Modeling and replicating statistical topology and evidence for CMB nonhomogeneity

Robert J. Adler[a,1], Sarit Agami[a], and Pratyush Pranav[a]

[a]Andrew and Erna Viterbi Faculty of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

Edited by Larry Wasserman, Carnegie Mellon University, Pittsburgh, PA, and approved August 29, 2017 (received for review April 25, 2017)

Under the banner of "big data," the detection and classification of structure in extremely large, high-dimensional, data sets are two of the central statistical challenges of our times. Among the most intriguing new approaches to this challenge is "TDA," or "topological data analysis," one of the primary aims of which is providing nonmetric, but topologically informative, preanalyses of data which make later, more quantitative, analyses feasible. While TDA rests on strong mathematical foundations from topology, in applications, it has faced challenges due to difficulties in handling issues of statistical reliability and robustness, often leading to an inability to make scientific claims with verifiable levels of statistical confidence. We propose a methodology for the parametric representation, estimation, and replication of persistence diagrams, the main diagnostic tool of TDA. The power of the methodology lies in the fact that even if only one persistence diagram is available for analysis—the typical case for big data applications—the replications permit conventional statistical hypothesis testing. The methodology is conceptually simple and computationally practical, and provides a broadly effective statistical framework for persistence diagram TDA analysis. We demonstrate the basic ideas on a toy example, and the power of the parametric approach to TDA modeling in an analysis of cosmic microwave background (CMB) nonhomogeneity.

persistence diagrams | Gibbs measures | topological data analysis | statistical topology | CMB nonhomogeneity

As a consequence of the current explosion in size, complexity, and dimensionality of data sets, there has been a growing need for the development of powerful but concise summary statistics and visualization methods that facilitate understanding and decision-making. A singularly novel approach, which has been particularly promising in areas as widespread as biology and medicine (1–3), neurophysiology, (4), cosmology (5, 6), and materials science (7), has been via the application of the powerful, and rather abstract, concepts of algebraic topology to develop what generally now falls under the label of "topological data analysis" (TDA). While approaching complex data from a topological viewpoint is not entirely new—it underlies Tukey's "Exploratory data analysis" of the 1960s (8) and the more recent approach by Friston and coworkers to brain imaging data (9)—TDA differs from all its forebears in its sophisticated exploitation of recent developments in computational topology. In particular, much of TDA has become almost synonymous with an analysis based on some version of persistent homology (10–12), represented visually as barcodes, persistence diagrams (PDs), or related representations (13–17).

With relatively few exceptions, notably refs. 17–22 (see additional citations in *SI Appendix*, *Homology and Persistent Homology*) TDA has not used statistical methodology as part of its approach, and, as a consequence, has typically been unable to associate clearly defined levels of statistical significance to its discoveries. While there may be a variety of reasons for this, one of the main obstacles to doing so is that the mathematical challenges involved in computing the statistical distributions of topological quantifiers have so far proven to be intractable. This is despite the fact that the measure-theoretic issues involved in defining probability measures which support notions such as expectations, variances, percentiles, and conditional probabilities have been effectively solved, for example, refs. 23–25.

One approach adopted by refs. 18–20 and others to circumvent these difficulties has been to reduce persistence diagrams to a single test statistic, often related to bottleneck norms, and then to adopt standard statistical resampling methods to analyze this statistic. If multiple diagrams are available, then the resampling can be done on them. However, since TDA is typically used in areas of very large data sets, the availability of replicates is rare, and consequently this approach is impracticable in most applications. An alternative approach is to (sub)sample points from the persistence diagram, and compute statistics on the subsamples. The problem with this approach, however, is that the true random object here is the full persistence diagram, and thus it is often unclear what is the precise meaning of the statistics produced this way.

We introduce an approach, based on generating a sequence of persistence diagrams which has similar statistical properties to those of the one generated by the data. The individual concepts underlying the method are not difficult, and follow a number of clearly defined stages. First, a parametric model is adopted that is sufficiently flexible to model an extremely wide class of persistence diagrams. The model we use is a class of Gibbs distributions, since these have a long history of success in modeling point sets (ref. 26 and its bibliography), which, essentially, is what a persistence diagram is. Having estimated parameters, we then exploit the fact that Gibbs distributions lend themselves to simulation by Markov chain Monte Carlo (MCMC) methods, and apply MCMC to produce a simulated sequence of diagrams.

Significance

Under the general heading of "topological data analysis" (TDA), the recent adoption of topological methods for the analysis of large, complex, and high-dimensional data sets has established that the abstract concepts of algebraic topology provide powerful tools for data analysis. However, despite the successes of TDA, most applications have lacked formal statistical veracity, primarily due to difficulties in deriving distributional information about topological descriptors. We present an approach, Replicating Statistical Topology (RST), which takes the most basic descriptor of TDA, the persistence diagram, and, using models based on Gibbs distributions and Markov chain Monte Carlo, provides replications of it. These allow for formal statistical hypothesis testing, without requiring costly, or perhaps intrinsically unavailable, replications of the original data set.

Author contributions: R.J.A. designed research; R.J.A. and S.A. performed research; S.A. and P.P. analyzed data; and R.J.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: radler@technion.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1706885114/-/DCSupplemental.

