIL, IK, WIK/PI, and CORR tests. X-axis represents the true PercFil (10%, 15%, 20%, 25%, 30%), compared to the null PercFil of 10%; The vertical axis shows the $\log_{10}$(p-values). The lines plots the median $\log_{10}$ p-value and error bars show the 25th and 75th percentiles of the 1000 iterations. (f) is a comparison of the best results for each approach.

4.2.1 Standardization of persistence diagrams

In the methods proposed above, difference in scale can also result in rejection when, in fact there are not statistically significant topological differences. For example, suppose we are considering two datasets - each has points randomly sampled from the perimeter of a circle with a radius of 1 and 10 respectively. It may or may not be desirable to conclude that the two datasets come from different persistence diagram generators (i.e. conclude $\mathcal{P}^{(1)} \neq \mathcal{P}^{(2)}$) since inference would be based on geometrical (scaling) differences rather than topological differences. If we wish to focus only on topological differences, we propose a possible preprocessing step to normalize scaling. 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 could be used for the standardization. The exception is with the CORR, in which the point cloud itself is standardized to $[0,1] \times [0,1] \times [0,1]$.

We repeated the simulation study from §4.2 except including our standardization; the results are displayed in Figure 8. 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: 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, the KC, WIK, and other SIL tests have essentially constant p-values across the percFil variation, losing effectiveness in distinguishing differences. One interpretation may be that these tests captured only geometric dissimilarities in the unstandardized setting.

5 Application: Cosmological Simulation Data

In order to study differences in LSS under varying cosmological models, cosmologists run large-scale, computationally intensive simulations using the varying cosmological model

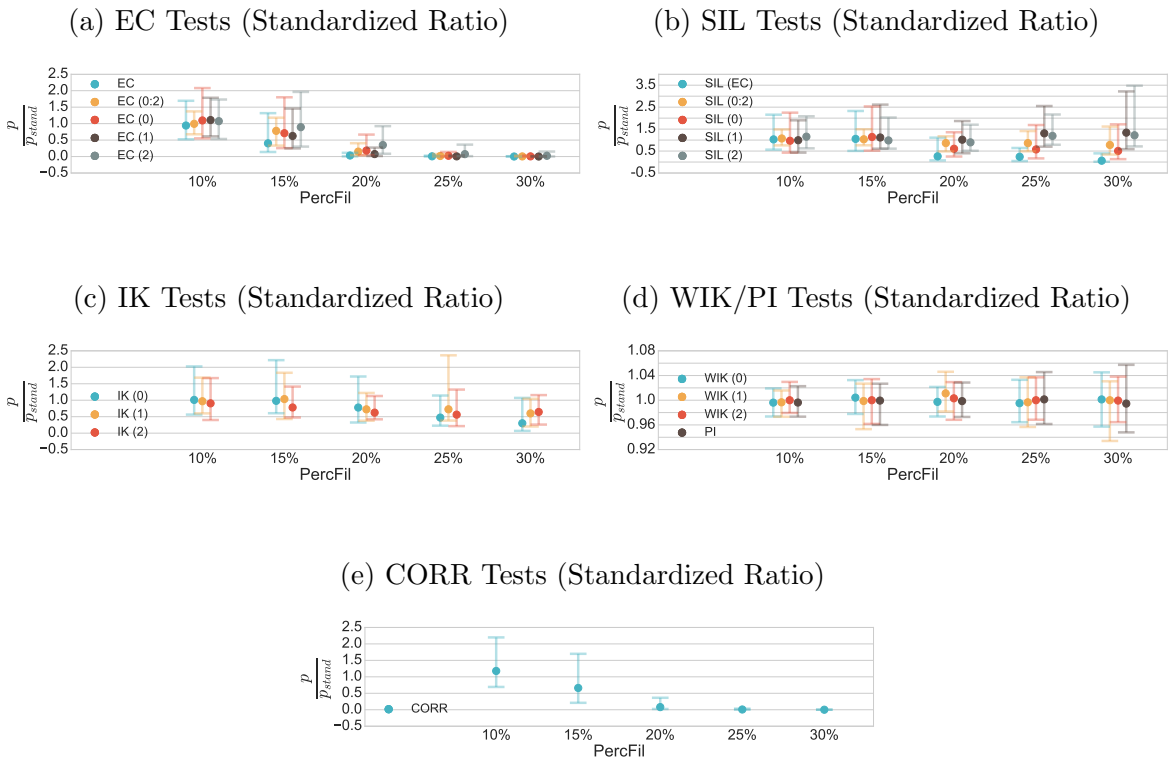


Figure 8: Results for EC, SIL, IK, WIK, PI, and CORR tests. X-axis represents the true PercFil (10%, 15%, 20%, 25%, 30%), compared to the null PercFil of 10%; The vertical axis shows the ration of the unstandardized p-values (from Figure 7) over the standardized p-values. The lines plots the median ratio and the error bars show the 25th and 75th percentiles of the 1000 iterations.

settings as inputs. In this work, we are interested in how the LSS under the assumption of CDM compares to the LSS under the assumption of WDM. We analyze the N-body simulations of structure formation displayed in Figure 1. The simulation box is 100 comoving[1] Mpc on each 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 $8.8 \times 10^6\ M_{sun}$. 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.

Both the WDM and CDM simulations make use of the same initial phases (giving them similar large-scale features), and differ in that the latter has wave amplitudes rescaled using the transfer function of a 3.3keV thermal relic, with 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 9 shows the persistence diagrams for the WDM and CDM simulations in Figure 1. The CDM data appear denser than the WDM data, but seem to share similar internal structure (due to the same initial phases); 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 using DTM with a resolution of 2 Mpc for each side of the simulation cube and a tuning parameter of 0.001.

Because we only have one realization of the WDM simulation and one of the CDM simulation, we consider a modification of the two-sample hypothesis tests used in the simulation study of Section 4. Each simulation cube is divided into disjoint sub-cubes

[1]Because of the expansion of the universe, distances between objects change across time. Comoving distance removes the effect of the expansion. This is in contrast to a *proper distance*, which changes across cosmic time. For more discussion on distances in astronomy, see Hogg (1999).

(a) WDM Persistence Diagram (b) CDM Persistence Diagram

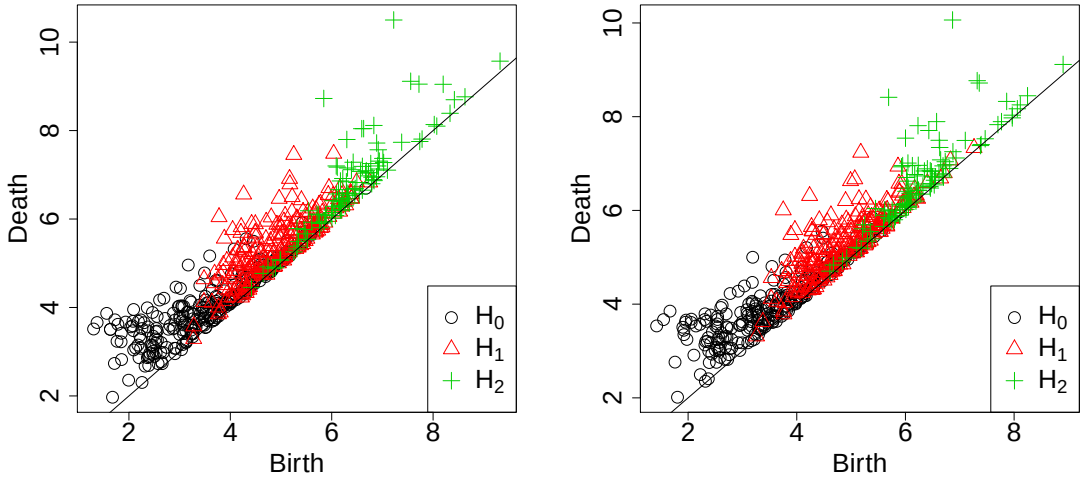


Figure 9: (a) Persistence diagram for the WDM simulation of Figure 1a; (b) Persistence diagram for the CDM simulation of Figure 1b.

producing 2^3 and 4^3 samples. An additional issue is that the WDM and CDM simulations were generated using the same initial phase - this results in a correlation of the largest-scale structures between the WDM and CDM simulations. Rather than performing two-sample T-tests, the sub-cubes of the two simulation cubes are *paired* with the corresponding cube in the other simulation.

Table 2 shows both the standardized (right) and unstandardized (left) results of the proposed hypothesis tests on the cosmological simulation data.. The unstandardized results suggest that with a higher number of sub-cubes, we are able to focus on smaller-scale topology, finding more significant differences. As expected, standardizing the persistence diagrams results in higher p-values on average. Judging from our most sensitive test statistics, Table 2 suggest that statistically significant differences in topology exist between WDM and CDM realizations.