Since the underlying raison d'être of this approach is that persistence diagrams provide an excellent summary of topology, and statistics computed off the diagrams themselves furnish even more succinct summaries, we call this procedure "replicating statistical topology" (RST), and its introduction and descriptions of its implementation are the main contributions of the paper. We believe that these ideas, integrated with the existing techniques of TDA, provide another significant contribution toward putting TDA on a more solid statistical footing. To support this, we treat one toy example, showing that the technique works as predicted, and then study the fascinating and important topic of nonhomogeneities in the cosmic microwave background (CMB) radiation via parameter estimation of our Gibbs model.

TDA and Persistence Diagrams

As homology is an algebraic method for describing topological features of shapes and functions, so persistent homology is an extension of this method for both enriching these descriptions and for describing how topology undergoes changes. We shall use it to describe the upper-level sets of real-valued functions f defined over a space $\mathcal{X}$, namely, sets of the form $A_u = \{x \in \mathcal{X} : f(x) \geq u\}$. In basic homology, the topology of each A_u is often summarized by its Betti numbers, β_k, $k = 0, \ldots \dim(\mathcal{X})$. The first of these, β_0, counts the number of connected components in A_u, and, roughly speaking, the remaining β_k count the number of $(k+1)$-dimensional "holes" in A_u. Persistent homology goes farther, and keeps track of how homological features, including quantifiers such as the Betti numbers, persist and occasionally change as the level u drops, giving a richer, more dynamic view of topology. A very brief description of the underlying principles of persistent homology, with pointers to the literature, is given in *SI Appendix*.

Persistent homology is undeniably the most popular tool in the burgeoning area of TDA, one of the main reasons for which is the fact that it is easily visualized via barcode diagrams. Continuing with the upper-level set example of the previous paragraph, and starting with $u = +\infty$ and then descending, each bar in such a diagram is an interval that starts (is "born" at) a level $u = b$, at which a new aspect of the homology of A_u appears, and ends ("dies" at) a lower level $u = d < b$, as this aspect disappears. A mathematically equivalent, but visually distinct, representation of barcodes is as PDs of the points (b, d). We shall assume that the reader has some familiarity with these concepts, but we now look at an instructive, and easy, example in Fig. 1, needed later.

In Fig. 1, *Left*, we see a sample $\tilde{x}_N = \{x_1, \ldots, x_N\}$ of $N = 800$ points from two circles, of diameters 4 and 2. A random sample of 500 points were chosen at random from the larger circle, and 300 were chosen from the smaller one. In Fig. 1, *Middle*, we see the corresponding kernel density estimate, defined by

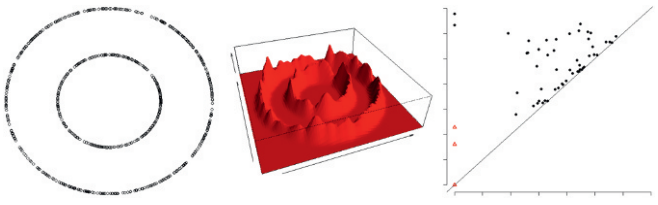

Fig. 1. A random sample from two circles: 500 points from the larger circle and 300 from the smaller one (*Left*) with a kernel density estimate (*Middle*) and the PD for its upper-level sets (*Right*). Black circles in the PD are H_0 persistence points, and red triangles are H_1 points. Birth times are on the vertical axis.

$$\hat{f}_N(p) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2\pi h^2} e^{-\|p-x_i\|^2/2h^2}, \qquad [1]$$

where $h > 0$ is a bandwidth parameter ($h = 0.1$ for Fig. 1). In Fig. 1, *Right*, we have the corresponding PD of the upper-level set filtration of $\hat{f}_N$, with the black circles indicating H_0 (zeroth homology) persistence and the red triangles corresponding to H_1, in both cases trying to capture the underlying homology of the two circles. As described above, each point in the diagram is a birth–death pair (b, d). The accepted paradigm of TDA, in examples of this type, is that points in the PD far away from the diagonal $b = d$ are meaningful, while points close to the diagonal, which represent short-lived topological phenomena, are not. Thus, since we know that the upper-level sets of $\hat{f}_N$ are characterized by having two main components, each of which contains a single one-cycle (hole), we expect to see two black circles and two red triangles somewhat isolated from the other points in the diagram. This is, in fact, the case.

While the PD in Fig. 1 performs as expected, and it is easy to identify the points that, a priori, we knew had to be there, there are many other points in the diagrams. Were we not in the situation of knowing ahead of time which points were "really" significant, it would not have been clear how to discount the additional points. We will treat this issue in *Example 1: Two Circles*.

There is another class of problems, in which the general structure of the PD, including points near the diagonal, is of more importance than the behavior of a handful of outliers. These are typically problems in which the topology is complex, and occurs at a number of different scales. Examples include tree-structured data objects such as brain artery trees (27), and a variety of cosmological structures (5, 6), including the CMB that we treat in *Example 2: CMB Nonhomogeneity*.

First, however, we must describe the general approach.

Gibbs Measures for Persistence Diagrams

Given a finite collection $\tilde{X}_N = \{X_1, \ldots, X_N\}$ of continuous random variables, with joint probability density $\varphi_\Theta(\tilde{x}_N)$, dependent on a multidimensional parameter Θ, we say that φ_Θ is a Gibbs distribution if it is written in the form

$$\varphi_\Theta(\tilde{x}_N) = \frac{1}{Z_\Theta}\exp(-H_\Theta(\tilde{x}_N)), \qquad [2]$$

where the "Hamiltonian" $H_\Theta : \mathbb{R}^N \to \mathbb{R}$ describes the "energy" of configurations $\tilde{x}_N$. Throughout the paper, we treat N as fixed, although a richer, more realistic model for PDs would treat it as random. The reasons for, and consequences of, this are discussed in *SI Appendix, Probability Models for Persistent Diagrams*. The normalization Z_Θ, actually a function of Θ, is known as the partition function, and is infamously hard to evaluate. All this is standard fare (28). However, we shall use Gibbs distributions to provide a model for PDs that look like those in Fig. 1. (There is another important family of PDs that arise from the construction of simplicial complexes over point sets, and these, at least for their H_0 diagrams, have all points with birth times identically zero. These are much easier to analyze, since they are effectively one-dimensional, and we shall treat them in a separate publication.)

As in all applications of Gibbs distributions, success depends on an appropriate choice for the energy function. Here is a way to do it for $\tilde{x}_N$, a set of N points in a subset $\mathcal{X}$ of $\mathbb{R}^2$. First, for $x \in \mathcal{X}$ and for $k \geq 1$, let $x^{nn}(k) \in \mathcal{X}$ be the kth-nearest neighbor to x, and set

$$\mathcal{L}_{\delta,k}(\tilde{x}_N) = \sum_{x \in \tilde{x}_N} \|x - x^{nn}(k)\| \mathbb{1}_{\{\|x - x^{nn}(k)\| \leq \delta\}}.$$

Restricting ourselves, for reasons to become clear soon, to $\mathcal{X}=\mathbb{R}\times\mathbb{R}_+$, define, for $x=(x^{(1)},x^{(2)})\in\mathcal{X}$,

$$\sigma_H^2=\sum_{x\in\tilde{x}_N}\left(x^{(1)}-\bar{x}^{(1)}\right)^2,\quad \sigma_V^2=\sum_{x\in\tilde{x}_N}\left(x^{(2)}\right)^2,$$

where $\bar{x}^{(1)}=N^{-1}\sum_{i=1}^N x_i^{(1)}$.

We now define a Hamiltonian, at effective interaction distance δ, up to cluster size $K\geq 0$, and, with interaction parameter $\Theta=(\theta_H,\theta_V,\theta_1,\ldots,\theta_K)$, as

$$H_{\delta,\Theta}^K(\tilde{x}_N)=\theta_H\sigma_H^2+\theta_V\sigma_V^2+\sum_{k=1}^K\delta^{-2}\theta_k\mathcal{L}_{\delta,k}(\tilde{x}_N). \quad [3]$$

The parameters here all have clear meanings. The horizontal spread about the mean of the points is controlled by σ_H^2, and the vertical spread is controlled by σ_V^2 [not centered because of the assumed nonnegativeness of the components $x^{(2)}$]. Each θ_k is a measure of energy density, controlling the probability of clusters of size $k+1$, with $\theta_k<0$ favoring such clusters, and $\theta_k>0$ lowering their probabilities. As noted in *SI Appendix, Probability Models for Persistent Diagrams*, working with energy densities rather than absolute energies (i.e., without the δ^{-2} factor in Eq. **3**) leads to more-robust numerical procedures.

Now, given a PD $\widetilde{B}=\{(b_i,d_i)\}_{i=1}^N$, define a new set of N points $\tilde{x}_N=\{x_i\}_{i=1}^N$, with $x_i^{(1)}=d_i$ and $x_i^{(2)}=b_i-d_i$. This (invertible) transformation has the effect of moving the points in Fig. 1 downward, so that the diagonal line projects onto the horizontal axis, but still leaves a visually informative diagram, which we shall call the projected PD (PPD). The statistical model we take for PPDs is a Gibbs distribution Eq. **2** with Hamiltonian Eq. **3**.

While this may seem a rather arbitrary form for the distribution of a PPD, there are a number of facts justifying it. The first is the trivial observation that any multivariate distribution can be written in the form of Eq. **2**, simply by taking $H_\Theta\equiv-\ln(\varphi_\Theta)$ and $Z_\Theta=1$. Moreover, "cluster expansions" of this form have been successfully used in statistical mechanics for the best part of a century as a basic approximation tool in the study of particle systems. More specifically, for the model to be rich enough for TDA, we need to choose the Hamiltonian from a parameterized family that comes close to spanning all "reasonable" functions on PPDs. However, we know from ref. 29 that the ring of algebraic functions on the space of PPDs is spanned by a family of monomials closely related to functions of the form of Eq. **3**. Finally, there is the issue, discussed in detail in *SI Appendix, Probability Models for Persistent Diagrams*, that we will often use these distributions not as exact models for PPDs but rather as a tool in a perturbative analysis. In these cases, the convenience of the models is more important than whether or not they provide a perfect fit to PPD data.