The results from Table 2 suggests that there are topological differences in the LSS under CDM and WDM. Next we explore where these differences in structure are most pronounced. The bottleneck distances between the persistence diagrams for each sub-cube

Unstandardized			Standardized		
Test	2^3 Sub-cubes	4^3 Sub-cubes	Test	2^3 Sub-cubes	4^3 Sub-cubes
EC	1.2e-06	7.4e-26	EC	1.1e-04	7.3e-12
$EC_{0:2}$	2.1e-05	7.4e-28	$EC_{0:2}$	1.3e-05	4.0e-14
EC_0	3.3e-08	3.8e-30	EC_0	1.4e-06	1.6e-12
EC_1	1.8e-05	8.2e-21	EC_1	0.024	0.001
EC_2	0.340	0.088	EC_2	0.004	2.0e-07
Sil_{EC}	7.7e-08	2.5e-20	Sil_{EC}	0.0122	0.005
$Sil_{0:2}$	1.9e-06	1.1e-33	$Sil_{0:2}$	4.3e-04	8.1e-12
Sil_0	3.0e-08	1.5e-34	Sil_0	6.5e-06	1.8e-11
Sil_1	1.2e-05	2.9e-23	Sil_1	0.028	3.9e-07
Sil_2	0.93	0.035	Sil_2	0.003	1.5e-04
KC_0	0.442	0.000	KC_0	0.986	0.856
KC_1	0.192	0.000	KC_1	0.962	0.956
KC_2	0.248	0.002	KC_2	0.962	0.260
WIK_0	0.084	0.000	WIK_0	0.092	0.000
WIK_1	0.051	0.000	WIK_1	0.066	0.000
WIK_2	0.496	0.000	WIK_2	0.459	0.001
PI	0.923	0.281	PI	0.999	0.306
CORR	6.7e-04	7.4e-16	CORR	0.0289	0.918

Table 2: P-values from hypothesis tests on the unstandardized (left) and standardized (right) WDM and CDM simulations using 2^3 and 4^3 sub-cubes. A p-value of 0.000 comes from a permutation test with no positive examples.

for H_0, H_1, and H_2 are computed, where the bottleneck distance between two persistence diagrams $\mathcal{D}_{1h}$ and $\mathcal{D}_{2h}$ for a homology dimension h is

$$\text{Bottleneck}(\mathcal{D}_{1h}, \mathcal{D}_{2h}) = \inf_{\eta:\mathcal{D}_{1h}\to\mathcal{D}_{2h}} \sup_{x\in\mathcal{D}_{1h}} \|x - \eta(x)\|_\infty$$

where η defines a bijection between the two persistence diagrams (allowing for matches to the diagonal if there is an imbalance of features on the diagrams), and $x \in \mathbb{R}^2$ (Edelsbrunner & Harer 2010).

The results for the full simulation cubes, 2^3 sub-cubes, and 4^4 sub-cubes are displayed in Figure 10. The standardized persistence diagrams result in smaller bottleneck distances, as expected. It is also clear that there are regions of topologically similar areas and partitions of topologically dissimilar areas between the WDM and CDM simulations. In Figure 11, the EC functions and two-point correlation functions are displayed for the sub-cubes with the highest bottleneck distance.

Jessi: add more discussion here.

6 Conclusion

In this paper, we presented a hypothesis testing framework built 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 topological hypotheses in a diverse array of fields that greatly improve over currently existing methods.

Jessi: add more discussion here.