The determination of δ depends on the number and spread of the points in the PD. In practice, theoretical results (compare the reviews in refs. 30 and 31) suggest taking δ of the form

$$\delta=\frac{\delta^*}{N^{\alpha_{k,d}}}\max\left(\max|x_i^{(1)}-x_j^{(1)}|,\ \max|x_i^{(2)}-x_j^{(2)}|\right), \quad [4]$$

where $\alpha_{0,d}=1/d$, $\alpha_{k,d}=k/(k+1)d$ for $k\geq 1$, d is the dimension of the data underlying the PD, and δ^* is a data-independent tuning parameter. The terms inside the brackets in Eq. **4** scale for the order of magnitude of the data, which is unimportant topologically. (For cases for which d is unknown, setting $d=2$ seems to work in practice, merely leading to larger than usual values of δ^*, as does ignoring the fine structure of $\alpha_{k,d}$ and taking it to be $N^{-1/2}$, as a global default.)

Pseudolikelihood. Given H_Θ as a parametric model, we now turn to the estimation problem. Unfortunately, estimation of the parameters by a method such as direct maximum likelihood is precluded by the fact that we don't have an analytic form for Z_Θ, nor is there any way to compute it numerically in any reasonable time frame.

The standard way around this problem, which we adopt, is the pseudolikelihood approach (32, 33). This originated in the context of point cloud data with spatial dependence, which is, essentially, a description of a PD. In particular, it exploits the inherent spatial Markovianness of a Gibbs distribution to replace the overall probability of, in our case, a random PPD $\tilde{X}_N$ by the pseudolikelihood

$$L_{\delta,\Theta}^K(\tilde{x}_N)\triangleq\prod_{x\in\tilde{x}_N}f_\Theta\left(x\middle|\mathcal{N}_{\delta,K}(x)\right), \quad [5]$$

where $\mathcal{N}_{\delta,K}(x)$ denotes the points among the K nearest neighbors of x in $\tilde{x}_N$ which are of distance no greater than δ from x. If no such points exist, then we take $\mathcal{N}_{\delta,K}(x)=\emptyset$. The conditional, local densities $f_\Theta\left(x\middle|\mathcal{N}_{\delta,K}(x)\right)$ are given by

$$\frac{\exp\left(-H_{\delta,\Theta}^K\left(x\middle|\mathcal{N}_{\delta,K}(x)\right)\right)}{\int_{\mathbb{R}}\int_{\mathbb{R}_+}\exp\left(-H_{\delta,\Theta}^K\left(z\middle|\mathcal{N}_{\delta,K}(x)\right)\right)\,dz^{(1)}dz^{(2)}}, \quad [6]$$

and the conditional Hamiltonians $H_{\delta,\Theta}^K\left(x\middle|\mathcal{N}_{\delta,K}(x)\right)$ by

$$\theta_H\left[x^{(1)}-\bar{x}^{(1)}\right]^2+\theta_V\left(x^{(2)}\right)^2+\sum_{k=1}^K\delta^{-2}\theta_k\mathcal{L}_{\delta,k}\left(\mathcal{N}_{\delta,K}(x)\right).$$

Model Specification and Parameter Estimation. While it might be expected that PPDs originating from different physical phenomena might require quite different models, we have found, in all of the examples that we tried, that taking $K=2$ or 3 in Eq. **3**—so that the largest cluster size was 3 or 4—was both efficient and sufficient. If a lower K was appropriate, then the estimation procedure described above estimated the higher-order parameters θ_k as close to zero. In this case, using standard, automated statistical procedures such as AIC, BIC, etc. (cf. ref. 34), we often deleted the corresponding clusters from the Hamiltonian. Overall, we found the procedure not to be sensitive to either these small parameters or the specific procedure adopted for deleting them. After considerable experimentation, we found that working with all parameters appearing when $K=3$, regardless of their absolute value, was the easiest thing to do. We also found that taking $K>3$ did little to improve the simulation procedure, and typically led to manifestations of overfitting. In *SI Appendix, Probability Models for Persistent Diagrams*, we describe some of this experimentation, giving examples of when these models are, and are not, successful in fitting PDs.

RST and MCMC

We refer the reader to refs. 35 and 36 for technical background to this section, in which we describe a standard Metropolis–Hastings MCMC approach to replicating PDs. In particular, see ref. 35, section 10.3.3, in which the particular approach we take is called "Metropolis-within-Gibbs" and its properties are discussed.

Given a pseudolikelihood as in *Pseudolikelihood* (with known or estimated parameters), generating simulated replications of the associated point set via MCMC is not hard, but first we need some definitions.

First, given a $\tilde{x}_N$, define a "proposal distribution" $q(\cdot|\tilde{x}_N)$ as the bivariate Gaussian density, with mean vector and covariance matrix identical to the empirical mean and covariance of the points in $\tilde{x}_N$, but restricted to $\mathbb{R}\times\mathbb{R}_+$. Next, for two points $x,x^*\in\mathbb{R}\times\mathbb{R}_+$, define an "acceptance probability," according to which we will replace $x\in\tilde{x}_N$ by x^*, leading to the updated PPD

$\tilde{x}_N^*$, as

$$\rho(x, x^*) = \min\left\{1, \frac{f_\Theta(x^*|N_{\delta,K}(x)) \cdot q(x|\tilde{x}_N^*)}{f_\Theta(x|N_{\delta,K}(x)) \cdot q(x^*|\tilde{x}_N)}\right\}.$$

[Note that integration in the denominator of f_Θ in Eq. **6** depends on x only through its neighborhood, and so cancellation in the ratio means that it does not enter into the computation of $\rho(x, x^*)$. This makes MCMC for pseudolikelihoods much more computationally feasible than for full likelihood models.]

The basic step of the algorithm, which describes the update of the point set $\tilde{x}_N = (x_1, \ldots, x_N)$, is then *Algorithm 1*.

Algorithm 1. MCMC step updating diagram for $\tilde{x}_N$

1: $k = 0$
2: $k \leftarrow k + 1$
3: Choose x^* according to $q(\cdot|\tilde{x}_N)$
4: Compute $\rho(x_k, x^*)$
5: Choose U a standard uniform variable on $[0, 1]$
6: **if** $U < \rho(x_k, x^*)$, **then** set $x_k = x^*$
7: **if** $k < N$ **then** go to Step 2

To obtain N approximately independent PPDs, we adopt a procedure dependent on parameters n_b, n_r, and n_R: Starting with the original PPD, run *Algorithm 1* for a burn-in period. Then, starting with the final PPD from the burn-in, run the algorithm a further n_b times, choosing the last output of this block of n_b iterations as the first simulated PPD. Repeat n_r times, each time starting with the most recently simulated PPD, namely, the output of the previous block. Finally, replicate the entire procedure n_R times, for a total of $n = n_r \times n_R$ simulated PPDs. The optimal choice of n_b, n_r, and n_R typically depends on the specific problem, and the behavior of the Markov chain being simulated. See *SI Appendix, Probability Models for Persistent Diagrams* for details and practical guidelines for choosing these parameters.

Given the collection of n simulated PPDs, we convert each PPD back to a regular PD with the mapping $x \to (x^{(1)} + x^{(2)}, x^{(1)}) = (b, d)$ of its component points, and write $\mathcal{S}(\widetilde{B}) = \{\widehat{B}_1, \ldots, \widehat{B}_n\}$ for the resulting collection of simulated PDs generated from $\widetilde{B}$. These form the first-level output of the RST procedure.

The higher levels are driven by the specific application, but the basic idea is to compute simpler, real, vector, or function-valued statistics off the simulated PDs, and take their empirical distribution as an estimate of the true, underlying distribution of the statistic. The same statistic, computed off the original PD, can then be tested for statistical significance against this empirical distribution in standard fashion. Diagrammatically, treating persistence-based TDA as a sequence of three steps,

$$\text{physical phenomenon} \to \text{PD} \to \text{analysis},$$

RST comes in at the second stage, providing additional information on the variation of PDs to feed into the analyst's preferred method. This is best described by example.

Examples

We treat two examples. One is a toy problem, for which the true situation is known, to see how and if RST works. The second studies the topology of CMB, and the analysis required is far more subtle. Details of both are given in *SI Appendix*. For both cases, we emphasize the point implied in the preceding paragraph, that our main interest is in the replication of the PDs, and not the particular method of statistical analysis following that.

Example 1: Two Circles. As a simple (but representative) test case, take a random sample from two circles, as in Fig. 1. Note that, while there are many points corresponding to the H_0 homology, there are only three for H_1. Furthermore, the H_1 points are all closer to the diagonal boundary, and less prominent. [These are common phenomena for barcodes, addressed theoretically in a number of studies (e.g., ref. 37).] Consequently, the RST procedure will not work for H_1 in this example. However, we do not know of, nor can imagine, any statistical procedure that can reach a meaningful conclusion based on so few points. (The procedures such as those described in refs. 18 and 19 require some form of replication, usually via a bootstrapping approach, of the original data set. This is precisely what we are trying to avoid.) On the other hand, a homology which has, at most, three generators is small enough to be investigated ad hoc, and statistical procedures are hardly needed.

However, there are certainly enough H_0 points in Fig. 1 to fit a spatial model. Before we do this, note that there are two points (at the top left) that we know to be significant, since we know, a priori, that the data come from points on two circles. However, there are a number of other points far away from the diagonal, and, were we ignorant of the true situation, it would not be clear whether they were significant or not.

Adopting the approach of RST, we estimated the parameters for a Gibbs distribution for the model with pseudolikelihood Eq. **5** for the H_0 data, taking $K = 3$. For three different scenarios, we generated 1,000 simulated PDs from this model, each with a burn-in period of 1,000 iterations, with (n_b, n_r, n_R) given by (500, 20, 50), (500, 40, 25), or (500, 100, 10).