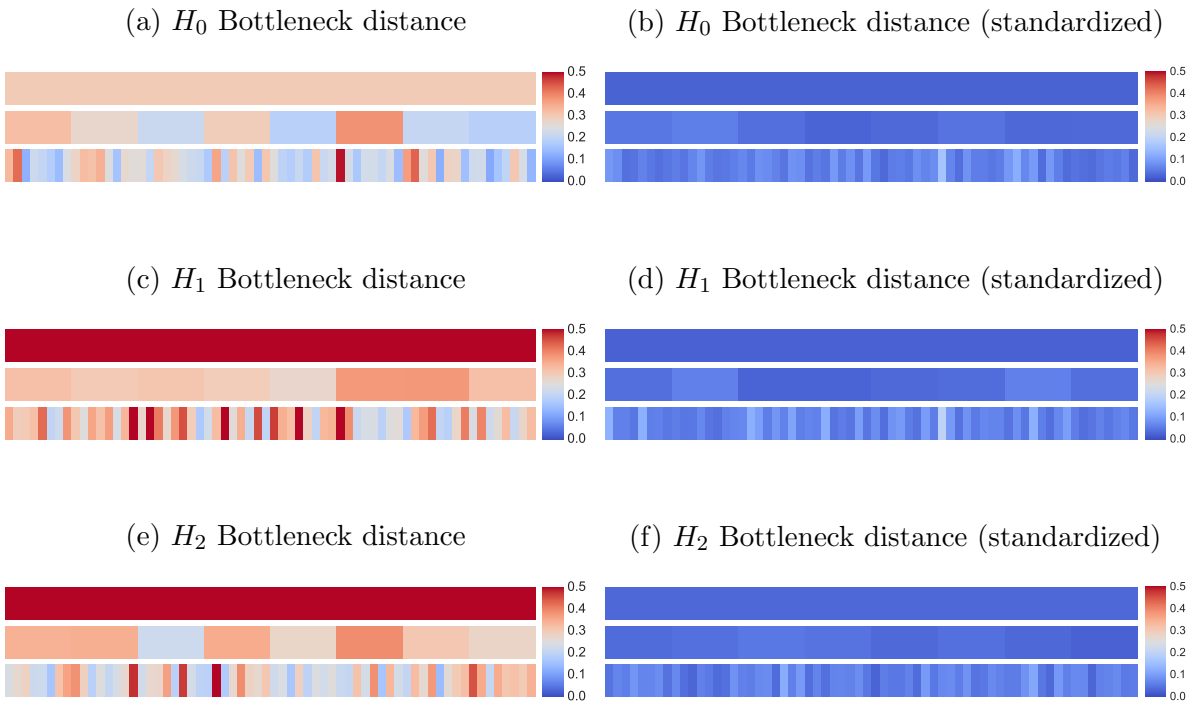
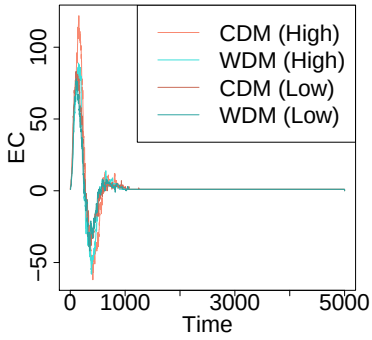


Figure 10: Bottleneck distances between the persistence diagrams of the CDM and WDM simulations for the full cube (first panel), 2^3 sub-cubes (2nd panel), and 4^3 sub-cubes (bottom panel). The left column (a, c, e) use the raw persistence diagrams while the right column (b, d, f) use the standardized persistence diagrams. Row one (a, b) are for H_0, row two (c, d) is for H_1, and row three (e, f) is for H_2.

(a) EC (High)



(b) CORR (High)

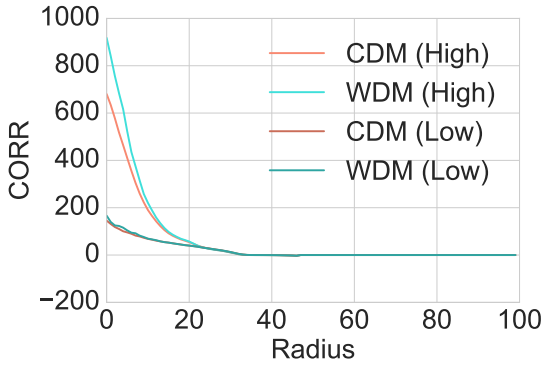


Figure 11: A comparison of EC and CORR between the cube with the highest bottleneck distance (High) and the cube with the lowest (Low). (a) and (b) clearly depict larger differences between CDM (High) and WDM (High) slices than between CDM (Low) and WDM (Low).

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* .

Baugh, C. (2006), 'Correlation function and power spectra in cosmology', *Encycl. of Astronomy and Astrophysics. IOP Publishing, Bristol* .

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.

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.

Hogg, D. W. (1999), 'Distance measures in cosmology', *arXiv preprint astro-ph/9905116* .

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.

Jarvis, M. (2015), 'TreeCorr: Two-point correlation functions', Astrophysics Source Code Library.

Jarvis, M., Bernstein, G. & Jain, B. (2004), 'The skewness of the aperture mass statistic', *Monthly Notices of the Royal Astronomical Society* **352**(1), 338–352.

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.

Mileyko, Y., Mukherjee, S. & Harer, J. (2011), 'Probability measures on the space of persistence diagrams', *Inverse Problems* **27**(12), 124007.

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

Sánchez, A. G., Scóccola, C., Ross, A., Percival, W., Manera, M., Montesano, F., Mazzalay, X., Cuesta, A., Eisenstein, D., Kazin, E. et al. (2012), 'The clustering of galaxies in the sdss-iii baryon oscillation spectroscopic survey: cosmological implications of the large-scale two-point correlation function', *Monthly Notices of the Royal Astronomical Society* **425**(1), 415–437.

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.

Turner, K., Mileyko, Y., Mukherjee, S. & Harer, J. (2014), 'Fréchet means for distributions of persistence diagrams', *Discrete & Computational Geometry* **52**(1), 44–70.

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.