Using these three sets of simulations, we computed a number of statistics, but report on only one set here: the order statistics of the distances of the points in the PD to the diagonal. Given the points (b_i, d_i) of the PD, these are T_j, the jth largest among the differences $|d_i - b_i|, j = 1, \ldots, N$. Empirical distributions of the order statistics are then trivial to derive from the simulations of the PDs, and the order statistics calculated off the original PD can be compared with these. The results, for all three scenarios, showed that T_1 and T_2 were highly significant (the largest P value reached in any of the six cases was 0.003). The P values for T_3 were all in the range $(0.036, 0.041)$, and so T_3 was marginally significant at the standard 5% level. In none of the three scenarios was T_4 significant. Details of the analysis and the results are given in *SI Appendix, Probability Models for Persistent Diagrams*. These include a comparison with the bootstrap, confidence interval-based techniques of ref. 18. Using the same kernel bandwidth for the density estimate Eq. **1** that we used, these techniques identified only one point in each of the H_0 and H_1 diagrams as significant, indicating an underlying set topologically equivalent to a single circle, but missing the second circle. A similar analysis, using the related techniques in ref. 19, identified one H_0 point but no H_1 points at all. Adopting a different bandwidth, however, identified two points in each diagram, when using the methods of ref. 18.

In summary, blindly applying RST to simulate PDs, and taking the simplest of all statistics, showed (correctly) strong statistical evidence for two connected components in the topological space (two circles) which generated the PD, with borderline (but misleading) evidence for a third component. Despite the fact that the PD has a number of points far from the diagonal, and quite close to the third-farthest point (Fig. 1), these were (correctly) considered statistically insignificant. Thus, in this toy example, with the simplest of statistical quantifiers, RST competes favorably with existing methodologies.

Example 2: CMB Nonhomogeneity. Current cosmological theory describes a phase of rapid inflation in the primordial universe roughly 10^{-35} s after its birth. Spontaneous quantum fluctuations in what was then a high-energy, uniform, pseudovacuum universe resulted in minute perturbations in its density field. Eventually, aided by gravitational amplification, these

fluctuations led to the complicated, inhomogeneous structure of the cosmic web of planets, stars, galaxies, etc., which make up today's universe.

The CMB is the thermal radiation, generated as the universe cooled, some 300,000 years after the hypothesized Big Bang. Amazingly, it is measurable still today, and, since the temperature fluctuations in the observable CMB follow the pattern of the quantum perturbations from the inflationary era, it enables the mapping of the fluctuations in the distribution of matter in the early universe.

CMB data are directional, measuring fluctuations in radiation coming into a satellite from different directions. The first, satellite-based, detailed measurement of the CMB was carried out by the Cosmic Background Explorer (COBE) probe in the early 1990s, followed a decade later by the Wilkinson Microwave Anisotropy Probe experiment. Most recently, the high-precision Planck mission measured temperature anisotropies of the CMB to an accuracy of 10^{-5} degrees. They are measured at seven different frequency bands, and a resolution of 5' (5 arc minutes, or 5/60 of a degree), representing the most detailed and precise measurement of the CMB temperature anisotropies till date. Common to all of these, however, is that each CMB measurement is that of a function on a sphere, as in Fig. 2.

There are many mathematical models for CMB, the most common being a homogeneous, isotropic Gaussian random field (38–42). Both assumptions of Gaussianity and homogeneity have been challenged recently, from both theoretical and empirical viewpoints (43, 44), and it is homogeneity that we now address, parametrically, using our Gibbs model. (Non-Gaussianity has been addressed, geometrically, in refs. 45 and 46.)

To test homogeneity, we first cut out a ring around the equator of $\pm 30^\circ$, leaving "northern" and "southern" 60° spherical caps of data. The reason behind ignoring the central ring is that much of the data here are not from actual observations, which are unavailable due to confounding effects such as the Milky Way, but are "reconstructions" using one of a variety of techniques (47). Since all of these techniques are based on both Gaussian and homogeneity assumptions, the central ring should not show any deviation from the assumptions. Our aim is to test whether or not the CMBs in the two caps can be assumed to be realizations of the same stochastic process.

The next step is to generate five smoothed versions of the CMB in each cap, which we do with five different Gaussian kernels, with full width at half maximum 300', 180', 120', 90', and 60'. The highest level of smoothing (300') suppresses most of the fine-scale variation seen in Fig. 2, while the 60' level leads to no visually distinguishable difference. For each such smoothing, we produce PDs generated by the upper-level set filtration, for both H_0 and H_1, leading to a total of $20 = 5 \times 2 \times 2$ PDs. Although

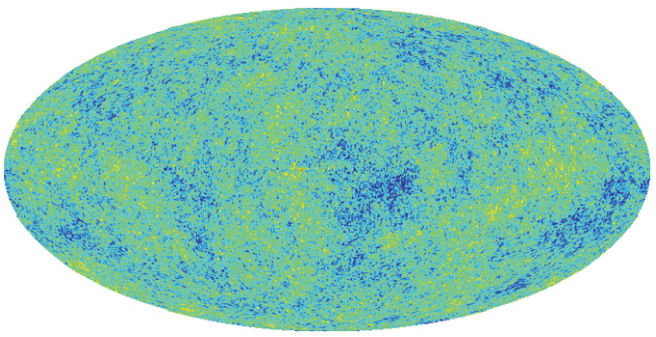

Fig. 2. A reconstructed version of CMB data from the Planck experiment, created using the Commander rule technique, seen in 2D Mollweide projection.

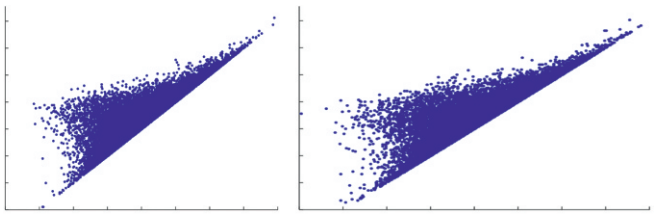

Fig. 3. H_1 PDs for unsmoothed CMB data, northern cap (*Left*) and southern cap (*Right*). There are approximately 27,000 points in each diagram. As usual, the vertical axis gives birth points, and the horizontal axis gives death points.

the aims there are different, details of the numerical procedure can be found in ref. 6, and an example of two PDs is given in Fig. 3.

The two PDs of Fig. 3 are quite similar, and, apart from a handful of outliers, it is hard to see any consistent differences between them. However, fitting a Gibbs model with pseudo-likelihood Eq. **5**, again taking $K = 3$, to each of the 20 PDs yields some interesting results, summarized in Table 1. Each such model involves five free parameters (we treat δ as a nuisance parameter only), and Table 1 gives the number of such parameters that, for each smoothing, and for each PD (H_0 or H_1), were found to be significantly different between the models for the northern and southern caps at a 5% significance level. Two different significance tests were used. Test 1 estimated the variance of the parameter estimates via empirical Fisher information, while Test 2 used external estimates of these variances based on simulations. Details are given in *SI Appendix, Example 2 - Analyzing CMB Data*. For reasons described there, we have more faith in the second, more conservative, of these tests.

The results provide statistical confirmation for differences between northern and southern PDs, with the most significant differences at the intermediate smoothing levels. (This is clearest from the P values associated with the test results, given in *SI Appendix*, Tables S2 and S3.) While we do not have a definitive physical explanation for this, it is most likely due to the effect of interactions between objects that evolved due to the true primordial CMB fluctuations and foreground phenomena that evolved at later epochs. However, whatever the cosmological reason underlying Table 1, the implication is that it is unreasonable to blandly assume that the northern and southern cap CMB maps are realizations of the same stochastic process. In other words, a hypothesis of homogeneity is questionable.

From the point of view of this paper, however, our main discovery is not cosmological but lies in demonstrating the ability of the Gibbs model, which assumes nothing about the original data nor about how PDs express properties of the underlying data, to differentiate between complex structures using purely topological methods. Consequently, we believe that the approach described

Table 1. Number of north–south CMB parameter differences for two tests, five levels of smoothing, and both homologies

		Level of smoothing				
Test	Homology	300′	180′	120′	90′	60′
Test 1	H_0	3	2	4	2	2
	H_1	0	3	3	2	4
Test 2	H_0	1	1	1	0	0
	H_1	0	2	0	0	0

See *SI Appendix, Example 2 - Analyzing CMB Data* for further details.

here will open the door to developing a wide variety of (semiparametric) statistical methods for further applications of TDA.

ACKNOWLEDGMENTS. Two splendid referees are responsible for making us work hard, think a lot, and produce over 30 pages of *Supporting Information*, which should clarify most of the imprecision forced on us by PNAS's six-page format. Among others, we thank Jose Blanchet, Herbert Edelsbrunner, Jingchen Liu, Anthea Monod, Sayan Mukerjee, and Katherine Turner for helpful conversations in various stages of this research, which was supported in part by the research projects Stochastic Algebraic Topology and Its Applications II (AFOSR, FA9550-15-1-0032) and Understanding Random Systems via Algebraic Topology (European Research Council Advanced Grant 320422).

1. Nicolau M, Levine A, Carlsson G (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci USA* 108:7265–7270.
2. Chan J, Carlsson G, Rabadan P (2013) Topology of viral evolution. *Proc Natl Acad Sci USA* 110:18566–18571.
3. Giusti C, Pastalkova E, Curto C, Itskov V (2015) Clique topology reveals intrinsic geometric structure in neural correlations. *Proc Natl Acad Sci USA* 112:13455–13460.
4. Curto C (2017) What can topology tell us about the neural code? *Bull Am Math Soc* 54:63–78.
5. Sousbie T (2011) The persistent cosmic web and its filamentary structure–I. Theory and implementation. *Mon Not Roy Astron Soc* 414:350–383.
6. Pranav P, et al. (2017) The topology of the cosmic web in terms of persistent Betti numbers. *Mon Not Roy Astron Soc* 465:4281–4310.
7. Hiraoka Y, et al. (2016) Hierarchical structures of amorphous solids characterized by persistent homology. *Proc Natl Acad Sci USA* 113:7035–7040.
8. Tukey JW (1962) The future of data analysis. *Ann Math Statist* 33:1–67.
9. Kilner JM, Friston KJ (2010) Topological inference for EEG and MEG. *Ann Appl Stat* 4:1272–1290.
10. Edelsbrunner H (2014) *A Short Course in Computational Geometry and Topology*, Springer Briefs in Applied Sciences and Technology (Springer, New York).
11. Edelsbrunner H, Harer J (2008) Persistent homology—A survey. *Surveys on Discrete and Computational Geometry*, Contemporary Mathematics (Am Math Soc, Providence, RI), Vol 453, pp 257–282.
12. Edelsbrunner H, Harer J (2010) *Computational Topology: An Introduction* (Am Math Soc, Providence, RI).
13. Carlsson G (2009) Topology and data. *Bull Am Math Soc* 46:255–308.
14. Carlsson G (2014) Topological pattern recognition for point cloud data. *Acta Numer* 23:289–368.
15. Ghrist R (2008) Barcodes: The persistent topology of data. *Bull Am Math Soc* 45:61–75.
16. Ghrist R (2014) *Elementary Applied Topology* (Createspace, North Charleston, SC).
17. Wasserman L (2018) Topological data analysis. *Annu Rev Stat Appl*, in press.
18. Chazal F, et al. (2014) Robust topological inference: Distance to a measure and kernel distance. arXiv:1412.7197v1.
19. Fasy B, et al. (2014) Confidence sets for persistence diagrams. *Ann Statist* 42:2301–2339.
20. Robinson A, Turner K (2013) Hypothesis testing for topological data analysis. arXiv:1310.7467v2.
21. Bobrowski O, Mukherjee S, Taylor J (2017) Topological consistency via kernel estimation. *Bernoulli* 23:288–328.
22. Bubenik P (2015) Statistical topological data analysis using persistence landscapes. *J Mach Learn Res* 16:77–102.
23. Mileyko Y, Mukherjee S, Harer J (2011) Probability measures on the space of persistence diagrams. *Inverse Probl* 27:124007.
24. Munch E, et al. (2015) Probabilistic Fréchet means for time varying persistence diagrams. *Electron J Stat* 9:1173–1204.
25. Turner K, Mileyko Y, Mukherjee S, Harer J (2014) Fréchet means for distributions of persistence diagrams. *Discrete Comput Geom* 52:44–70.
26. Banerjee S, Carlin BP, Gelfand AE (2015) *Hierarchical Modeling and Analysis for Spatial Data*, Monographs on Statistics and Applied Probability (CRC, Boca Raton, FL), 2nd Ed, Vol 135.
27. Bendich P, Marron JS, Miller E, Pieloch A, Skwerer S (2016) Persistent homology analysis of brain artery trees. *Ann Appl Stat* 10:198–218.
28. Georgii HO (2011) *Gibbs Measures and Phase Transitions*, de Gruyter Studies in Mathematics (Walter de Gruyter, Berlin), 2nd Ed, Vol 9.
29. Adcock A, Carlsson E, Carlsson G (2016) The ring of algebraic functions on persistence bar codes. *Homology Homotopy Appl* 18:381–402.
30. Kahle M (2014) Topology of random simplicial complexes: A survey. *Algebraic Topology: Applications and New Directions*, Contemporary Mathematics (Am Math Soc, Providence, RI), Vol 620, pp 201–221.
31. Bobrowski O, Kahle M (2014) Topology of random geometric complexes: A survey. arXiv:1409.4734.
32. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *J Roy Stat Soc Ser B* 36:192–236., and discussion (1974) 36:192–225, and reply (1974) 36:225–236.
33. Chalmond B (2003) *Modeling and Inverse Problems in Imaging Analysis*, trans Maître H, Applied Mathematical Sciences (Springer, New York), Vol 155.
34. Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference* (Springer, New York), 2nd Ed.
35. Robert CP, Casella G (2004) *Monte Carlo Statistical Methods*, Springer Texts in Statistics (Springer, New York), 2nd Ed.
36. Brooks S, Gemna A, Jones G, Meng XL (2011) *Handbook of Markov Chain Monte Carlo* (Chapman and Hall, Boca Raton).
37. Bobrowski O, Kahle M, Skraba P (2017) Maximally persistent cycles in random geometric complexes. *Ann Appl Probab* 27:2032–2060.
38. Smoot GF, et al. (1992) Structure in the COBE differential microwave radiometer first-year maps. *Astrophys J* 396:L1–L5.
39. Bennett CL, et al. (2003) First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Preliminary maps and basic results. *Astrophys J Suppl* 148:1–27.
40. Spergel DN, et al. (2007) Three-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Implications for cosmology. *Astrophys J Suppl* 170:377–408.
41. Komatsu E, et al. (2011) Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Cosmological interpretation. *Astrophys J Suppl* 192:18.
42. Planck Collaboration (2015) Planck 2015 results. XIII. Cosmological parameters. arXiv:1502.01589v3.
43. Eriksen HK, Hansen FK, Banday AJ, Gorski KM, Lilje PB (2004) Asymmetries in the cosmic microwave background anisotropy field. *Astrophys J* 605:14–20.
44. Park CG (2004) Non-Gaussian signatures in the temperature fluctuation observed by the Wilkinson Microwave Anisotropy Probe. *Mon Not Roy Astron Soc* 349:313–320.
45. Planck Collaboration (2016) Planck 2015 results. XVI. Isotropy and statistics of the CMB. *Astron Astrophys* 594:A16.
46. Buchert T, France MJ, Steiner F (2017) Model-independent analyses of non-Gaussianity in Planck CMB maps using Minkowski functionals. arXiv:1701.03347.
47. Planck Collaboration (2013) Planck 2013 results. XXIII. Isotropy and statistics of the CMB. arXiv:1303.5083v3.